Review of Philosophy and Psychology
https://doi.org/10.1007/s13164-021-00571-4

Experimental Philosophy and the Incentivisation Challenge: a Proposed Application of the Bayesian Truth Serum

Philipp Schoenegger[1]

Accepted: 5 July 2021/Published online: 14 August 2021

Abstract

A key challenge in experimental social science research is the incentivisation of subjects such that they take the tasks presented to them seriously and answer honestly. If subject responses can be evaluated against an objective baseline, a standard way of incentivising participants is by rewarding them monetarily as a function of their performance. However, the subject area of experimental philosophy is such that this mode of incentivisation is not applicable as participant responses cannot easily be scored along a true-false spectrum by the experimenters. We claim that experimental philosophers' neglect of and claims of unimportance about incentivisation mechanisms in their surveys and experiments has plausibly led to poorer data quality and worse conclusions drawn overall, potentially threatening the research programme of experimental philosophy in the long run. As a solution to this, we propose the adoption of the Bayesian Truth Serum, an incentive-compatible mechanism used in economics and marketing, designed for eliciting honest responding in subjective data designs by rewarding participant answers that are surprisingly common. We argue that the Bayesian Truth Serum (i) adequately addresses the issue of incentive compatibility in subjective data research designs and (ii) that it should be applied to the vast majority of research in experimental philosophy. Further, we (iii) provide an empirical application of the method, demonstrating its qualified impact on the distribution of answers on a number of standard experimental philosophy items and outline guidance for researchers aiming to apply this mechanism in future research by specifying the additional costs and design steps involved.

✉ Philipp Schoenegger
ps234@st–andrews.ac.uk

[1] The University of St Andrews, St Andrews, UK

1 Experimental Philosophy and the Incentivisation Challenge

In this section, we motivate the main claim of this paper, namely that a lack of attention to incentivisation mechanisms has harmed experimental philosophical inquiry. In the following section, we propose the adoption of a mechanism that ought to allay those concerns and may prove central in future endeavours to make experimental philosophy a more robust research programme. Then, we provide an empirical application of this mechanism in the context of experimental philosophy and lastly outline some potential challenges to its implementation.[1]

The vast majority of experimental philosophy research employs a survey methodology, though there are some notable recent exceptions (e.g. Rubin, O'Connor, & Bruner 2019; Diaz 2019; Alsmith & Longo 2019) to this claim.[2] Generally speaking, participants (be it lay populations or expert philosophers) are often expected to evaluate thought experiments, cases of philosophical interest, or statements more broadly (cf. e.g. Nahmias et al. 2005). For example, participants may be asked to score the morality of an action on a Likert scale, indicate on a binary choice item whether an agent knows that p or not, or rank outcomes based on their plausibility. Underlying these tasks is the experimenters' aim to collect data from participants who take their time with the study, engage with it seriously, consider the cases carefully, and answer in an honest and considered manner. This is central to ensure the collection of data sufficiently relevant for philosophical theorising. As philosophy departments have not had a long history of recruiting students for their empirical studies, participants are, for better or for worse, frequently sourced from online platforms like MTurk or Prolific. They are then compensated for completing the survey and often only fail to receive payment in case they fail an attention check, i.e. a question that attempts to estimate whether participants pay sufficient attention to the instructions.

Several substantial methodological criticisms of experimental philosophy's modus operandi have been raised before, for example by Polonioli (2017), Stuart et al. (2019), and Woolfolk (2013). Specifically, Woolfolk (2013) has argued convincingly that one of the main challenges that experimental philosophy faces is the credibility of self-report questionnaires. They claim that because experimental philosophy does not (and in many instances cannot) embrace behavioural measures, self-report data are the only data available and thus we might not be able to rely on them as heavily as we do. As evidence for the worry about self-report, Woolfolk cites experimenter demand (the effect that participants answer in the way they think the experimenters want them to answer) and social desirability (answering according to some conception of a social norm or standard) (Woolfolk 2013, 80), as well as more general worries about validity and reliability.[3] Further, Cullen (2010) has argued that experimental philosophy's

[1] For helpful comments and suggestions, we thank Theron Pummer, Ben Grodeck, Raimund Pils, Emilia Wilson, and Enrico Galvagni. Further, we want to thank Tom Heyman for providing exceedingly helpful reviews.

[2] For the majority of this paper, we will continue to refer to this as the standard methodology in experimental philosophy. This is because while non-survey methods are becoming more common, surveys are still in the majority and, more importantly, the vast majority of claims made in this paper generalise to most other methods as well.

[3] Especially the more general question of reliability within experimental philosophy research is one that is both uniquely pressing but also outside of the scope of this paper.

results are distinctly sensitive to minimal changes in question design, further suggesting that a straightforward survey design may fail to capture what experimental philosophers truly wish to investigate. These methodological worries have been raised in addition to several philosophical concerns (e.g. Kauppinen 2007; Ludwig 2010; Horvath 2010) that have proffered criticism of the role of intuitions in philosophy and the value of experimental work to philosophy more generally.

In this paper, we want to present a further challenge to experimental philosophy that has thus far not been properly addressed. Specifically, we argue that experimental philosophy fails to take incentivisation concerns seriously. Further, we do not only aim to point out a methodological challenge for experimental philosophy; we also propose a potential solution that can readily be applied by researchers on a vast majority of research projects currently running under the label of experimental philosophy. In later sections, we also show its application in this context directly by reporting a study with this mechanism applied.

The core concept at work in the critique we want to highlight is incentive compatibility. Specifically, we claim that incentive compatibility has been crucially neglected by experimental philosophers so far and that this has had non-negligible effects on data quality. In a nutshell, an incentive compatible experimental design is one with an equilibrium in which each participant chooses the action that truthfully and fully reveals their preferences (Toulis et al. 2015). This may sometimes come in conflict with a participant's desire to maximise financial payoff in an experiment or other contravening preferences like impressing the experimenter. If these two actions are identical, a design is incentive compatible as the incentive to maximise payoff is compatible with the incentive to reveal their true preferences. In other words, if a participant can maximise their payoffs in an experiment by stating and acting according to their true preferences, a design is incentive compatible (Harrison 2006). Importantly then, the goal of any experimental mechanism designed to be incentive compatible is to provide participants with incentives such that they can respond honestly and reveal their true preferences in their choices in the experiment or survey while also maximising their payoff (or chances thereof).

Experiments that are designed without incentive compatibility in mind are such where participants' honest revelation of preferences and the actions that maximise their payoffs may sometimes come apart. This may then distort the results gathered as (some) participants will forgo honest and careful responses to questions if they can instead maximise their earnings (per hour) by answering quickly or giving responses that they think will result in higher monetary rewards. For example, as an experimenter, one may want participants to truthfully state their views on the topic of abortion on a number of questions relating to the situations in which it might be permissible. However, due to the nature of participant recruitment (say, on Amazon Mechanical Turk) and the social norms and financial circumstances present in this research environment, participants have no incentive to provide their honest and thorough responses if they know that any answer on a 5-point Likert scale is usually enough to be awarded compensation for as long as any potential attention check is passed. Any detailed vignette that lays out the moral arguments for or against a certain proposition is merely an obstacle to be overcome. The participants' action that maximises their payoff is answering the questions as quickly as possible, irrespective of their actual views on abortion or the details of the case presented. As such, the data generated by this process

do not properly capture the true preferences of the participant and may, as such, be a distorted representation that will further challenge the validity of any conclusions drawn from these data and any meaningful philosophical work that could later build directly on these results.

In the specific case of experimental philosophy, the incentivisation challenge arises especially out of a combination of two factors: (i) the nature of experimental philosophy studies and (ii) its typical participants. While these are not the only underlying reasons for the incentivisation challenge, they represent the core of the challenge for experimental philosophy. Let us now take those in turn more slowly.

(i) Broadly speaking, experimental philosophy focuses on lay intuitions about thought experiments or cases and statements more generally. Participants are expected to evaluate these vignettes (paragraphs describing a case or outlining an argument) with regard to philosophically relevant aspects, for example, whether an actor is morally responsible for an act when an element of luck is involved, whether someone knows that p in situations of deception, or whether someone has free will in a fully determined universe. These make up the vast majority of experimental work historically and still represent a sizable portion of current work: For example, all entries in an early collected volume relied on this methodology (Knobe and Nichols 2008). However, note that recently experimental philosophy has also begun adopting methods from different adjacent sciences (cf. Fischer and Curtis 2019). While the field is slowly moving to a more varied repertoire of methods, the questionnaire/survey style methodology remains dominant and will arguably retain its dominant position for the foreseeable future. Importantly, all of these research designs aim at collecting subjective data about participants' views, judgements, or intuitions. That is, the experimenters have no way of checking whether participants respond 'correctly' to such questions and are instead interested in (a change in) their judgements and intuitions or their relationship to other objects of interest. That is, the subject matter studied is inherently subjective.

(ii) Regarding study samples, experimental philosophy has followed (social) psychology's focus on undergraduate student populations as well as online samples. The former's strong unrepresentativeness and the resultant challenges for external validity based on this sampling bias have been debated widely in psychology (cf. e.g. Gray and Keeney 2015; Nielsen et al. 2017). Online samples, while at least possibly allaying some worries about a distinct skew towards sampling from young, relatively privileged, Anglospheric undergraduate populations (cf. Hauser and Schwarz 2016), pose further challenges that exacerbate issues related to the incentive structure of online studies. If participants fully anonymously take studies online from their own devices at their own time, their primary incentive is to minimise time spent on the studies to maximise payoffs by increasing the number of studies taken, potentially concurrently. Further, due to the volume of tasks 'workers' on Amazon Mechanical Turk (MTurk) complete, they are relatively adept at identifying and passing basic attention checks as these will have been passed numerous times before (Chmielewski and Kucker 2020). This means that their action aimed at maximising payoff is to complete as many studies as possible while putting in as little effort as possible. However, their honest preference revealing action would warrant more careful reading and

> consideration of the questions posed as some level of reflection is necessary to even respond to the types of questions posed, for a cursory reading of a vignette on moral responsibility is insufficient for a meaningful response on issues like those studied by experimental philosophers.

Considerations of (ii) pose a central challenge to experimental philosophy as drawing out philosophical findings from disinterested online samples rushing to completion by providing minimally sufficient answers is a non-trivial feat. However, paired with (i), the fact that the answers sought by experimental philosophers are subjective judgements or intuitions, most solutions employed by other fields to address these challenges cannot be easily implemented for experimental philosophy.[4] That is, responses concerning objective answers might be easier to incentivise, e.g. by rewarding correct answers, leading to a re-alignment of the actions aimed an honest revelation of their preferences and the actions aimed at maximising payoffs, as participants can simply maximise their payoffs by attempting to answer questions as correctly as they can. However, focusing on intuitions and judgements makes it even harder to incentivise truth-telling as experimenters cannot reward participants based on how accurate their responses are due to the nature of the questions asked.

One natural response to this line of reasoning might be to blame this on the survey methodology and argue for a move towards different type of methods like behavioural laboratory experiments. However, while this may allay some of the problems raised about the survey approach (cf. Cullen 2010; Woolfolk 2013), moving away from the current standard would lead to many of the same challenges while also resulting in a method that may not be suited to answering many of the types of questions experimental philosophers are typically interested in. Even so, it would still face the incentivisation challenge because even in behavioural or behavioural-adjacent designs, one cannot rely on an easy incentivisation mechanism based on the truth of participants' answers as the focus of experimental philosophy remains on subjective data that are hard to incentivise; after all, intuitions cannot simply be graded as true/false by an experimenter, nor can their value be assessed against the background of majority rule. If most of experimental philosophy research then cannot be properly incentivised, the whole field faces the incentivisation challenge and one might want to be even more sceptical of the findings presented in the literature so far than one might have previously been.

In simple terms, the main challenge facing experimental philosophy, as we see it, is that participants are recruited for subjective evaluation tasks in such a way that their participation incentives (receiving payment) do not align with truthful and complete revelation of their honest views (carefully considering cases and answering honestly), which in turn threatens the quality of data used in experimental philosophy.

This incentivisation challenge has not received much attention in the literature so far. Importantly enough, this challenge has also been largely neglected by the discipline that experimental philosophy has drawn most heavily from: psychology. As Hertwig

[4] For the remainder of the paper, we will understand intuitions and judgements more generally as subjective and will further contrast them with objective answers exist, for example answers to arithmetic questions. By 'subjective' we do not mean to make any claim about the nature of philosophical concepts. Rather, we simply understand by that that experts and researchers have not reached a consensus as to, for example, what knowledge is or how to identify the good. As such, answers cannot be evaluated against some standard that would then be used to determine financial payoff. Answers of this type are what we have in mind when we use the term 'subjective.'

and Ortmann (2001) outline, the experimental practises of psychology are substantially different than those of neighbouring economics, even though both share a focus on experimental data on a number of overlapping areas like the heuristics and biases literature or pro-social behaviour more generally. Relevantly to the purposes of this paper, economics has a strong disciplinary norm that forbids deceiving participants as well as a long tradition in offering substantial monetary performance incentives. Hertwig and Ortmann (2001) claim that psychology's lack of these strict practices might explain some of the variability in findings in the psychological literature, which may extend to and explain disparate outcomes on failures of replication. Irrespective of the truth of this comparative claim, we argue that experimental philosophy need not inherit disciplinary norms from one discipline only and ought to draw more freely from other disciplines, like economics.

This has not happened yet, and a proper engagement with the incentivisation challenge has not been present in the literature so far. For example, despite offering a thorough methodological critique of experimental philosophy, Woolfolk does not mention the issue of incentivisation even once, despite proposing that experimental philosophy should adopt the practices and methods of the "biobehavioral sciences" (Woolfolk 2013, 86). Further, Pölzler (forthcoming) points to problems of insufficient effort responding, though only mentions compensation, not incentivisation, which are related but distinct notions. Others have, however shortly, picked up the issue: For example, Hassoun (2016) argues that even though some "economists […] worry that we need to give subjects some incentive to reveal their true views", one might justifiably remain sceptical that economic incentives are needed in the context of experimental philosophy. They claim that "we can sometimes be fairly confident in people's reports about what they believe" (Hassoun 2016, 237). We suggest that no sufficiently strong reason has been offered for this view and that, prima facie, we ought not to be so confident, especially in the light of current participant recruitment practises and their corresponding incentive structures outlined above. Thus, experimental philosophy should follow the best practises of disciplines that have grappled with similar issues and constraints: We claim that experimental philosophy ought to address the incentivisation challenge head-on in pursuit of better data and stronger conclusions drawn as opposed to ignoring it and simply assuming that it does not matter to the overall research programme of employing empirical data in philosophy.

2 Incentive Compatible Mechanism Design and the Bayesian Truth Serum

There are a number of ways that researchers in several disciplines have attempted to design incentive-compatible mechanisms that reward participants for answering honestly in cases in which participants respond with regard to subjective data (i.e. their opinions, intuitions, judgements, predictions, etc). For example, some mechanisms reward participants by how common their answers are in the full sample gathered. In other words, participants are rewarded by choosing answers that are most often chosen by other participants too. This, of course, introduces a number of additional challenges and might incentivise the wrong type of responses. These challenges would be especially difficult for experimental philosophers to resolve as participants' actions

aimed at maximising payoffs might not be aligned with the actions that would reveal their true preferences; after all, participants might simply attempt to guess what most people would answer as opposed to providing their own honest thoughts and judgements which may be different than the majority's. Other mechanisms like the peer prediction method (Miller et al. 2005) improve upon this by introducing proper scoring rules, but even methods like these or the related Delphi method (Linstone and Turoff 1975), a further variation on the prediction approach, still favour the consensus answer to some extent and are plausibly inadmissible for our current goals as participants' actions revealing their true preferences may come in conflict with their actions aimed at maximising payoff when payoff can be maximised by guessing the most popular option.

A promising improvement upon these types of mechanisms is the Bayesian Truth Serum (BTS), introduced by Drazen Prelec (2004). The Bayesian Truth Serum is a survey scoring method developed and already applied in marketing research and economics. It is designed to reward honesty of responders in survey designs where the subject matter is subjective and cannot be compared against an objective baseline. It has been used in predicting new product adoption (Howie et al. 2011), in perceptual deterrence studies in criminological research (Loughran et al. 2014), in eliciting expert beliefs about future energy prices (Zhou et al. 2019), and in estimating the prevalence of questionable research practices in psychology (John et al. 2012). In all of these cases, participants could not be standardly incentivised to provide true answers due to the nature of the questions being subjective or not straightforwardly verifiable in a relevant time frame. However, when it comes to new products or expert beliefs, it is crucial for researchers to elicit honest answers, just as it would be for experimental philosophers concerned with thought experiments (cf. also Prelec et al. 2017) as the answers might have direct impact, for example on future public policy in the case of expert beliefs on energy prices or on philosophical theorising in the case of findings in experimental philosophy.

The Bayesian Truth Serum (BTS), as introduced by Prelec (2004) aims to close this gap and provides an incentive-compatible mechanism aimed at eliciting honest responses that would be applicable to the domain of subjective questions and answers. In essence, the Bayesian Truth Serum works by rewarding answers that are surprisingly common monetarily. The underlying reasoning for rewarding surprisingly common answers is based on Bayesian claims about population frequency. Specifically, about the fact that one's own view of a topic ought to be underestimated by other agents. This is because one takes one's own view as an informative sample of one, which in turn leads one to overestimate the true frequency of one's view, leading others to underestimate one's own view. Put simply, the "highest predictions of the frequency of a given opinion […] in the population should come from individuals who hold that opinion" (Prelec 2004, 462).

This mechanism underlying the Bayesian Truth Serum builds on two assumptions. First, the sample size gathered has to be large enough such that a single respondent cannot tactically answer questions in such a way that their prediction of the empirical distribution meaningfully impacts the distribution of all responses.[5] The second

[5] While formally the result is stated with a countably infinite population (Prelec 2004, 463), it also generalises to large finite samples and, as we shall later see, even smaller samples under certain conditions.

assumption is that all respondents share a common prior and that this belief is then updated according to Bayes' rule. Accordingly, personal opinions and views are thus treated as "impersonally informative signals" (Prelec 2004, 463) concerning the population distribution. In other words, a respondents' view, opinion, or intuition is treated as providing evidence about the population distribution in an impersonal way such that other respondents who answer identically will also draw the same inference from that.

The important takeaway from this model is that one would expect individuals to overestimate the population frequency of their own view because their own view on the matter is directly taken into account by themselves. In other words, the Bayesian Truth Serum draws on the notion that individuals believe that their own personal beliefs are disproportionally held amongst others because they themselves have that view. Conversely then, individuals ought to think that their own beliefs are underestimated by the wider population. The crucial point here is that one's honest view is very likely also that view which is surprisingly common because one should conclude that the "true popularity of [one's view] is underestimate[d] by the [wider] population. Hence, one's true opinion is also the opinion that has the best chance of being surprisingly common" (Prelec 2004, 462). Rewarding participants based on this 'surprisingly common' criterion thereby merges the participants' honest preference revealing and payoff maximising actions, ensuring incentive-compatibility. Being monetarily rewarded based on this criterion now means that participants are no longer incentivised to guess the majority view (as they were under a majority incentivisation method) or to complete the survey in as little time as possible (as they were in the case of no specific incentivisation mechanism). Rather, answering honestly now maximises payoffs.[6] Importantly, note that in actual implementations of the mechanism (as opposed the theoretical statements in idealised settings) participants are not informed of the specifics of the mechanism but are instead merely told that an algorithm scores their answers according to honesty and that they are rewarded based on this. We will pick this up in more detail later.

In order to arrive at a measure of surprisingly common answers, the BTS works by having participants answer all questions of the survey or experiment as they would in a standardly constructed study. Participants are also asked to provide a prediction as to the underlying distribution of answers on every item. The BTS then assigns high scores to answers that are more frequently endorsed than collectively predicted (i.e. surprisingly common) and rewards participants based on this. The score capturing the notion of a "surprisingly common" response (Prelec 2004, 462) is the information score, or i-score, according to which participants are directly rewarded. Calculating the information score of any given answer requires the calculation of two terms: First, the relative frequency of that answer, and second, the geometric mean of that answer's predicted frequency. The former is simply calculated by drawing on the actual distribution of answers and taking the frequency of any given answer. The latter is importantly different from the arithmetic mean and is calculated by taking the n^{th} root of the product of all predicted answers (1…n). The i-score is then calculated by log $\left(\frac{Relative\ Frequency\ of\ Answer}{Geometric\ Mean\ of\ Answer's\ Predicted\ Frequency}\right)$ and participants are directly rewarded based

[6] Specifically, truth telling maximises expected payoff under the assumption that everyone is responding truthfully: truth-telling is a Bayesian Nash equilibrium.

on their information score and their prediction accuracy (i.e. how well they predict others' responses).[7] Importantly though, participants are not informed of the specifics of this mechanism and are only told that answering honestly will improve their payoff as determined by this score.

Consider the following example as an illustration of how this mechanism of assigning an information score to surprisingly common answers works in practice. In our highly idealised example there are five participants, call them A, B, C, D, and E. They are asked which is the better philosopher, John Locke or Mary Astell. We assume that A and D honestly believe John Locke to be the better philosopher while the others honestly prefer Mary Astell. In addition to responding to this question, they are also asked to provide an estimation of how the distribution of others in this sample will look like. Below you find a table with their hypothetical responses.[8]

Let us now discuss the case of participant A in more detail. A responds to the main question asking who they believe the best philosopher to be with 'John Locke.' Asked how many of the participants will overall believe John Locke to be the best, A responds '65%,' conversely assigning '35%' to Mary Astell. A treats his endorsement of Locke as informative about the underlying distribution in forming this estimation and thus has an estimation that is higher than that of others who prefer Mary Astell or the overall aggregate estimation. In aiming to calculate A's information score, we first need to calculate the geometric mean of the endorsement predictions for 'John Locke' in the full sample. The geometric is calculated by taking the n^{th} root of the product of n terms. In our example above, this would be $\sqrt[5]{65*30*20*55*30} = 36.45$. We can now calculate the information score of A's choice of 'John Locke;' Recall that the information score of an answer is calculated by $\log\left(\frac{Relative\ Frequency\ of\ Answer}{Geometric\ Mean\ of\ Answer's\ Predicted\ Frequency}\right)$. In our illustration, two people picked 'John Locke' as the best philosopher of all time (making the relative frequency of an answer 40%), and the geometric mean of this answers predicted frequency was 36.45 As such, $\log\left(\frac{40}{36.45}\right) = .04$. Being greater than '0', this makes the answer surprisingly common, which would be rewarded financially (participant's total i-scores across answers are summed up and added to a prediction accuracy score to determine the final pay-out), meaning that A has an incentive to state their honest view despite being in the minority. Note though that the same also applies to B, whose belief is in the majority. For B's endorsement of Astell has an information score of $\log\left(\frac{60}{57.29}\right) = .02$. As such, both those in the minority and in the majority have incentive to state their true views.

Recall that one of the Bayesian insights that this mechanism heavily draws on was that participants treat their own view as an informative sample of one. This claim, necessary for the mathematical statement of the incentive-compatibility and the theoretical strength of the mechanism, is borne out empirically in the false consensus effect that shows that people do overestimate the consensus for their own views or behaviours both when they are in the majority and when they are in the minority (cf. Ross et al.

[7] In this paper, we focus specifically on the information score as this is the main contribution of the Bayesian Truth Serum.

[8] Note however that this is a very stylised example with low sample and made-up answers purely presented for illustration purposes.

1977; Mullen et al. 1985; Choi and Cha 2019). This effect has been studied extensively in hypothetical choice scenarios as well as behavioural contexts like authentic conflict situations (Ross et al. 1977), and a multitude of explanations have been raised. For example, some argue that selective exposure and cognitive availability play a role in producing this effect, while others prefer motivational explanations (cf. Marks and Miller 1987; Sherman et al. 1983). This effect has also been directly replicated cross-culturally by Liebrand et al. (1986) and is still widely found in variety of contexts long after it has first been studied in the 1980s (e.g. Coleman 2018; Welborn et al. 2017). The Bayesian Truth Serum relies on a Bayesian interpretation of this phenomenon, i.e. that individuals treat their own views as informative. This Bayesian interpretation for the false consensus effect had been floated before, for example by Dawes (1989, 1990) or Krueger and Zeiger (1993) as well as others. However, Prelec's Bayesian Truth Serum is the first direct application of this Bayesian insight which directly motivates the scoring method of the information scores that reward participants whose answers are unexpectedly common.

Crucially, the Bayesian Truth Serum has also received some direct empirical validation above and beyond numerous direct applications in peer-reviewed research and the mathematical proofs published previously. Most prominently, Frank et al. (2017) directly validated the BTS mechanism in large-scale online human experiments relevantly similar to those likely to be employed by experimental philosophers. Over a series of studies with a total sample size of 8765 participants, they find that the instructions relating to the BTS mechanism reliably increase honest responses in sessions where the true distribution is known (e.g. a binomial distribution in a study where participants have to report the outcomes of coin flips), while also significantly impacting the distribution in studies where the underlying distribution cannot be known. Further, the BTS was also shown to improve honest responses in other settings, such as a recognition questionnaire (Weaver and Prelec 2013), while also outperforming other competing mechanisms like the solemn oath in other designs (cf. e.g. de-Magistris and Pascucci 2014), in which participants swear an oath to answer truthfully. Importantly, in this validation study and in further applications, participants are not directly informed about the mechanism. That is, participants do not have to understand this mechanism and are not facing additional incentives aimed at potentially outsmarting the mechanism. They are generally only told that an algorithm aimed at rewarding honesty is being applied to their study and that a certain fraction of them (those ranking highest on this metric) will be rewarded monetarily. As such, any effect is, in essence, an effect based on the instructions given, not technically of the mechanism.

We suggest that based on the arguments laid out in the first part of the paper and the empirical validation provided by others, the Bayesian Truth Serum might directly address and at least partially alleviate the incentivisation challenge for experimental philosophy by providing an incentive-compatible mechanism aimed at rewarding honest responses in subjective data collection. Recall that we argued that the main challenge is one in which the sample recruited for experimental philosophy studies have to answer subjective questions that elude any easy incentivisation. With the Bayesian Truth Serum, participants are told that their honesty will be rewarded financially and are thus facing an incentive structure in which they maximise their payoff by reporting their honest views as, per Bayesian assumption, those are more

likely to be surprisingly common and will be rewarded most, thus bringing their honest preference revealing and payoff maximising actions in line with each other. This means that participants will plausibly no longer simply aim to complete as many studies as possible in as short a time as is feasible without being denied payment, but rather they will aim to answer as honestly as possible in order to receive the highest possible payment allocated by the Bayesian Truth Serum.

3 Empirical Application of the Bayesian Truth Serum in Experimental Philosophy

In order to make the adoption of this mechanism for experimental philosophy further plausible above and beyond the theoretical reasoning and previous validation studies outlined above, we present a direct empirical application of the Bayesian Truth Serum in the context of experimental philosophy. While this cannot be a validation in the strict sense of the word,[9] we take this study to provide evidence for the ability of the instructions used alongside the Bayesian Truth Serum to meaningfully impact participant behaviour and, coupled with the theoretical upsides outlined above, ought to strengthen the case in favour of a widespread adoption of this mechanism in experimental philosophy. Further, while there have been several direct validations (Frank et al. 2017) and applications in experimental research (Howie et al. 2011; John et al. 2012; Loughran et al. 2014; Zhou et al. 2019; Barrage and Lee 2010; Weiss 2009) we believe showing its impact in the same context that we propose it should be applied in is important. In this empirical application, we thus employ the Bayesian Truth Serum in a setting standardly used in experimental philosophy.

3.1 Methods

We recruited 425 participants via Prolific to participate in our study. 23 of these participants failed the attention check (asking them explicitly to indicate 'Strongly Agree' on a Likert scale) and were excluded prior to analysis. The remaining 402 participants (73.9% female, 25.9% male) had an average age of M = 36.17 (SD = 11.73). Participants were randomly selected into the Control (n = 191) or the BTS treatment condition (n = 211) at the beginning of the experiment.[10]

Participants in the Control group received a standard reminder that they ought to read the following cases and consider them carefully before giving their answers. Those in the BTS condition received two paragraphs adopted from Frank et al. (2017), informing the participants that their honesty is being evaluated by an algorithm designed to test for this and that those who score in the top third according to this metric will receive a bonus of £1 (Fig. 1).[11] Those in the BTS condition were also asked to provide their predictions as to the underlying distribution of answers in this sample to calculate the information scores. All participants were paid £0.75 for their

[9] This is because unlike previous validation studies who tested honesty in cases where the underlying distribution was known (coin flips and dice rolls) this is not possible in the context of experimental philosophy, as outlined in the earlier sections of this paper.

[10] Ethics approval has been received under SA15351.

[11] Participants' final score was determined by their total i-scores and their total prediction accuracy.

Table 1 Illustration

Participants	A	B	C	D	E
Responses					
John Locke	X			X	
Mary Astell		X	X		X
Predictions					
John Locke	65%	30%	20%	55%	30%
Mary Astell	35%	70%	80%	45%	70%

Illustrative example of the BTS mechanism

participation and the £1 bonus was paid out in addition to this to those scoring in the top third of information scores in the BTS condition.

To test the Bayesian Truth Serum in the context of experimental philosophy, we selected a variety of vignettes from seven articles that had been published in either The Review of Philosophy and Psychology or Philosophical Psychology within the last ten years. Specifically, we included vignettes on attributions of knowledge-how in conditions of luck (Carter et al. 2019) – example 1, modesty (Weaver et al. 2017) – example 2, freedom of choice in situations of nudging (Hagman et al. 2015) – example 3, the moral permissibility of torture (Spino and Cummins 2014) – example 4, the correspondence theory of truth (Barnard and Ulatowski 2013) – example 5, moral responsibility (De Brigard and Brady 2013) – example 6, and determinism (Nadelhoffer et al. 2020) – example 7. Likert scales were used for all seven examples (six used a 7-point Likert scale ranging from 1 = "Strongly disagree" to 7 = "Strongly agree" while one used a 5-point Likert scale ranging from 1 = "Strongly disagree" to 5 = "Strongly agree"). See Appendix 2 for a full list of all seven examples. All examples were presented in a random order to participants.

3.2 Results

We conducted pairwise comparisons of the distributions of answers using Pearson's χ^2 goodness-of-fit tests. We coded the answer distribution in the Control condition as the expected distribution and compared the BTS treatment distribution of answers to it. We find that in four out of the seven cases, the distribution of answers in the BTS treatment differed significantly from the distribution in the Control at $p < .001$.[12] See the Appendix 1, Tables 3 and 4 for full descriptive data on participant responses and predictions.

Importantly, the direction of the effects are not identical across examples. For instance, the BTS condition shows an increase of answers in the middle of the scale compared to the Control, especially '4 – Neither agree nor disagree' in example 1, but the reverse is true for example 5 or 7, see Fig. 2. Further, it is important to note that this test does not treat the responses as an interval, i.e. significant changes may not be capturing an increase or decrease in agreement with the statements overall, but rather a mere change in response pattern which we believe to be the adequate test for this

[12] Because we have set the criterion of significance at '<.05' we will treat the results for example 2 of $p = .05$ as not significant.

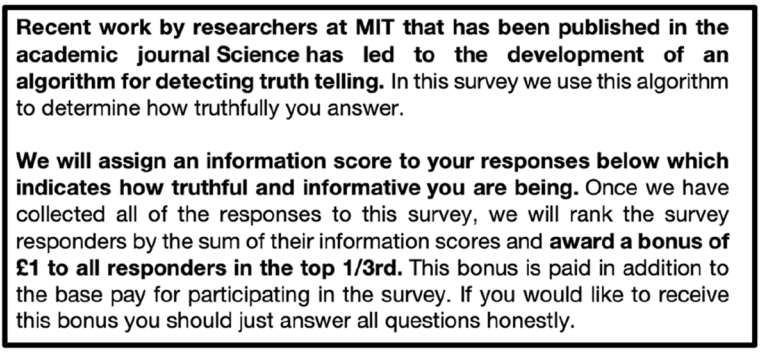

Fig. 1 BTS Treatment. Notes: Message displayed to participants in the Bayesian Truth Serum treatment, informing them about the reward mechanism

situation as no directionality of honesty can be established a priori. A two-sample Kolmogorov-Smirnov test, designed to test the same question for continuous data, does not show any significant effects on all seven examples, see Appendix 1, Table 5 for further detailed test statistics. We also report a Chi-squared test for independence. The results of this test show a significant change in responses in two out of the seven items (specifically example 1 and example 5 on knowledge-how and the correspondence theory of truth), see Appendix 1, Table 6 for full test statistics.[13]

Overall, participants in the treatment condition ($M = 13.61$, $SD = 7.80$) spent significantly more time (in minutes) in the experiment than those in the control condition ($M = 8.26$, $SD = 5.53$), $t(400) = -7.859$, $p < .001$, $d = .79$, leading to participants in the Control having a higher average pay at £5.45 per hour compared to the BTS condition with £4.76 per hour, even factoring in the additional bonuses. This difference in time is best explained by the fact that those in the BTS condition had to provide predictions on each item, while those in the Control did not.

3.3 Discussion

Our main finding in this empirical application is that the instructions of the Bayesian Truth Serum mechanism influence responses in a number of standard experimental philosophy items. Four out of seven examples showed strong and statistically significant changes in response distributions as determined by a χ^2 goodness-of-fit test. Importantly though, this effect shows substantial heterogeneity, with some examples suggesting a move towards the centre (example 1) and others a move towards the poles (examples 5 and 7). However, almost half the examples show no significant change at all. Note further that because this research is between-subject, this is not to be taken as informative about individual changes in behaviour. Additionally, because we fail to observe significant differences on the two-sample Kolmogorov-Smirnov test and a limited set of significant differences on the test for independence, we should also be cautious generally to make claims of an increase or decrease in agreement across conditions or of a large-scale robust effect prior to additional replications. Nonetheless, we take these results to establish empirically that the instructions relating to the

[13] We thank a reviewer for recommending to also include this test.

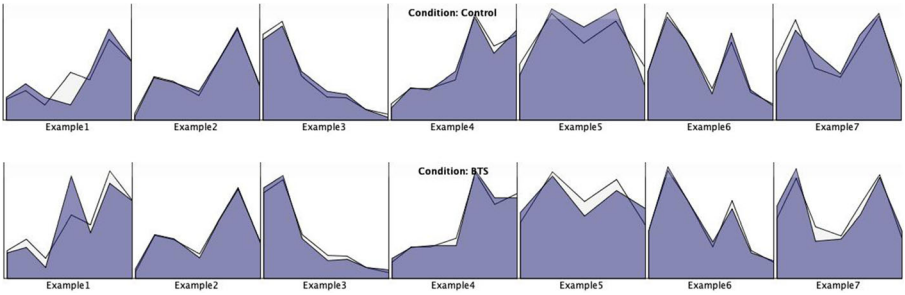

Fig. 2 Differences in Distributions. Notes: The black distribution outline is the distribution of answers merged between both conditions. Coloured top panel distribution shows distribution of Control. Coloured bottom panel distribution shows distribution of BTS treatment

Bayesian Truth Serum may have an impact on participant behaviour in the standard setting of an experimental philosophy study in some of the examples studies here.

One may conclude from the examples above (Table 2), that there exists a difference between the content present in the examples that show a difference in distribution compared to the Control and those that do not. For example, the vignettes on moral responsibility, moral permissibility, or virtue do not show a statistically significant effect, perhaps contrary to expectation, while those on knowledge and determinism do. However, because this study was not designed to detect a disparate effect on the philosophical content, we caution against drawing any conclusions from this; it may simply be that factors not directly controlled for mediate this effect. As such, specifically directed research that holds constant the length of the vignettes, difficulty, and other potential confounds would be needed to establish a potential disparate effect of the mechanism on different philosophical topics, and no conclusions about this ought to be drawn from our data.

Further research above and beyond direct replications may also attempt to more intentionally disentangle the effect of the Bayesian Truth Serum from that of a higher bonus or of the prediction task generally. Regarding the effect of the bonus, given that those in the BTS condition actually received a lower pay per hour, even including the bonusses on the aggregate level, one might hold that those worries might not be of too great a concern. Further, Frank et al. (2017) already investigated this question and found that "honesty was not increased as a result of increased participant pay in the

Table 2 Differences of distributions

	χ^2	Exact Sig. (2-tailed)
Example 1 (knowledge-how)	199.089	.001****
Example 2 (virtue)	12.577	.050*
Example 3 (freedom of choice)	28.667	.001****
Example 4 (moral permissibility)	8.367	.212
Example 5 (correspondence theory of truth)	49.933	.001****
Example 6 (moral responsibility)	5.758	.451
Example 7 (determinism)	24.697	.001****

Pearson's χ^2 goodness-of-fit statistics for pairwise distribution comparisons coding the Control condition as the expected distribution and comparing the BTS to it. *$p < .1$, **$p < .05$, ***$p < .01$, ****$p < .001$

control treatment to match the pay of participants in BTS treatments in expectation" (Frank et al. 2017, 10), suggesting that this might generalise and that this worry ought not be too concerning for experimental philosophers.

In order to give readers a concrete number regarding the additional payments needed to implement this mechanism, we will run through the expenses of this application in detail in the hope that it may generalise to future research and prove helpful for interested readers. For simplicity's sake, however, we will assume that the exclusions were applied evenly across conditions and that there were exactly 200 participants in each treatment. Using Prolific in early 2021, the total cost of paying 200 participants £0.75 for participation adds up to £210 including platform fees. The same base cost applies to the BTS treatment, though paying out £1 to 67 participants (a third of all participants) increases the cost by £93.80 (including the bonus charge applied by the platform when paying out bonuses) to £303.80. Of course, base payment as well as bonus payment may vary, though in the example shown here and in future studies using a similar proportion of base pay to bonus, one can expect an increase in research costs of about 30% (30.88% in this case specifically).

There is an additional caveat regarding the interpretation of these findings that is worth highlighting: This empirical showcase cannot technically demonstrate that the Bayesian Truth Serum changes response patterns. Rather, what is shown here is that the instructions, see Fig. 1, have this documented effect. The Bayesian Truth Serum is, in this context, just a post-hoc mechanism for determining the distribution of bonus payments. As such, it is more accurate to claim that this paper has demonstrated the effect of the accompanying treatment text, as opposed to the Bayesian Truth Serum.

However, the claim that it is the Bayesian Truth Serum that has this effect may turn out to be true in future settings if this mechanism is adapted widely and understood by a significant portion of participants. If all (or most) researchers use this mechanism, and it is generally understood by participants that they do indeed maximise their payoff by answering honestly, then one may hold that it is the Bayesian Truth Serum that has this effect. One challenge in situations of such widespread uptake would then be that some researchers might prefer to use only the instructions but not actually reward participants according to this mechanism; after all, calculating and paying the additional rewards is a non-trivial effort. As such, it is crucial for a community of researchers (within and across disciplinary boundaries) to agree to using this mechanism to avoid these potentially hazardous side-effects of individual researchers defecting to lower their own research costs while also potentially lowering the efficacy of the instructions in turn.[14]

Overall, we take these results to add to the literature showing that the instructions given to participants that inform them of the payment mechanism of the Bayesian Truth Serum can impact answers in the context of experimental philosophy, where no such empirical investigation had been conducted thus far.

4 Objections and Implementation Concerns

We claim that the paper so far has provided two interrelated reasons for the adoption of the Bayesian Truth Serum in experimental philosophy. First, it directly addresses the

[14] We kindly thank a reviewer for pressing us on this important point.

incentivisation challenge by offering an incentive-compatible mechanism where none was used before to the plausible detriment of data quality. Second, we have shown empirically that implementing this mechanism meaningfully changes responses to standard examples of experimental philosophy. In this last section we will discuss the implementation of the Bayesian Truth Serum in experimental philosophy more broadly and address some concerns to provide a fuller picture of our proposal that experimental philosophers ought to address the incentivisation challenge by adopting the Bayesian Truth Serum.

As we have argued above, adopting the Bayesian Truth Serum would meet the incentivisation challenge head on as the BTS is one of the few incentive compatible mechanisms that could theoretically be adopted given the type of data that experimental philosophers are interested in collecting. Experimental philosophy has closely modelled itself after psychology in the past. For example, when the replication crisis hit psychology, experimental philosophers also adopted the Open Science best practices and replicated numerous studies (cf. e.g. Cova et al. 2021; Schönegger and Wagner 2019, Kim and Yuan 2015; Seyedsayamdost 2015) in attempt to respond to it. However, there is no reason why experimental philosophy could not adopt this incentivisation mechanism and draw on the literature on incentivisation more generally like is done in the field of economics. We hope that the arguments outlined in the first part of the paper have provided ample reason to adopt this mechanism. Having presented the empirical application of the method above, we argue that the fact that we have shown that answers differ significantly ought to further provide evidence that this mechanism, at least in its current form and wording, is worth implementing, as the theoretical upsides of incentivising honesty at least plausibly pay off as advertised theoretically. It is important to again point out that the above application is not (and indeed cannot) be a strict validation study as conducted before in different contexts, as one cannot directionally check if people answer more honestly in the context of experimental philosophy. Additionally, as discussed above, what has really been tested is whether the treatment text that was given to participants had an impact on responses. However, adding this suggestive picture to the clear theoretical upsides does, in our view, provide a strong position for those wishing to adopt the Bayesian Truth Serum in experimental philosophy, especially given the relative lack of alternatives in addressing the incentivisation challenge.

We now want to turn our focus to some questions of implementation that have not been addressed above. Specifically, implementing the Bayesian Truth Serum would necessitate the following changes for researchers in experimental philosophy: First, researchers would have to modify their research to ask participants to state not only their straightforward answers, but also to provide predictions as to the frequency of answers. Second, researchers would have to provide an explanatory statement (like the one presented above and adopted from previous studies) to inform participants that they will be rewarded (at least in part) based on their honesty as measured by the information score. Third, researchers will have to acquire the funds to pay out these additional payments according to this criterion, which might increase the costs of research.

First, in order for researchers to properly conduct the calculations needed to reward surprisingly common answers, they have to expand their questionnaires by including prediction items to each question. That is, after giving participants a vignette and a question as to the content, researchers further have to include an additional question

asking the participants about a prediction of other participants' responses on this very question. Specifically, participants have to state their predicted distribution of answers on each item. This change is minor, yet an increase in completion time and as such participant payment is to be expected, especially when pay per hour has a lower bound on many online platforms. Further, as the number of questions presented to the participants will have substantially increased, researchers might find it beneficial to lower the overall number of items to ensure the surveys and experiments finish at a reasonable rate and in a manageable time without risking widespread drop-out of participants or fatigue effects. For an example of how such an item asking participants for their prediction of the frequencies would look like for Example 1, see the Appendix 3.

Second, researchers will have to explain to participants that they are being rewarded by this mechanism. We suggest that simply adopting a message like the one used in the empirical application (Fig. 1) suffices for the time being, as it has been shown both in this context and others (e.g. Frank et al. 2017) that this message alone elicits a change in behaviour from participants. There is no real difficulty in implementing this and one might not need to consider further trade-offs at this point in time. However, as adoption increases and participants become more aware of this mechanism, a change in messages may be appropriate. Importantly, when presenting participants with this message, the researchers need not explain the reward mechanism of information scores to participants – participants are merely rewarded based on it and informed of the overall reward system, but not the details of it. In other words, what matters in the application of this mechanism is not its Bayesian assumptions or the intricacies of how the information score is calculated, but simply that participants believe that responding truthfully maximised their payoff (Loughran et al. 2014, 687). That this is the case has been shown extensively in different contexts (Howie et al. 2011; John et al. 2012; Loughran et al. 2014; Zhou et al. 2019), and our empirical application presented in this paper further makes this adoption plausible directly in the context of experimental philosophy.

Third, researchers will have to change the way that they reward participants. Not only will participants spend more time in the study (due to asking them to predict other's answers), but researchers will also have to reward participants based on their information score. This will take some additional effort and time, but we argue that this cost is worth it if one can employ an incentive-compatible mechanism that plausibly improves data quality. As our estimations above, this might be best anticipated as an 30% increase in costs (if the proportions between base pay and bonus pay are similar like in our empirical application). Further, we claim that contra those suggesting that one might simply reward participants a bonus without actually calculating the information score, i.e. deceiving participants about the nature of the incentivisation mechanism, might have long-term effects on the participant pool, above and beyond research ethical concerns, relating to their level of trust in the experimenters and their motivation to guess the experimenter's intentions. For an overview of the empirical literature on deception in experimental research, cf. Hertwig and Ortmann (2001). This is also a potential solution to the worry outlined above, that when most have adopted this mechanism, some individuals may prefer to free-ride and not actually pay out the bonuses. We believe that introducing and fostering norms of honesty, transparency, and accountability might go a long way towards a better and more open science, while also offering some protection against this worry.

Lastly, we want to address some further objections and worries about our proposal. First, it is important to point out that the standard BTS is not incentive compatible for small samples as per the assumption of large samples above, which may be thought of as a challenge to the general implementability as experimental philosophy has sometimes relied on smaller samples. As a response to this more generally worry, Witkowski and Parkes (2012) introduced the robust Bayesian Truth Serum (RBTS) aimed at providing an incentive compatible mechanism for small samples, i.e. sample size $n \geq 3$. Importantly, the RBTS is only applicable to research designs that elicit binary information. As with the standard Bayesian Truth Serum, the RBTS collects the same types of information from respondents, i.e. their information responses and their predictions about the underlying distribution of answers. Their contribution is inducing a "shadow posterior" (Witkowski and Parkes 2012, 2) by coupling an agent i's information report with a prediction report of another agent, j, where the prediction of agent j is adjusted towards agent i's information report. The payment is determined by a binary quadratic scoring rule (Witkowski and Parkes 2012, 5). For a proof that RBTS is strictly Bayes-Nash incentive compatible, see Witkowski and Parkes (2012). The main upside of RBTS is that it allows experimental designs to ensure that truth-telling is incentive compatible even in small sample sizes, for as long as the information input is binary.[15] However, we claim that these solutions might not be necessary if experimental philosophy continues on its path of increasing the sample sizes, as doing so has numerous additional methodological upsides and enabling the application of the standard BTS would only add to these. Additionally, because the main effect is driven by the instructions given to the participants, this may be thought of as a secondary concern.

Second, one might see the above implementation worries as one of a trade-off between scientific standards and cost, as implementing the BTS increases participant time, participant payment, as well as researcher time. However, we claim that not incentivising research because it would be cheaper is clearly not a sufficiently strong reason to outweigh the scientific upside of properly aligning honest preference revealing actions with payoff maximising actions, especially if no other, cheaper alternatives aimed at addressing the incentivisation challenge are currently available or in use. Strengthening the scientific basis of experimental philosophy ought to be our foremost concern as we move towards establishing this subfield in the longer term, far outweighing more financial and practical worries, even if that would result in less studies being conducted and published in the short term. In other words, we claim that while our proposal would bring with it additional costs of time and money, doing so would be worth it as we gain additional evidentiary strength in our research programme overall.

Lastly, one may question the Bayesian assumption of the BTS mechanism. Specifically, the criticism could be either about the descriptive claim that individuals do, in fact, give significantly higher estimates about population rates of a view when they

[15] Under two further assumptions, Radanovic and Faltings, 2013proposes a more general RBTS mechanism that extends also to non-binary signals in small samples: First, the self-dominant assumption states that any agent believes that any observation of a given value, such as one's response to a question, is most likely also to be observed by another agent. Second, the self-predicting assumption more weakly states that any agent believes that observing a certain value means that another agent is most likely to observe that same value. Under those, Radanovic and Faltings, 2013shows, the RBTS can be applied more generally.

themselves hold that view, or about the normative claim that rational agents ought to treat their own view as an informative sample of n = 1. The latter type of objection would practically be denying the assumption and would as such be detrimental, though we see no reason to deny this basic Bayesian claim. A further objection might be the claim that people do not, in fact, act according to this Bayesian rule. That is, one might object that BTS relies on a descriptively wrong picture of human action. If this objection were to go through, this might be more problematic for our proposal. In response to this objection, we can again draw on the psychological literature on the false census effect and show that the effect exists empirically. This effect, sometimes labelled a bias, describes the tendency to report one's own views or actions as being more common than they actually are (Ross et al. 1977; Mullen et al. 1985; Choi and Cha 2019). This gives the assumption made by the BTS some prima facie justification and may be sufficient to defend against this objection. However, even if this response is not sufficient, recall that both the present empirical application and previous implementations and validations have shown significant changes in response distributions, suggesting that the mechanism might work even if some of the assumptions might not hold true fully. Though here it is again important to point out that what these studies have ultimately shown is that the instructions themselves have this effect, not technically the mechanism itself.

Overall, we believe that the Bayesian Truth Serum, first introduced by Prelec (2004), provides a plausible solution to the incentivisation challenge raised against experimental philosophy. We have argued that its application is not unproblematic and its implementation not without trade-offs. However, these challenges can be overcome and introducing an incentivisation mechanism to experimental philosophy that can be applied widely is worth the increased costs of time and money, especially as no competitor approach has yet been proposed, which leaves experimental philosophy open to the incentivisation challenge outlined at the beginning. It may be that in the future, other incentive-compatible mechanisms will be proffered that also address the incentivisation challenge in a better way, and we would very much welcome this. However, until such a development, we believe that the Bayesian Truth Serum is the only plausible candidate and ought to adopted by experimental philosophers, irrespective of the additional costs incurred by researchers.

5 Conclusion

In this paper, we have argued that experimental philosophy is facing and has thus far failed to engaged with the incentivisation challenge, according to which failing to provide an incentive compatible mechanism harms the quality of empirical research in philosophy. As a solution, we have suggested that researchers ought to implement the Bayesian Truth Serum mechanism, an incentive compatible payoff mechanism that rewards participants based on how surprisingly common their answers are – drawing on the Bayesian insight that the highest prediction of a frequency of any given question ought to come from those who hold that view. We then went on to present an empirical application of this mechanism' in the context of experimental philosophy, showing how this mechanism's instructions can change the distributions of answers across numerous standardly studied questions in experimental philosophy. We have closed

with a discussion of some of the implementation challenges but concluded that implementing the Bayesian Truth Serum would improve scientific practice in experimental philosophy and leave the sub-discipline in a stronger position going forward, and that the costs of this proposal are negligible in comparison.

Appendix 1 – Supplementary Statistics

Table 3 Response frequencies

Examples	(1)	(2)	(3)	(4)	(5)[a]	(6)	(7)
Control (n = 191)							
Frequency							
1 – Strongly disagree	.079	.010	.277	.031	.115	.126	.099
2	.126	.147	.325	.084	.288	.267	.194
3	.079	.131	.168	.079	.241	.204	.147
4 - Neither	.052	.099	.099	.126	.288	.068	.099
5	.162	.204	.089	.267	.068	.225	.183
6	.314	.319	.037	.173	–	.079	.230
7 – Strongly agree	.188	.089	.005	.241	–	.031	.047
Skewness	−.672	−.449	.886	−.609	−.006	.347	−.101
Kurtosis	−.903	−1.042	−.027	−.466	−.961	−1.007	−1.318
BTS (n = 211)							
Frequency							
1 – Strongly disagree	.066	.033	.313	.052	.171	.123	.156
2	.081	.152	.355	.081	.265	.289	.237
3	.028	.137	.137	.085	.161	.204	.081
4 – Neither	.265	.071	.062	.085	.227	.095	.085
5	.118	.199	.066	.280	.175	.180	.137
6	.246	.308	.038	.209	–	.066	.218
7 – Strongly agree	.194	.100	.028	.209	–	.043	.085
Skewness	−.588	−.440	1.289	−.704	.063	.504	.050
Kurtosis	−.535	−1.089	.981	−.401	−1.286	−.756	−1.489

Frequency of answers endorsed, skewness, and kurtosis of full answer distributions

[a] Example 5 had a 5-point Likert scale, ranging from 1 = "Strongly disagree" to 5 = "Strongly agree

Table 4 Predicted frequencies

Examples	(1)	(2)	(3)	(4)	(5)[a]	(6)	(7)
Geometric Mean							
Predicted frequency							
1 – Strongly disagree	.202	.149	.085	.201	.194	.096	.129
2	.183	.237	.105	.204	.200	.120	.147
3	.142	.177	.116	.176	.206	.146	.133
4 – Neither	.223	.138	.162	.134	.196	.171	.165
5	.127	.136	.16	.119	.163	.179	.145
6	.113	.149	.217	.113	–	.191	.179
7 – Strongly agree	.107	.096	.205	.095	–	.134	.155

Geometric means of predicted frequencies. As geometric means cannot be computed when at least one entry is '0', we removed all frequency predictions of '0,' which is why the sum of all predicted frequencies does not equal to 1

[a] Example 5 had a 5-point Likert scale, ranging from 1 = "Strongly disagree" to 5 = "Strongly agree

Table 5 Differences of distributions (K-S)

	D	Asymptotic Sig. (2-tailed)
Example 1 (knowledge-how)	.107	.198
Example 2 (virtue)	.034	1.000
Example 3 (freedom of choice)	.066	.773
Example 4 (moral permissibility)	.032	1.000
Example 5 (correspondence theory of truth)	.107	.199
Example 6 (moral responsibility)	.046	.984
Example 7 (determinism)	.100	.267

Independent samples Kolmogorov-Smirnov goodness of fit test between the control and the BTS treatment. *$p < .1$, **$p < .05$, ***$p < .01$, ****$p < .001$

Table 6 Chi-square test for independence

	χ^2	Asymptotic Sig. (2-tailed)
Example 1 (knowledge-how)	37.785	.001****
Example 2 (virtue)	3.484	.746
Example 3 (freedom of choice)	6.877	.332
Example 4 (moral permissibility)	3.843	.698
Example 5 (correspondence theory of truth)	16.229	.003***
Example 6 (moral responsibility)	2.607	.856
Example 7 (determinism)	11.067	.086*

Chi-Square Test for Independence between the control and the BTS treatment. *$p < .1$, **$p < .05$, ***$p < .01$, ****$p < .001$

Appendix 2 – Vignettes

EXAMPLE 1. Charlie needs to lern how to change a lightbulb, and so he goes to the 'how-to' section in his local library. He finds a shelf full of identical looking books titled Home Repair. In each of those books are step-by-step instructions on the way to change a lightbulb-we'll call the way the book describes way 'w'. Unbeknownst to Charlie, all the copies of Home Repair on the shelf are fakes, except for one. Pranksters have placed these copies there, and these fake copies contain mistaken step-by-step instructions on the way to change a lightbulb. Since Charlie does not known this, he reaches up and grabs the copy of Home Repair nearest to him. By sheers luck, he selects the only copy in the entire library that contains genuine and reliable step-by-step instructions for charging a lightbulb. Had Charlie picked up any of the other guides-which he so easily could have-he would have believed the mistaken instructions were correct.
Do you agree with the following statement? 'Charlie knows how to change a lightbulb.'

EXAMPLE 2. Jamie is a very good rollerblader, but she doesn't known that she is very good at rollerblading. When praised for her rollerblading, she says, "Thank you! But I don't think I'm very good at rollerblading."
Do you agree with the folowing statement? 'When Jamie said, ""Thank you! [...]." she was being modest.'

EXAMPLE 3. Many countries have a problem with their citizens not paying taxes, which costs society a considerable amount of money. Some countries have therefore started to send out information to the taxpayers with the encouraging message "To pay your taxes is the right thing to do". The idea with this intervention is to give tax evaders a bad conscience and therefore increase their motivation to pay their taxes.
To what extent do you think that the described policy restricts the individual's freedom of choice?

EXAMPLE 4. A bomb has been planted in a crowded section of a major city and a terrorist is currently in custody. Beyond a reasonable doubt, this terrorist has the information required for successfully defusing the bomb. If the bomb explodes, it is estimated that thousands of people will die.
If torture is successfully employed against the terrorist, he will provide the information needed to defuse the bomb in time. All of the available evidence from previous situations like this indicates that torture has a high probability of making the terrorist provide the needed information to defuse the bomb in time. Alternative investigational and interrogative methods other than tortures still remain as options. These methods have a low probability of success.
Give the situation described above, please indicate which rating best describes your judgement of the following statement: 'It is morally permissible to torture the terrorist.'

EXAMPLE 5. Bruno has just finished painting his house. Bruno painted his house the same color as the sky on a clear summer day. Bruno claims his house is blue.
With respect to the case of Bruno, how much do you agree or disagree with the following statement? "If a claim reports how the world is, then it is true.'

EXAMPLE 6. Mary is the single mother of two: Mark, 7, Sally, 4. Mary works most of the day, and although she is known for being fairly patients and good natured, over the last year she has exhibited some unusually aggressive behavior toward her neighbor. Last week, when she came back from work late at night, she couldn't drive into her garage because her neighbor had blocked her driveway with his new BMW Enraged, she stepped on the gas pedal and crashed her car into her neighbor's. Unfortunately, her neighbor was still inside the car (it was too dark for anyone to see him), and both his legs were seriously broken in several places. Now he is not only suing her for several thousand dollars, but he's also pressing charges. However, a neurologist examined her brain and discovered that, in the last year, Mary has been developing a rare tumor in her frontal lobe. Since the frontal lobe is necessary for emotional suppression-that is, the capacity

to control one's emotions-the neurologist claim that, unlike a healthy person, Mary was completely unable to control her rage and her desire to smash the car. "In fact", he says, "any person with this kind of tumor", facing the exact same situation, would have done exactly what Mary did. She didn't have done otherwise. "If Mary is found responsible for her actions, she may be sent to a federal medical facility for the next 6 months". There she could receive medical treatment, but she won't be able to seen her children. Unfortunately, during that that time, they would be living with Social Services, in what might be a much worse environment for them.

Do you agree with the following statement? 'Mary is normally responsible for crashing her car into her neighbor's.'

EXAMPLE 7. Imagine Jim lives in a causally closed universe. In this universe, given the physical state of the universe, the laws of the universe, and the fixity of the past, at any given moment the universe is closed, like a train moving down the tracks. Whenever Jim makes a decision to act in a particular way, it's always the case that he could have acted differently only if something leading up to decision had been different. In short, at any given moment, there is one and only one choice and action genuinely open to Jim. Moreover, if you knew absolutely everything about both the history of the universe and about Jim, you could always know in advance what Jim is going to decide to do. He is not the only deciding factor when it comes to what he does. Given the way the world was long before Jim was born, everything in his life is in the cards, so to speak. Jim can make choices, but these choices are the only choices open to him. Now, for illustrative purposes, imagine that Jim decides to take his dog for a walk in the park.

Do you agree with the following statement? 'Jim's choices and decisions make a difference in what he does.'

Appendix 3 – Additional Question (Example 1)

Do you agree with the following statement? 'Charlie knows how to change a lightbulb.'

	Strongly agree	Agree	Somewhat agree	Neither agree nor disagree	Somewhat disagree	Disagree	Strongly disagree
Enter your answer here.	○	◉	○	○	○	○	○

Out of the following options, which percentage of other participants in this study, do you think, chose which option as the answer to the above question?

Remember that your estimates have to sum to 100%.

Strongly agree	15
Agree	25
Somewhat agree	12
Neither agree nor disagree	24
Somewhat disagree	21
Disagree	2
Strongly disagree	1
Total	100

References

Alsmith, A.J.T., & Longo, M.R. 2019. Using VR Technologies to Investigate the Flexibility of Human Self-Conception. In E. Fischer & M. Curtis (Eds.) *Methodological advances in experimental philosophy* (pp. 153–174). Bloomsbury Publishing.

Barnard, R., and J. Ulatowski. 2013. Truth, correspondence, and gender. *Review of Philosophy and Psychology* 4 (4): 621–638.

Barrage, L., and M.S. Lee. 2010. A penny for your thoughts: Inducing truth-telling in stated preference elicitation. *Economics Letters* 106 (2): 140–142.

Carter, J.A., D. Pritchard, and J. Shepherd. 2019. Knowledge-how, understanding-why and epistemic luck: An experimental study. *Review of Philosophy and Psychology* 10 (4): 701–734.

Chmielewski, M., and S.C. Kucker. 2020. An MTurk crisis? Shifts in data quality and the impact on study results. *Social Psychological and Personality Science* 11 (4): 464–473.

Choi, I., and O. Cha. 2019. Cross-cultural examination of the false consensus effect. *Frontiers in Psychology*: 1–13.

Coleman, M.D. 2018. Emotion and the false consensus effect. *Current Psychology* 37 (1): 58–64.

Cova, F., Strickland, B., Abatista, A., Allard, A., Andow, J., Attie, M., ... & Zhou, X. 2021. Estimating the reproducibility of experimental philosophy. *Review of Philosophy and Psychology, 12*(1), 9–44.

Cullen, S. 2010. Survey-driven romanticism. *Review of Philosophy and Psychology* 1 (2): 275–296.

Dawes, R.M. 1989. Statistical criteria for establishing a truly false consensus effect. *Journal of Experimental Social Psychology* 25 (1): 1–17.

Dawes, R.M. 1990. The potential Nonfalisity of the false Consesus effect. In *Insights in decision making: A tribute to Hillel J Einhorn*, ed. R. Hogarth, H.J. Einhorn, and R.M. Hogarth. University of Chicago Press.

De Brigard, F., and W.J. Brady. 2013. The eEfect of what we think may happen on our judgments of responsibility. *Review of Philosophy and Psychology* 4 (2): 259–269.

de-Magistris, T., and S. Pascucci. 2014. The effect of the solemn oath script in hypothetical choice experiment survey: A pilot study. *Economics Letters* 123 (2): 252–255.

Diaz, R. 2019. Using fMRI in Eperimental Philosophy: Epxloring the Prospects. In E. Fischer & M. Curtis (Eds.) *Methodological advances in experimental philosophy* (pp. 131–152). Bloomsbury Publishing.

Fischer, E., & Curtis, M. (Eds.). (2019). Methodological advances in experimental philosophy. Bloomsbury Publishing.

Frank, M.R., M. Cebrian, G. Pickard, and I. Rahwan. 2017. Validating Bayesian truth serum in large-scale online human experiments. *PLoS One* 12 (5): e0177385.

Gray, K., and J.E. Keeney. 2015. Impure or just weird? Scenario sampling Bias raises questions about the Foundation of Morality. *Social Psychological and Personality Science* 6 (8): 859–868.

Hagman, W., D. Andersson, D. Västfjäll, and G. Tinghög. 2015. Public views on policies involving nudges. *Review of Philosophy and Psychology* 6 (3): 439–453.

Harrison, G. W. (2006). Making choice studies incentive compatible. In Valuing environmental amenities using stated choice studies (pp. 67–110). Springer, Dordrecht.

Hassoun, N. (2016). Experimental or empirical political philosophy. In Sytsma J. & Buckwalter W. (Eds). A companion to experimental philosophy, 234-246.

Hauser, D.J., and N. Schwarz. 2016. Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods* 48 (1): 400–407.

Hertwig, R., and A. Ortmann. 2001. Experimental practices in economics: A methodological challenge for psychologists? *Behavioral and Brain Sciences* 24 (3): 383–403.

Horvath, J. 2010. How (not) to react to experimental philosophy. *Philosophical Psychology* 23 (4): 447–480.

Howie, P.J., Y. Wang, and J. Tsai. 2011. Predicting new product adoption using Bayesian truth serum. *Journal of Medical Marketing* 11 (1): 6–16.

John, L.K., G. Loewenstein, and D. Prelec. 2012. Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science* 23 (5): 524–532.

Kauppinen, A. 2007. The rise and fall of experimental philosophy. *Philosophical Explorations* 10 (2): 95–118.

Kim, M., & Yuan, Y. (2015). No Ccross-Cultural Differences in the Gettier Car Case Intuition: A Replication Study of Weinberg et al. 2001.

Knobe, J., and S. Nichols. 2008. *Experimental philosophy*. Oxford: Oxford University Press.

Krueger, J., and J.S. Zeiger. 1993. Social categorization and the truly false consensus effect. *Journal of Personality and Social Psychology* 65 (4): 670–680.

Liebrand, W.B., D.M. Messick, and F.J. Wolters. 1986. Why we are fairer than others: A cross-cultural replication and extension. *Journal of Experimental Social Psychology* 22 (6): 590–604.

Linstone, H.A., and M. Turoff, eds. 1975. *The Delphi method*, 3–12. Reading, MA: Addison-Wesley.

Loughran, T.A., R. Paternoster, and K.J. Thomas. 2014. Incentivizing responses to self-report questions in perceptual deterrence studies: An investigation of the validity of deterrence theory using Bayesian truth serum. *Journal of Quantitative Criminology* 30 (4): 677–707.

Ludwig, K. 2010. Intuitions and relativity. *Philosophical Psychology* 23 (4): 427–445.

Marks, G., and N. Miller. 1987. Ten years of research on the false-consensus effect: An empirical and theoretical Review. *Psychological Bulletin* 102 (1): 72–90.

Miller, N., P. Resnick, and R. Zeckhauser. 2005. Eliciting informative feedback: The peer-prediction method. *Management Science* 51 (9): 1359–1373.

Mullen, B., J.L. Atkins, D.S. Champion, C. Edwards, D. Hardy, J.E. Story, and M. Vanderklok. 1985. The false consensus effect: A meta-analysis of 115 hypothesis tests. *Journal of Experimental Social Psychology* 21 (3): 262–283.

Nadelhoffer, T., S. Yin, and R. Graves. 2020. Folk intuitions and the conditional ability to do otherwise. *Philosophical Psychology* 33 (7): 968–996.

Nahmias, E., S. Morris, T. Nadelhoffer, and J. Turner. 2005. Surveying freedom: Folk intuitions about free will and moral responsibility. *Philosophical Psychology* 18 (5): 561–584.

Nielsen, M., D. Haun, J. Kärtner, and C.H. Legare. 2017. The persistent sampling Bias in Developmental Psychology: A call to action. *Journal of Experimental Child Psychology* 162: 31–38.

Polonioli, A. 2017. New issues for new methods: Ethical and editorial challenges for an experimental philosophy. *Science and Engineering Ethics* 23 (4): 1009–1034.

Pölzler, T. (forthcoming). Insufficient effort responding in experimental philosophy. In Lombrozo, T., Knobe, J., & Nichols, S. (Eds.), Oxford studies in experimental philosophy, volume 4. Oxford: Oxford University Press.

Prelec, D. 2004. A Bayesian truth serum for subjective data. *Science* 306 (5695): 462–466.

Prelec, D., H.S. Seung, and J. McCoy. 2017. A solution to the single-question crowd wisdom problem. *Nature* 541 (7638): 532–535.

Radanovic, G., & Faltings, B. (2013). A robust Bayesian truth serum for non-binary signals. In proceedings of the 27th AAAI conference on artificial intelligence (AAAI'13) (no. CONF, pp. 833-839).

Ross, L., D. Greene, and P. House. 1977. The "false Consensus effect": An egocentric Bias in social perception and attribution processes. *Journal of Experimental Social Psychology* 13 (3): 279–301.

Rubin, H., O'Connor, C., & Bruner, J. 2019. Experimental economics for philosophers. In E. Fischer & m. Curtis (Eds.) Methodological advances in experimental philosophy (pp. 175–206). Bloomsbury Publishing.

Schönegger, P., and J. Wagner. 2019. The moral behavior of ethics professors: A replication-extension in German-speaking countries. *Philosophical Psychology* 32 (4): 532–559.

Seyedsayamdost, H. 2015. On gender and philosophical intuition: Failure of replication and other negative results. *Philosophical Psychology* 28 (5): 642–673.

Sherman, S.J., C.C. Presson, L. Chassin, E. Corty, and R. Olshavsky. 1983. The false consensus effect in estimates of smoking prevalence: Underlying mechanisms. *Personality and Social Psychology Bulletin* 9 (2): 197–207.

Spino, J., and D.D. Cummins. 2014. The ticking time bomb: When the use of torture is and is not endorsed. *Review of Philosophy and Psychology* 5 (4): 543–563.

Stuart, M.T., D. Colaço, and E. Machery. 2019. P-curving X-phi: Does experimental philosophy have evidential value? *Analysis* 79 (4): 669–684.

Toulis, P., Parkes, D. C., Pfeffer, E., & Zou, J. (2015, June). Incentive-compatible experimental design. In Proceedings of the sixteenth ACM conference on economics and computation (pp. 285-302).

Weaver, R., and D. Prelec. 2013. Creating truth-telling incentives with the Bayesian truth serum. *Journal of Marketing Research* 50 (3): 289–302.

Weaver, S., M. Doucet, and J. Turri. 2017. It's What's on the Inside that Counts... Or is It? Virtue and the Psychological Criteria of Modesty. *Review of Philosophy and Psychology* 8 (3): 653–669.

Weiss, R. R. J. (2009). Optimally aggregating elicited expertise: A proposed application of the Bayesian truth serum for policy analysis (Doctoral dissertation, Massachusetts Institute of Technology).

Welborn, B.L., B.C. Gunter, I.S. Vezich, and M.D. Lieberman. 2017. Neural correlates of the false consensus effect: Evidence for motivated projection and regulatory restraint. *Journal of Cognitive Neuroscience* 29 (4): 708–717.

Witkowski, J., & Parkes, D. C. (2012). A robust Bayesian truth serum for small populations. In proceedings of the 26th AAAI conference on artificial intelligence (AAAI'12).

Woolfolk, R.L. 2013. Experimental philosophy: A methodological critique. *Metaphilosophy* 44 (1–2): 79–87.

Zhou, F., L. Page, R.K. Perrons, Z. Zheng, and S. Washington. 2019. Long-term forecasts for Fnergy commodities Price: What the experts think. *Energy Economics* 84: 104484.