

CHANGING USE OF FORMAL METHODS IN PHILOSOPHY: LATE 2000S VS. LATE 2010S

SAMUEL C. FLETCHER

UNIVERSITY OF MINNESOTA, TWIN CITIES

scfletch@umn.edu

JOSHUA KNOBE

YALE UNIVERSITY

joshua.knobe@yale.edu

GREGORY WHEELER

FRANKFURT SCHOOL OF FINANCE & MANAGEMENT

g.wheeler@fs.de

BRIAN ALLAN WOODCOCK

UNIVERSITY OF MINNESOTA, TWIN CITIES

brianwoodcock99@gmail.com

forthcoming in Synthese

ABSTRACT: Traditionally, logic has been the dominant formal method within philosophy. Are logical methods still dominant today, or have the types of formal methods used in philosophy changed in recent times? To address this question, we coded a sample of philosophy papers from the late 2000s and from the late 2010s for the formal methods they used. The results indicate that (a) the proportion of papers using logical methods remained more or less constant over that time period but (b) the proportion of papers using probabilistic methods was approximately *three times* higher in the late 2010s than it was in the late 2000s. Further analyses explored this change by looking more closely at specific methods, specific levels of technical engagement, and specific subdisciplines within philosophy. These analyses indicate that the increasing proportion of papers using probabilistic methods was pervasive, not confined to particular probabilistic methods, levels of sophistication, or subdisciplines.

KEYWORDS: Formal methods in philosophy, digital humanities, metaphilosophy, Philosophical Studies, logic, probability

1 INTRODUCTION

A few decades ago, the use of formal methods in philosophy was dominated by the application of a small number of closely related techniques. Above all, philosophers were concerned with the application of *logic*. Indeed, to the extent that philosophers turned to other areas for formal tools to conduct their research, it was often to areas that were intimately connected to logic (e.g., set theory).

Of course, philosophers recognized that other formal methods could sometimes be helpful in addressing philosophical questions, and some philosophical work did make use of such methods. Still, there was clearly a qualitative difference in prominence. The use of logic was an absolutely central aspect of philosophy, while the use of these other mathematical tools was more peripheral.

A question now arises as to whether there has been a change in the use of formal methods in philosophy over time. Is the use of formal methods in philosophy still dominated by logic, or has there been a shift toward diversification, with an increased use of other formal tools? If so, what tools? Have there been any shifts in the subdisciplines of philosophy that use these formal methods, or the level of sophistication of the methods used?

To address these questions, we looked at a sample of philosophy papers from the latter half of the 2000s and from the latter half of the 2010s. We coded these papers for whether they used logical methods and also for whether they used formal methods other than logic (probability theory, decision theory, game theory, statistics, causal modeling, etc.). We then asked whether the use of these different methods was changing over time.

In taking a data-oriented approach to our metaphilosophical questions, this paper joins a nascent body of work that uses empirical methods to study the changing nature of philosophy through large-scale analysis of articles published in philosophy journals over time. Malaterre, Chartier, and Pulizzotto (2019), for instance, explored the evolution of topics published in the journal *Philosophy of Science* over an eighty-year span, 1934–2015. They took advantage of sophisticated topic-modeling software that implements unsupervised machine learning algorithms to discover topics from common repeated word patterns in a large corpus.¹ Such tools are not (yet) clearly applicable to the present project, in which we are interested not in topics generally but in particular methods.

In work closer methodologically to our own, Bonino, Maffezioli, and Tripodi (2020) investigated the proportion of philosophy papers that use logical methods in five leading analytic journals over seventy years, from 1941–2010.² They were also interested in perceptions about trends concerning the centrality of logic and the sophistication of its usage. They sampled articles over a longer time period from one journal per year and from articles with the subject “logic” appearing in the Philosopher’s Index (<https://philindex.org/>), whereas we examined all the articles from one journal for years much closer together and within a much shorter time period. They partitioned logical methods into much finer-grained topics, but did not investigate non-logical methods.³ By contrast, the present study is centrally concerned with the role of these other, non-logical formal methods.

2 METHODS

2.1 JOURNAL AND YEAR SELECTION

Since it was not feasible to survey all the articles published within Anglophone philosophy in the time period of interest, we needed to tractably select a sufficiently representative sample. For this purpose, we decided to sample from one journal to ensure sufficient continuity in editorial policies. That journal would have to satisfy at least three criteria: it should

1. be a generalist journal widely recognized for the quality of its publications;

¹For other works on the history of philosophy through topic modeling, see those of Betti and van den Berg and colleagues (Betti and van den Berg 2014; Betti and van den Berg 2016; Betti, van den Berg, Oortwijn, and Treijtel 2019) and the other contributions to the second part of Fischer and Curtis (2019).

²They selected *The Journal of Philosophy*, *Mind*, *The Philosophical Review*, *The Proceedings of the Aristotelian Society*, and *Philosophy and Phenomenological Research*, so the corpus they investigate is distinct from the one we investigate.

³They did record whether articles have logic as their primary subject matter, although this is just one of the ten subdisciplines we consider, as mentioned above.

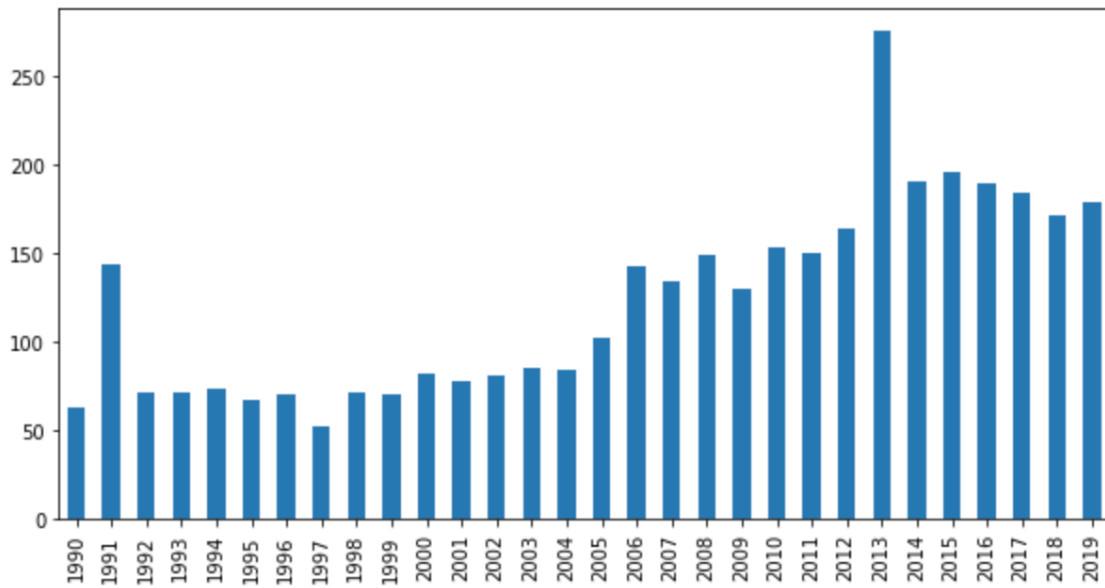


Figure 1: Number of papers published in *Philosophical Studies* between 1990 and 2019.

2. make its articles from the past couple of decades sufficiently accessible to download; and
3. publish a substantial number of articles per year.

We therefore selected *Philosophical Studies*. It is recognized as a generalist journal, has its catalog from the past couple of decades accessible through the Springer publishing system, and publishes over 100 articles per year—especially since 2005. (It was also used in an unpublished pilot study for the present project from 2019 by one of us, namely Josh Knobe.) It is, of course, not the *only* journal that satisfies the above criteria, and in principle we were open to including more journals depending on how quickly the data collection and classification described in the following subsections would proceed. However, we supposed it *prima facie* unlikely that we would proceed with such swiftness as to recommend selecting a second journal from which to sample. (As described below, that supposition was borne out.)

Having selected *Philosophical Studies*, our next goal was to select which years to classify. There were three main factors that drove our selection procedure:

1. Since many of the research questions we wished to answer involved intertemporal comparisons, we aimed to sample some years from the 2010s, some years from the 2000s, and if possible, some years from the 1990s.
2. Since *Philosophical Studies* began publishing over 100 articles per year regularly in 2005 (see Figure 1), we prioritized years late in each decade over years earlier in each decade.
3. In anticipation of the screening stage of our data collection (§2.3), in which we simply checked for the *use* of any formal methods in the articles reviewed, we

knew that we would set aside one year of the journal’s articles to train our formal methods screening team. The aforementioned unpublished pilot study used articles from 2016, so we decided to use this year for training purposes.

In light of these considerations, we prioritized years 2019, 2009, and 1999, followed by 2017, 2007, and 1997, then 2015, 2005, and 1995. However, as described in §2.2, we had technical difficulties scraping articles before 1999, and were only able to scrape articles in a slightly different order than originally planned. We were nevertheless still able to complete the screening process for 2019, 2017, 2015, 2009, 2007, 2005, and 1999.

2.2 DATA SCRAPING AND CLEANING

To collect the data, we first exported from *PhilPapers.org* a list of all *Philosophical Studies* articles published from 1990 to 2019 (cf. Figure 1). Next we began scraping the PDF articles from Springer’s website for *Philosophical Studies*, but were only able to access papers from 1999 to 2019 due to subscription restrictions at the Frankfurt School of Finance & Management. Given the pace of screening, as described in §2.3, we decided that we would not have sufficient time to screen data from 1995 and 1997, so we did not pursue an alternative means to acquire articles from the ’90s.

We then corrected our list of papers by cross-checking the PhilPapers data set against the populated subdirectories of downloaded articles. This process enabled us to eliminate duplicates and resolve inconsistencies.

2.3 SCREENING FOR USE OF FORMAL METHODS

After collecting PDFs and metadata for the articles to be classified, we assembled a team of philosophy undergraduates (juniors and seniors) at the University of Minnesota during the Fall, 2020 semester. For each article that a team member considered, we tasked them with determining *whether* the article uses formal methods in a way that contributes to the article’s arguments. We did not ask the team members to describe what methods, if any, were used, nor what subdiscipline of philosophy an article belonged to.

To provide guidance in this screening task, we developed a number of positive and negative heuristics that team members used to facilitate their judgment. The *presence* of formal notation was the central such positive heuristic, indicative that the article using the notation used formal methods. Formal notation consists in the presence of symbols, expressions, or graphics that are not properly a part of standard written English (or some other natural language) and are often considered as mathematical or logical. However, this positive heuristic had several negative heuristic defeaters:

- The symbols should not be merely labels or names for some arbitrary thing in a given class. If they could functionally be replaced with names, then they do not constitute a use of formal methods. For instance, sentences like “Consider a person p who intends to ϕ ” would not count by themselves as using formal methods for our purposes.
- The use of formal notation should not merely be in apposition to a non-formal description. For instance, consider a sentence like “Suppose that the electron is a state $|\psi\rangle + |\phi\rangle$, i.e., some superposition.” If only the non-formal description—that

the electron is in some superposition—plays a role in the subsequent argument, then the use of formal notation would not count as contributing to the article’s arguments.

- Formal notation used solely within a quotation counts only if the article’s argument draws from that notation. This precludes incidental inclusion of formal notation that plays no actual role in an article.

These counterbalancing negative heuristics preclude some sorts of minimal employments of formal notation from counting as using formal methods, but they allow for uses of formal notation that are very simple if those uses contribute in some way to the article’s argument. (As we describe in §2.4, we only subsequently classified articles that use formal methods by that usage’s level of sophistication.)

In order for team members to develop facility with and consensus on these heuristics, we employed three techniques. First, we began the semester with a meeting in which each team member was assigned to screen a pre-determined set of articles from 2016, some of which clearly used formal methods, some of which did not, and some of which were borderline cases, such as those only ultimately excluded by one the above negative heuristics. Second, all team members initially screened the same set of (“training”) articles from 2016, each reporting their evaluation in a separate Google sheet to which one of us, namely Sam Fletcher, had common access. All team members screened 93 articles from 2016, resulting in a Fleiss’ $\kappa = 0.708$, indicating substantial agreement.⁴ This provided input to the third technique, weekly meetings with the team members to discuss articles on which they had conflicting judgments.

We continued these meetings throughout the semester, even after team members received screening assignments beyond the initial training articles. (Conflicting ratings continued to provide discussion material for the aforementioned weekly meetings.) Our goal in giving team members assignments was to ensure that each article received at least **two** screening ratings. Team members worked at different rates throughout the semester and sometimes took their own initiative in screening further articles when they had finished an assignment, leading some articles to receive three screening ratings (and in some rare cases, more). At the end of the semester, about one sixth of the 2017 articles, all the 2007 articles, and most of the 2005 articles had only received one screening rating, so in the summer of 2021 we added a philosophy graduate student to the screening team to ensure that each of these articles received a second screening rating. That student underwent the same sort of training and weekly meetings as the rest of the screening team.

An article “passed” the screening process and was considered for the classification process described in the next subsection if and only if it was judged as using formal methods according to at least half of the screening ratings it received. For example, most articles received two screening ratings. In order for such an article to have passed to the classification process, it must have received at least one affirmative judgment that it used formal methods. And for any article that received three screening ratings, it must have

⁴These were the articles from the first six of twelve issues, with three exceptions of articles from those issues which not all team members rated. This kappa score includes the the screening ratings from the graduate student team member, whose participation we describe at the end of the following paragraph. See footnote 6 and the discussion surrounding it for more on conceptions of interrater reliability.

received at least two affirmative judgments that it used formal methods in order to have passed to the classification process. And so on.

2.4 CLASSIFICATION PROCESS

Above all, we were concerned with which formal methods the articles used, those users' levels of sophistication, and to which subdisciplines of philosophy those articles belonged. Instead of delineating these categories by introspection, we performed a pilot study with the 50 articles from 2016 that had previously passed the student screening process. Each of the four of us abductively determined our own delineation of these categories by examining these articles independently. Based on these separate delineations, we developed a consensus list of items within each category. Table 1 summarizes our consensus delineations.

Method	Level	Subdiscipline
<hr/> LOGIC FAMILY <hr/>		
None-Modal Logic	None (0)	Action & Free Will
Modal Logic	Fundamental (1)	Decision & Game Theory
Set Theory & Relations	Intermediate (2)	Epistemology
	Advanced (3)	Philosophy of Language
<hr/> PROBABILITY FAMILY <hr/>		
Probability Theory		Logic
Decision & Game Theory		Metaphysics
Statistics		Philosophy of Mind
Causal Modeling		Other
		Philosophy of Science
		Value Theory
<hr/>		
Other		
<hr/>		

Table 1: Classification Categories Used in Our Analysis

Within the **methods** category, we coded each article for whether it used the following: non-modal logic, modal logic, set theory and relations,⁵ probability theory, decision and game theory, statistics, causal modeling, or some other formal method. With one exception, these methods are not mutually exclusive, meaning that a single article could be coded as making use of more than one method. That exception is “non-modal logic” and “modal logic.” When any of us judged an article to use formal methods from logic, they decided whether those methods included ones from modal logic—whether syntactic, such as box and diamond operators, or semantic, such as Kripke or algebraic semantics. If the logical methods included one from modal logic, we would classify the article as using “modal logic”; otherwise, we would classify the article as using “non-modal logic.” So, articles coded as using “modal logic” may also use formal methods from parts of non-modal logic, but we never classified any article as using both.

⁵The method of “Set Theory & Relations” includes the general mathematical treatment of functions, n-tuples, and relations, as well as order and lattice theory.

Aside from this case, we therefore faced a question as to how much engagement with different methods would be required before a single article should be coded as using multiple methods. After all, some formal methods often and naturally include a cursory or elementary use of others. For example, probability theory routinely invokes elementary set theory or logic, and the Kripke semantics for modal logic invokes sets of worlds. We decided that formal methods assumed in the routine practice of another, like these, would not permit a second methods classification for the former. However, independent or atypical use of the such formal methods *would* demand this.

We grouped each of the methods, except for “other,” into one of two broader “families” of methods: the logic family and the probability family. All articles coded as using non-modal logic, modal logic, or set theory and relations were classified as using methods that fell into the logic family. All articles coded as using probability theory, statistics, decision and game theory, or causal modeling were classified as using methods that fell into the probability family. Like with the individual methods that comprise them, these two methods family classifications were neither mutually exclusive nor exhaustive. Thus, a single article could be classified as using methods that fell into the logic family and also methods that fell into the probability family. And, an article classified as using only “other” formal methods (e.g., real analysis, number theory, etc.) would not fall into either family. In the analyses reported below in §3, we will be concerned primarily with these broader families, but we will also provide information about the individual methods within each family.

Within the **levels** category, we coded each article on a single ordinal scale for the level of sophistication of its usage of formal methods: None (0), Fundamental (1), Intermediate (2), and Advanced (3). The “None (0)” rating allowed us to classify an article as not actually using formal methods, despite passing the screening process, as described in §2.3. These articles would be “false positives” for the use of formal methods produced by the initial screening process. Nevertheless, even when one of us applied a “None (0)” rating to an article, they still classified which formal methods that article would be using, under the supposition that it was to be classified as using formal methods. Although this may seem counterintuitive, it is probable that any article that passes the screening process comes close to employing formal methods substantively in its argument, even if we didn’t judge it ultimately to pass that threshold. Moreover, classifying the formal methods that an article used, if that article were to be classified as using a formal method, is potentially useful data when resolving certain conflicting level classifications (e.g., 0 versus 1), which we discuss in section 2.5.

For the remaining level classifications, we used the following heuristics, each illustrated by an example.

1. A fundamental use of formal methods requires only the fundamentals of a method to understand, such as that which a couple weeks’ experience in an average introductory course concerning the subject (or a couple hours of reading online) would afford. An example would be formalizing an argument in propositional logic to demonstrate that it has a particular form—say, that of *modus ponens*.
2. Understanding an intermediate use of formal methods requires experience closer to that of a full semester of an average introductory course on the method. An article employing multi-step formal proofs within a system of natural deduction

for first-order logic, for example, would likely be classified as intermediate.

3. An advanced use of formal methods is any which presupposes more experience than that of a full semester of an average introductory course on the method or some other advanced study to understand. For instance, an article comparing the metalogical features of some alternatives to classical logic, with novel proofs of properties such as completeness or compactness, would be classified as advanced.

Within the **subdiscipline** category, we coded each article for whether its investigations fell under the following: philosophy of action and free will, decision and game theory, epistemology, philosophy of language, logic, metaphysics, philosophy of mind, philosophy of science, value theory, or some other subdiscipline. Like with (most of) the classifications in the methods category, the classifications in the subdiscipline category were not mutually exclusive. An article might be classified appropriately as research in both epistemology and metaphysics, for instance. Also like with the methods category, we faced a question as to how much topical engagement with different subdisciplines would be required before a single article should be coded as falling within multiple subdisciplines. After all, some topics in some subdisciplines often and naturally include a cursory or elementary reference to topics in others. For example, many topics in the philosophy of language routinely invoke elementary considerations in logic. We decided that topics in one subdiscipline raised in the routine practice of another, like these, would not permit a second subdiscipline classification for the former. However, independent or atypical invocation of such topics would demand this.

Given these categories, we made classification assignments to each of the four of us authors using the following procedure. First, we had identified 245 articles as candidates for using a formal method—224 from the initial screening process and 21 articles from the added screening in the summer of 2021. Next, we shuffled these articles pseudorandomly. We did the same with a six-element list of all the possible pairs of classifiers (authors). Finally, we divided the randomized list of 245 articles into six nearly equal chunks, matching them in order with the six-element list of classifier (author) pairs. So, each classifier was assigned roughly half of the articles and each article was assigned to two classifiers. We followed this process after each of the initial screening for 224 articles and the added summer screening for 21 articles.

2.5 CLASSIFICATION RESOLUTION AND FINAL DATA SET

Inevitably, there were some differences between the two classifications each article received for each category. We report results for the interrater reliability of each classification item in Table 2, including for classification into one of the two families of methods, the “Logic Family” and the “Probability Family.”⁶ Roughly, measures of interrater

⁶ “Level” classification lends itself to two types of treatment. As a single, 4-level, ordinal variable, interrater reliability for the “Level” classification can be straightforwardly calculated using Spearman’s ρ (Sheskin 2011, sec. 29). However, the zero-versus-nonzero split was always interpreted with special significance. We saw two questions in this variable: (i) “Was a formal method used?” and (ii) “With what degree of sophistication was a formal method used when it was used?” Correspondingly, then, one might be interested in the reliabilities for those two decisions: (i) A nominal decision to classify an article as level 0 versus one of the upper levels, and (ii) an ordinal decision about which level to assign to an article that is not level 0.

reliability aim to capture the degree of consistency between raters, greater consistency indicating a higher quality measurement. Since each article was classified by two classifiers, *percent agreement* for a given category is simply the percentage of the articles in the category for which the classifiers’ evaluations agree. This, however, does not take into account the amount of agreement that would be due to chance if the classifiers had made their evaluations randomly. *Cohen’s kappa*, κ , takes chance agreement into account. It is a measure of the difference between the observed agreement and the agreement due to chance, standardized to lie on a scale from -1 to 1 , where 1 indicates perfect agreement, 0 indicates chance agreement, and negative values indicate potential systematic disagreement. Although there is some subjectivity in interpreting positive kappa values, the following is a common scheme: almost perfect agreement (0.81 – 0.99), substantial agreement (0.61 – 0.80), moderate agreement (0.41 – 0.60), fair agreement (0.21 – 0.40), and slight agreement (0.01 – 0.20).⁷ On this scheme, a kappa of 0.81 for the “Logic Family” and a kappa of 0.84 for the “Probability Family” both indicate almost perfect agreement.

Method	% agree	κ	Subdiscipline	% agree	κ
Logic Family	92%	0.81	Action	95%	0.54
Non-Modal Logic	77%	0.52	Decision	96%	0.67
Modal Logic	84%	0.59	Epistemology	90%	0.72
<u>Sets & Relations</u>	84%	0.34	Language	91%	0.67
Probability Family	93%	0.84	Logic	91%	0.54
Probability Theory	93%	0.71	Metaphysics	91%	0.79
Decision & Game	95%	0.71	Mind	96%	0.59
Statistics	100%	0.93	Other	96%	-0.02
<u>Causal Modeling</u>	98%	0.66	Science	95%	0.62
Other	92%	0.50	Value	95%	0.74

Level as Nominal	% agree	κ	Level as Ordinal	Spearman’s	p-value
0 vs Upper levels	85%	0.41	Upper levels (1, 2, 3)	$\rho = .58$	$p < .001$
			All levels (0, 1, 2, 3)	$\rho = .68$	$p < .001$

Table 2: Interrater reliabilities reported as both percent agreement and Cohen’s κ (except for the “level” category when treated as an ordinal variable, as discussed in footnote 6). These reliabilities concern all articles that passed the screening stage and underwent classification—except for the 1999 articles, for the reasons discussed at the end of section 2.5.

We assigned each article that received a discrepant pair of classifications to a third, distinct classifier, randomly selected among the two who had not classified that article. For almost all categories of interest, that distinct classifier resolved the discrepancy by selecting a value among the two discrepant possibilities. The only exception was when the discrepancy involved the “level” category with a separation of more than one ordinal

⁷For more on Cohen’s kappa, see Viera and Garrett (2005), and for more on measures of interrater reliability in general, see, e.g., Shoukri (2010, ch. 5–6).

value. In this case, the resolution could also include any value of “level” in the interval inclusively between the two discrepant values. This resolution process was followed after both the initial screening and the added summer screening.

After the completion of the initial round of resolutions, two things occurred which affected the shape of our final data set. First, we addressed some remaining data cleaning issues in the full set of 7 years of article data that had been collected for the study. We removed 5 (inexact) duplicate rows in the data set and 4 articles that, although they had appeared online in 2019, were not officially published until 2020. We also removed 17 articles that, upon investigation, were evidently not research articles and, hence, failed to qualify as the intended objects of the study. These articles came in two types. One included corrections to previously published research articles, signalled by words in the title such as “Erratum” and “Correction.” The other were introductions to special journal discussions signalled by the words “Preface” and “Introduction” in their titles. Hence, our overall data set for the 7 years of the study was reduced from 1001 to 975 articles.

Second, we decided to exclude the 1999 articles from our final analysis, reducing the total number of articles from 975 to 909. The absence of articles from 1995 and 1997 already made this sample from the 1990s seem deficient as compared with the samples from the 2000s and the 2010s. In addition, the small number of articles (66) from that year yielded an even smaller number of articles (17) actually using formal methods. So, we concluded that any results about the use of formal methods from that year alone may well not be representative of the late 1990s.

In summary, the final data set consisted of 909 articles from the chosen 6 years of the study (2005, 2007, 2009, 2015, 2017, 2019), of which 223 passed the screening process. Of those, we (the authors) determined through the classification and resolution processes that 191 genuinely used formal methods. The difference of 32 represents the number of false positives from the screening process, which is 14% of the 223 articles that passed the screening process. Our classifications identified 136 articles as using methods in the “Logic Family” and 61 articles as using methods in the “Probability Family.” Note that the sum of these (136 + 61) is not the same as the number which passed the resolution process (191) since we classified some articles as using methods belonging to both families and some articles as using some “Other” method not in either family.

3 RESULTS

Our most central question was about the change over time within each broad family of methods. The results for that question are displayed in Figure 2. As the figure shows, the proportion of papers using methods from the logic family in the most recent time period (2015–2019) is relatively close to what we found in the earlier time period (2005–2009). By contrast, the proportion using methods from the probability family in the most recent time period is *more than three times* as high as what we found in the earlier time period.

We conducted a series of statistical analyses to determine whether each of the changes displayed in this figure was statistically significant. To begin with, we simply looked at the change over time in the proportion of papers that use any formal methods at all. Binary logistic regression was employed for this analysis with year of publication as the predictor and whether an article used formal methods (1 vs. 0) as the dependent variable.⁸

⁸For more on the methods of binary and multinomial logistic regression used in our analyses, see, e.g.,

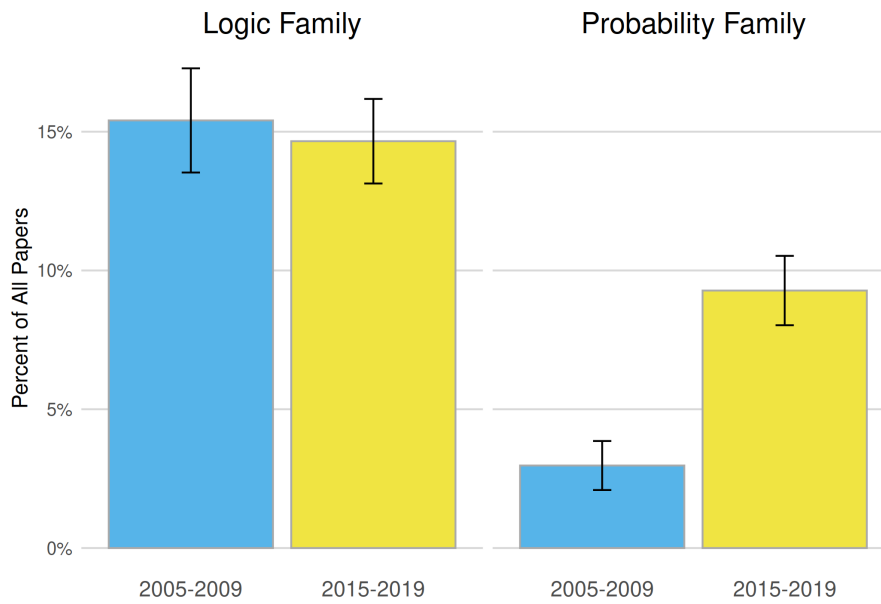


Figure 2: Papers in each family by time period. Error bars show standard error.

Note that in this analysis, we treated year of publication as a continuous variable (rather than dichotomizing into two time periods). Overall, we found no significant change over time in the proportion of papers that used formal methods: $\beta = 0.003$, $SE = 0.003$, $p = 0.29$, $OR = 1.007$, 95% CI [.998, 1.008].

Next, we looked separately at methods in the logic family and methods in the probability family. These analyses allow us to determine whether there is a significant change over time in the use of methods in the logic family and, separately, whether there is a significant change over time in the use of methods in the probability family. To answer these questions, we ran two separate binary logistic regression analyses. In one analysis, the dependent variable was use of methods in the logic family (1 vs. 0), and the predictor was year of publication. This analysis showed no significant change in the use of broadly logical methods over time: $\beta = -0.001$, $SE = 0.002$, $p = 0.67$, $OR = 1.00$, 95% CI [0.994, 1.003]. In the other analysis, the dependent variable was use of methods in the probability family, and again, the predictor was year of publication. This analysis showed that the use of methods in the probability family is increasing over time: $\beta = .006$, $SE = 0.002$, $p < 0.001$, $OR = 1.006$, 95% CI [1.002, 1.009]. These results confirm the informal impression one might get from looking at the figure. There was no significant change over time in the use of logical methods, but there was a very large and statistically significant change over time in the use of probabilistic methods.

A slightly different way of exploring this same change over time is to look separately at each time period and, within each time period, to compare the proportion of papers in the two different families. Using this approach, we can look at the earlier time period and ask whether there was greater use of logical methods than of probabilistic methods in that time period. Then we can look at the later time period and ask the same question.

Hosmer et al. (2013).

In the earlier time period, there was an extremely large effect whereby there were more papers using methods in the logic family (15.4%) than using methods in the probability family (3.0%): McNemar’s exact, $p < .001$, $OR = 6.11$, 95% CI [2.33,9.08].⁹ In the more recent time period, there is still a significant effect whereby there are more papers using methods in the logic family (14.7%) than in the probability family (9.3%), but the difference is now considerably smaller: McNemar’s exact, $p = .006$, $OR = 1.74$, 95% CI [1.16,2.66].

We were very struck by the dramatic change over time in the proportion of papers using methods from the probability family. To get a better sense of what might be driving that change, we can turn to our other variables (specific method, level, and subdiscipline). The hope is that breaking things down using those other variables will give us some further insight into the nature of the change over time.

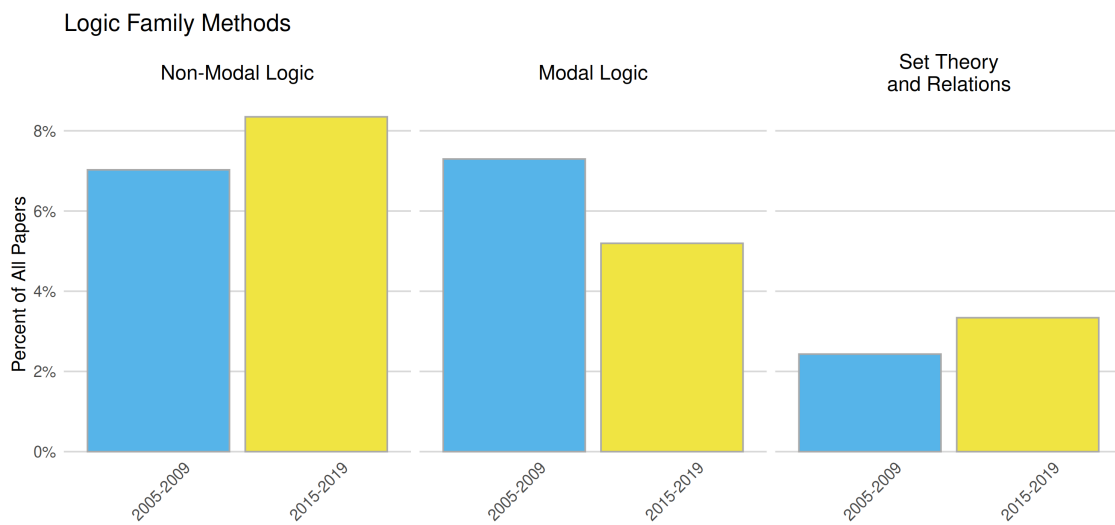


Figure 3: Papers within each specific method in the logic family by time period.

First, we looked at the change over time in the use of each specific **method** within each of the two families. Figure 3 shows the change over time in methods within the logic family. Figure 4 shows the change over time in methods within the probability family. Our aim in breaking things down in this way was in part to get a better understanding of the dramatic increase in use of methods from the probability family. Was this increase driven primarily by an increase in the use of one specific sort of probabilistic method? Or was there a broader change that involved an increase in the use of a number of different sorts of probabilistic methods?

As Figure 4 shows, there was an extremely sharp increase between the two time periods in all four of the different methods within the probability family. In the papers in the more recent time period, we find approximately twice as much use of probability theory, six times as much use of decision and game theory, five times as much use of statistics, and four times as much use of causal modeling. Given that these four different methods tend to be used by separate groups of philosophers and are usually applied to

⁹For more on McNemar’s exact test of marginal homogeneity, see, e.g., Sheskin (2011, sec. 20).

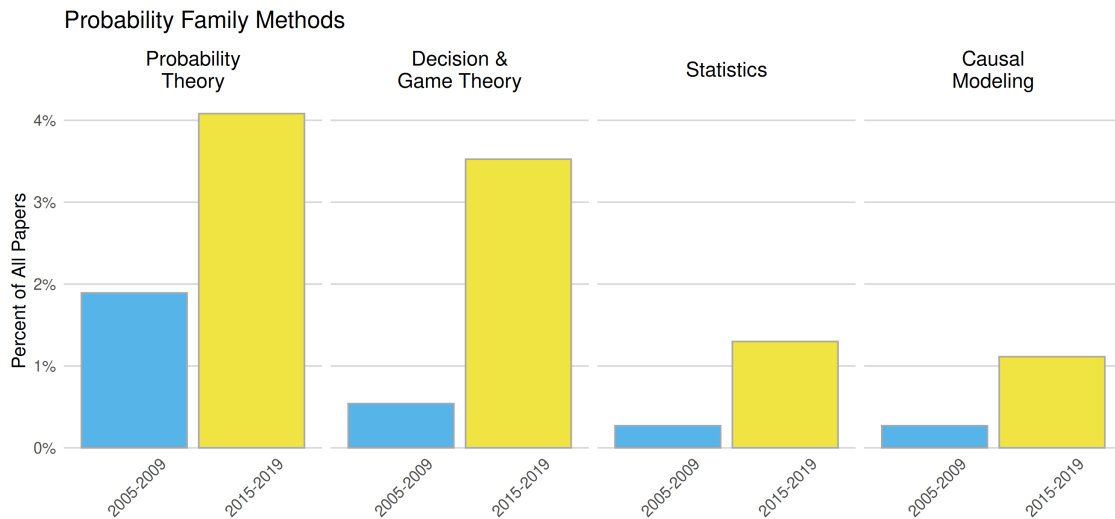


Figure 4: Papers within each specific method in the probability family by time period.

very different areas of philosophy, the fact that all four showed such large increases is quite striking.

To further explore these changes, we ran a separate binary logistic regression analysis for each separate method. The results are shown in Table 3. As the table shows, the only method for which there was a significant change considered just by itself was decision and game theory, which underwent an especially dramatic increase over this time period. A question therefore arises as to whether the entire increase in the use of methods within the probability family was driven by the increase in the use of decision and game theory. To address this question, we ran an additional analysis in which we excluded from the data set all papers that used decision and game theory and then asked whether there was still a significant increase in papers using methods in the probability family. This analysis showed that, even after this exclusion, the increase in probability family papers was still significant: $\beta = 0.003$, $SE = 0.001$, $p = 0.03$, $OR = 1.003$, 95% CI [1.0003, 1.0057]. In other words, when it comes to the other three probability family methods (probability theory, statistics, and causal modeling), no individual method showed a statistically significant increase considered just in itself, but the three methods do show a statistically significant increase when considered together.

Next, we examined the change over time in papers within each family at each of the three **levels**. Figure 5 shows the change over time in papers at each level within both the logic family and the probability family. The figure once again shows the sharp increase in papers within the probability family, but this time, one can see the increase broken down by level. The key question we want to address now is whether the increase was driven primarily by papers at one specific level or whether there was a more pervasive change, driven by an increase in papers in the probability family at multiple different levels.

	β	SE	t	p	OR	95% CI	
						Lower	Upper
Non-Modal Logic	0.0005	0.0018	0.2800	0.7790	1.0005	0.9970	1.0040
Modal Logic	-0.0017	0.0016	-1.0770	0.2820	0.9983	0.9952	1.0014
Set Theory	0.0008	0.0011	0.7200	0.4720	1.0008	0.9986	1.0030
Probability Theory	0.0019	0.0012	1.6750	0.0943	1.0019	0.9997	1.0048
Decision Theory	0.0028	0.0010	2.8400	0.0046	1.0028	1.0009	1.0048
Statistics	0.0009	0.0006	1.4250	0.1550	1.0009	0.9997	1.0021
Causal Modeling	0.0006	0.0006	1.0670	0.2860	1.0006	0.9995	1.0017

Table 3: Results for each separate binary logistic regression, taking each of the separate methods as a dependent variable and year of publication as a predictor.

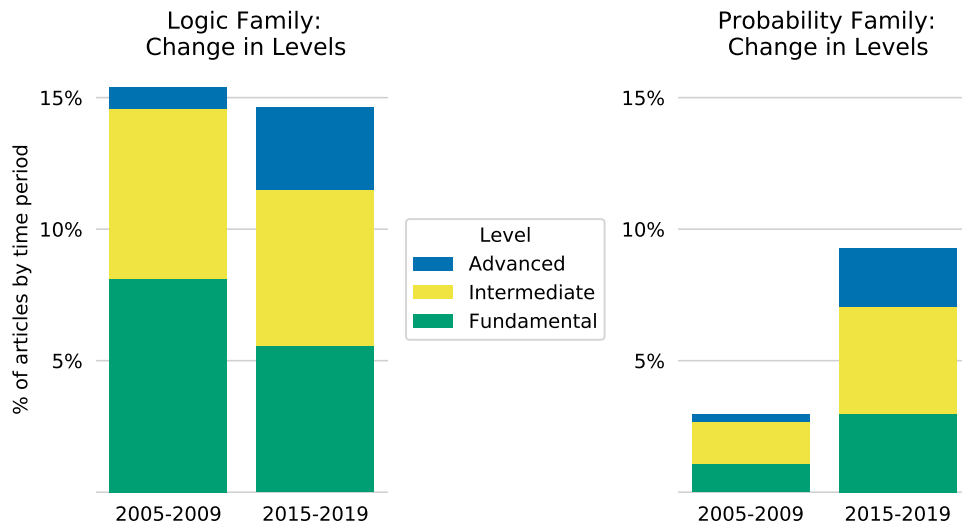


Figure 5: Percent of papers using methods in the logic family and in the probability family at each level by time period

To address this question, we conducted a separate multinomial logistic regression for each family, with level as the dependent variable and year of publication as the predictor. In both regressions, the reference level was “None (0)” (i.e., no use of methods within the family). Tables 4 and 5 show the results.

	β	SE	Wald	p	Exp(β)	95% CI	
						Lower	Upper
Fundamental	-0.038	0.026	2.087	0.149	0.963	0.914	1.014
Intermediate	-0.020	0.027	0.539	0.463	0.980	0.929	1.034
Advanced	0.151	0.061	6.088	0.014	1.163	1.032	1.311

Table 4: Results for multinomial logistic regression, taking level of use of methods in the logic family (fundamental, intermediate, advanced, or none) as a dependent variable and year of publication as a predictor. “None” served as the reference level.

	β	SE	Wald	p	Exp(β)	95% CI	
						Lower	Upper
Fundamental	0.096	0.052	3.432	0.064	1.101	0.994	1.219
Intermediate	0.082	0.043	3.653	0.056	1.085	0.998	1.180
Advanced	0.164	0.077	4.549	0.033	1.178	1.013	1.370

Table 5: Results for multinomial logistic regression, taking level of use of methods in the probability family (fundamental, intermediate, advanced, or none) as a dependent variable and year of publication as a predictor. “None” served as the reference level.

This approach allows us to ask whether or not there is a significant change in the proportion of papers in a given family at each level. So, for example, in looking at the logic family, we can ask whether there was a significant change in the fundamental level, whether there was a significant change in the intermediate level, and so forth. Within the logic family papers, there was a statistically significant increase in the proportion of papers at an advanced level, but the proportion of papers at the fundamental and intermediate levels decreased enough that there was no overall significant change in the proportion of logic family papers, even though the latter decreases themselves did not reach statistical significance. Within the probability family, there was a significant increase over time in papers at the advanced level. There were trends toward an increase in probability family papers at the fundamental level and the intermediate level, but neither of those two trends was statistically significant in itself.

To further explore these changes, we used the same approach that we used above in the analysis of individual methods. We wanted to know whether the increase over time in papers in the probability family was driven entirely by papers at the advanced level, so we excluded from the data set all papers from the probability family at the advanced level and then asked whether there was still a significant increase over time in probability family papers. This analysis showed that, even when we exclude probability papers at the advanced level, we still find a significant increase over time in papers in the probability family: $\beta = .0040$, $SE = 0.001$, $p = 0.007$, $OR = 1.0040$, 95% CI [1.0011, 1.0070]. Overall, then, the data do not suggest that the change over time in the proportion of

probability papers is driven by a change at one specific level. Instead, there appears to be an increase over time in papers at multiple different levels.

Finally, we looked at the proportion of papers within each family in each separate **subdiscipline** of philosophy. Figure 6 shows the results for the two time periods of interest. Comparison shows in which subdisciplines there was relative growth in the use of probabilistic methods. More detail is available in Table 6, which shows the use of each method (and not just method family) by subdiscipline as a percentage of all the papers in that time period. Once again, breaking things down in this way enables us to get a deeper understanding of the increase in the use of probabilistic methods. Was the change driven by increased use of probabilistic methods in one specific subdiscipline? Or was there an increase in those methods in a number of different subdisciplines?

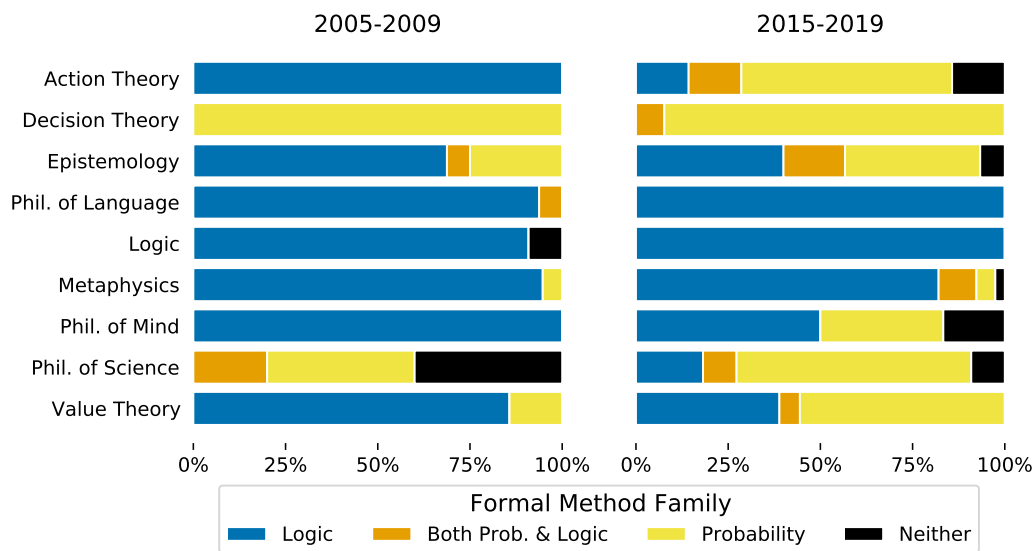


Figure 6: Percentage of papers within each family of formal methods in each subdiscipline of philosophy for the two time periods, 2005–2009 and 2015–2019.

Figure 6 shows that within each of the subdisciplines of action theory, epistemology, philosophy of mind, and value theory, the relative proportion of papers using probabilistic methods among those using any formal methods at all increased sizeably. To see this, in visual terms, note the substantial growth in the portions of the bars for these subdisciplines representing the use of probabilistic methods (either with or without the use of logical methods). In addition, the right hand side of Table 6 shows that, for these subdisciplines, the absolute proportion of papers using probabilistic methods among all papers in the time period investigated saw noticeable growth. The same part of the table yields the same conclusion for the subdisciplines of decision and game theory, metaphysics, and philosophy of science, even though these latter subdisciplines experienced little change in the relative proportion of their papers using probabilistic methods to those using any formal methods. In contrast, philosophy of language and logic remained bastions of the use of logical formal methods, as one might expect, both in absolute and relative terms.

(a) 2005–2009	Non-Modal Logic	Modal Logic	Sets & Relations	Probability Theory	Decision & Game	Statistics	Causal Modeling
Action	0.27%	0.27%	–	–	–	–	–
Decision	–	–	–	–	0.27%	–	–
Epistemology	0.81%	2.43%	–	1.08%	–	0.27%	–
Language	2.97%	1.35%	0.54%	0.27%	–	–	–
Logic	2.16%	0.54%	0.27%	–	–	–	–
Metaphysics	2.16%	2.16%	1.35%	–	–	–	0.27%
Mind	0.27%	0.27%	0.27%	–	–	–	–
Science	0.27%	–	–	0.81%	–	–	–
Value	0.54%	0.81%	0.27%	–	0.27%	–	–
(b) 2015–2019							
Action	0.19%	0.19%	–	–	0.19%	0.37%	0.37%
Decision	0.19%	–	–	0.19%	2.23%	0.19%	–
Epistemology	1.30%	1.86%	0.19%	2.60%	0.56%	–	–
Language	2.23%	0.19%	0.37%	–	–	–	–
Logic	2.04%	0.74%	0.37%	–	–	–	–
Metaphysics	3.53%	2.41%	1.86%	0.74%	–	0.19%	0.56%
Mind	0.19%	0.19%	0.19%	–	–	0.19%	0.19%
Science	0.19%	0.37%	0.19%	1.11%	0.19%	0.19%	0.19%
Value	0.74%	0.37%	0.56%	0.37%	1.11%	0.56%	–

Table 6: (a) Percent of all papers in 2005–2009 ($n = 370$) using each specific method by subdiscipline and (b), similarly, for 2015–2019 ($n = 539$). The dash symbol ‘–’ represents 0%. Note that each individual paper could be classified as using more than one method or belonging to more than one subdiscipline.

So, there appear to be methodological shifts in several subdisciplines of philosophy. Are there features in the data that might help us to understand what occurred locally to account for these shifts? Interestingly, although a number of different subdisciplines shifted toward greater use of probabilistic methods, papers in those different subdisciplines appear to be using probabilistic methods in quite different ways. Within action theory, there was an increase in the use of experimental philosophy.¹⁰ Pointing to the influence of the notion of expected utility, unsurprisingly the subdiscipline of decision and game theory saw an increased use of those formal methods, but so did value theory. Finally, in philosophy of mind, there was an increased engagement with methods from cognitive science.¹¹ The changes in these different subdisciplines might normally be regarded as more or less independent, but it is interesting to see that they all involve a shift in the same direction, namely, toward the increased use of methods within the probability family.

4 DISCUSSION

Our inquiry has been concerned with changes in the use of formal methods between 2005 and 2019. Looking at a sample of papers from that time period, we found that the use of methods in the logic family remained relatively constant, while the use of methods in the probability family increased dramatically. To better understand this change, we broke that sample down into those with more specific methods (statistics, causal modeling, etc.), specific levels of technical engagement (fundamental, intermediate, advanced), and specific subdisciplines (value theory, epistemology, etc.). These further analyses suggested that the change is surprisingly pervasive. There was a large increase in the use of methods from the probability family for a number of different methods, a number of different levels, and a number of different subdisciplines.

We do acknowledge that there are a number of limitations to our analyses. We describe the three most important, to our minds. First, our journal and year selections leave open the possibility that the trends we observe are not based on representative samples of work in anglophone analytic (or analytic-adjacent) philosophy during the time periods of interest. This could be due to any idiosyncratic editorial practices at *Philosophical Studies* or quirks of what sort of research tends to be submitted to it, such as biases for certain subdisciplines or geographic regions for authors. It is also plausible that the nature of our results would be different had we also examined journals specializing in or with an affinity for epistemology, logic, philosophy of language, or philosophy of science.

One obvious way to address this limitation would be to use more computational approaches. For example, it might be worthwhile to use the data set we did analyze to train a classifier to replace some or all of the human coding of papers, using our rater-reliability scores as a benchmark. This might permit analysis of not just all of the papers published

¹⁰In the earlier time period, there were only two papers in action theory that used formal methods, both in the logic family. By contrast, in the more recent time period, out of the seven papers using formal methods, we find five of them using methods in the probability family. Three of those used experimental methods.

¹¹Of the two articles in philosophy of mind in the later time period that used methods from the probability family, one used experimental methods from cognitive science and the other synthesized other reports of the results of applying such experimental methods.

in *Philosophical Studies* during this period, but also of those in the twenty leading generalist journals. The result would be a larger and more diverse sample of papers and hence, potentially, a more accurate picture of the change over time.

Second, our screeners' reliability ($\kappa = 0.708$), although substantial, was also not as high as we might have liked. This is ameliorated somewhat by demanding only that at least half of screeners rate an article to be using formal methods in order for it to advance to the classification stage, reducing the rate of false negatives at the cost of some false positives, which can be addressed at the classification stage. Future studies could do more to calibrate how many screenings each article would need in order to achieve a satisfactory screening error rate, given a certain reliability.

Third, while our classifications were very reliable when it came to judgments about the method families (i.e., the logic family and probability family), they were less so when it came to the individual methods within those families, the levels of sophistication, and the subdisciplines.¹² This qualifies the strength of the conclusions we can draw about changes within these categories. Future studies might improve this by implementing a lengthier "training" process before classification analogous to the one used to train the team for the screening process. This could bring classifiers closer in harmony regarding their conception of each classification.

Despite these limitations, our analyses still show significant evidence for certain broad trends. In particular, we find it quite striking that the proportion of papers using probabilistic methods increased so sharply over such a short period of time. This result becomes even more striking when one considers the fact that the change seems to be arising across different methods and subdisciplines that would not normally be regarded as intimately connected. The present findings thereby raise a number of questions that could be addressed in further work. The first concerns robustness and generalizability:

1. Can one confirm the trends we have discovered when taking into account broader and deeper samples of the published literature?

Assuming that they are so confirmed, there are three further questions that they naturally suggest:

2. What is the cause of these changes?
3. Can we predict whether the trends will continue or change in the near future?
4. What, if anything, does this signal about philosophy's changing formal methodology or its relationship with other parts of academic inquiry?

On the second question, regarding the cause of the changes, we can rule out some candidates. The changes do not likely reflect a *topical* fad, since the changes involve methods, not topics, and arise across a variety of subdisciplines of philosophy with little common topical overlap. They also are probably not due to external pressures or changes, such as from journal publication policies or government requirements.¹³ Thus,

¹²The exception to this seems to be articles that used statistics, which admit of many fewer borderline cases and are quite distinctive from other methods.

¹³By comparison, the International Committee of Medical Journal (ICMJ) Editors decided in 2004 to make clinical trial preregistration a precondition for publication in ICMJ journals, following developing laws mandating the same for certain clinical trials funded by public grants (DeAngelis et al. 2004; 2005).

the change seems to involve something within the discipline of philosophy that is leading philosophers to acquire and employ a new set of technical tools—a change that may reflect a move away from analytic philosophy’s traditional focus on the structure of language,¹⁴ toward modeling processes, or both (Wheeler 2013).

If we had been focusing only on the change observed for one specific method, or for one specific subdiscipline, certain other hypotheses would immediately suggest themselves. For example, if we were only trying to explain the increase in the use of probability theory, it would be natural to point to the rise of formal epistemology. Analogously, if we were only trying to explain the increase in the use of statistics, it would be natural to point to the rise of experimental philosophy. Similar hypotheses could be developed for decision theory, game theory, and causal modeling. But, clearly, there is a further fact here that each of these hypotheses would be failing to explain: We are seeing a dramatic increase in the use of *all* of these different methods in the *same* time period. Thus, it seems that we need some explanation as to why the use of all of these different probabilistic methods increased at the same time.¹⁵

Deepening the puzzle, it seems that there were similar developments in other disciplines at approximately the same time. Within psychology, there was a massive surge in research developing probabilistic models of cognition (Oaksford and Chater 2007; Tenenbaum, Kemp, Griffiths, and Goodman 2011). Similarly, within artificial intelligence, there was a sharp decrease in the use of logical methods (Hao 2019) and an explosion of new research using statistical methods (Krizhevsky, Sutskever, and Hinton 2012; LeCun, Bengio, and Hinton 2015). Thus, there is at least some reason to suspect that the change we observe in the discipline of philosophy is due to something far broader, which goes beyond philosophy specifically.

Turning now to the third question, we can ask about the trajectory of the discipline going into the future. If the use of methods within the probability family continues to increase at anything like the rate observed within the present study, then the use of methods in the probability family will soon be on par with or even outstrip the use of methods in the logic family overall and for certain subdisciplines. As Figure 6 shows, the methods used in decision theory, philosophy of language, logic, and metaphysics have not changed much in the time period we investigated. But in all other subdisciplines, the relative use of methods from the probability family has increased. Such a change would, of course, have major implications for the character of the discipline of philosophy more broadly.

We remain uncertain as to whether the use of methods in the probability family will continue to increase. Clearly, it is possible that the use of methods in this family will remain at its present level or even decrease. Perhaps we would be able to arrive at more accurate predictions about the future trajectory using the sorts of computational approaches sketched above. Or perhaps we would be able to do so if we had a better understanding of the underlying causes of the changes that have already occurred.

A continuing trend toward increased use of probabilistic methods would, if it does oc-

¹⁴As expressed, for example, by Michael Dummett (1978, p. 458), who claimed that the goal of analytic philosophy was to analyze the structure of thought, as opposed to the psychological process of thinking, by applying logic to the analysis of language.

¹⁵Intriguingly, Malaterre, Chartier, and Pulizzotto (2019) provide evidence for a similar change in the philosophy of science in the 1980s. So, whatever it is that is happening in so many areas of philosophy over the past decade or so, it is possible that it is similar to what happened in philosophy of science around thirty to forty years ago.

cur, have normative implications for certain practices within the discipline of philosophy. To take just one example, consider philosophy graduate education in formal methods. One traditional approach to this required some number of logic courses but no courses in other formal methods, especially those in the probability family. This approach might well have been sensible when the vast majority of philosophy papers using formal methods were using methods from the logic family. However, it seems that one important role of graduate education in philosophy is to prepare students to be able to engage with research using the methods actually used within the discipline. If methods from the probability family increasingly find usage across many subdisciplines of philosophy, it might therefore make sense to reconsider these graduate education requirements.

Finally, on the fourth question: There are many disciplines across the sciences and in the digital humanities that use methods in the probability family more widely than methods in the logic family.¹⁶ Increasing use of formal methods in the probability family might therefore signal increasing methodological confluence with other parts of the humanities and sciences that use formal methods, as we alluded above. Such a confluence might signal a changing relationship between philosophy and other scholarly and scientific disciplines, in contrast with analytic philosophy’s reputation as standing apart from those disciplines. It might be possible to investigate this question through estimating the rates of citation of philosophy articles using formal methods in non-philosophy journals and through the rates of co-authorship in such articles, including rates of co-authorship between those primarily affiliated with a university or college philosophy department and those primarily affiliated with some other department or institution.

ACKNOWLEDGEMENTS

We would like to thank the Minnesota screening team, including Kelly Adams, Femke Kuiling, Devin Lightheart, Marie Lundgren, and Sonja Zbinden, for their indispensable work to create our data set, and two reviewers for comments that helped improve the clarity and accessibility of our conclusions.

REFERENCES

- Betti, A. and H. van den Berg (2014). Modelling the history of ideas. *British Journal for the History of Philosophy* 22(4), 812–835.
- Betti, A. and H. van den Berg (2016). Towards a computational history of ideas. In L. Wieneke, C. Jones, M. Düring, F. Armaselu, and R. Leboutte (Eds.), *CEUR Workshop Proceedings*, Volume 1681. Aachen: CEUR-WS.
- Betti, A., H. van den Berg, Y. Oortwijn, and C. Treijtel (2019). History of philosophy in ones and zeros. In M. Curtis and E. Fischer (Eds.), *Methodological Advances in Experimental Philosophy*, pp. 295–332. London: Bloomsbury.
- Bonino, G., P. Maffezioli, and P. Tripodi (2020). Logic in analytic philosophy: A quantitative analysis. *Synthese forthcoming*, 1–38.

¹⁶Methods from the digital humanities have only made limited inroads in philosophy. In §1, we discussed some examples that use topic modeling. See also Buckner, Niepert, and Allen (2011) for an approach to creating a computational “ontology” for philosophy—a kind of automated map of the structure and relations of concepts in a domain.

- Buckner, C., M. Niepert, and C. Allen (2011). From encyclopedia to ontology: Toward dynamic representation of the discipline of philosophy. *Synthese* 182(2), 205–233.
- DeAngelis, C. D., J. M. Drazen, F. A. Frizelle, C. Haug, J. Hoey, R. Horton, S. Kotzin, C. Laine, A. Marusic, A. J. P. Overbeke, T. V. Schroeder, H. C. Sox, and M. B. Van Der Weyden (2005, 06). Is This Clinical Trial Fully Registered? A Statement From the International Committee of Medical Journal Editors. *JAMA* 293(23), 2927–2929.
- DeAngelis, C. D., J. M. Drazen, F. A. Frizelle, C. Haug, J. Hoey, R. Horton, S. Kotzin, C. Laine, A. Marusic, A. J. P. M. Overbeke, T. V. Schroeder, H. C. Sox, and M. B. Van Der Weyden (2004, 09). Clinical Trial Registration: A Statement From the International Committee of Medical Journal Editors. *JAMA* 292(11), 1363–1364.
- Dummett, M. (1978). Can analytical philosophy be systematic, and ought it to be? In *Truth and Other Enigmas*, pp. 437–461. Cambridge, MA: Harvard University Press.
- Fischer, E. and M. Curtis (Eds.) (2019). *Methodological advances in experimental philosophy*. London: Bloomsbury.
- Hao, K. (2019, January). We analyzed 16,625 papers to figure out where AI is headed next. *MIT Technology Review*.
- Hosmer, Jr, D. W., S. Lemeshow, and R. X. Sturdivant (2013). *Applied Logistic Regression* (third ed.). Wiley Series in Probability and Statistics. Hoboken, NJ: Wiley.
- Krizhevsky, A., I. Sutskever, and G. E. Hinton (2012). Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems*, Volume 25. Curran Associates, Inc.
- LeCun, Y., Y. Bengio, and G. Hinton (2015). Deep learning. *Nature* 521, 436–444.
- Malaterre, C., J.-F. Chartier, and D. Pulizzotto (2019). What is this thing called philosophy of science? A computational topic-modeling perspective, 1934–2015. *HOPOS: The Journal of the International Society for the History of Philosophy of Science* 9(2), 215–249.
- Oaksford, M. and N. Chater (2007). *Bayesian Rationality*. Oxford: Oxford University Press.
- Sheskin, D. J. (2011). *Handbook of Parametric and Nonparametric Statistical Procedures* (fifth ed.). Boca Raton, FL: CRC Press.
- Shoukri, M. M. (2010). *Measures of Interobserver Agreement and Reliability* (second ed.). Chapman & Hall/CRC Biostatistics Series. Boca Raton, FL: CRC Press.
- Tenenbaum, J. B., C. Kemp, T. L. Griffiths, and N. D. Goodman (2011). How to grow a mind: Statistics, structure, and abstraction. *Science* 331, 1279–1285.
- Viera, A. J. and J. M. Garrett (2005, May). Understanding interobserver agreement: The kappa statistic. *Family Medicine* 37(5), 360–363.
- Wheeler, G. (2013). Models, models and models. *Metaphilosophy* 44(3), 293–300.