An Epistemic Principle of Charity in Informal Argument Evaluation

Jens Koed Madsen (j.madsen2@lse.ac.uk)
Department of Psychological and Behavioural Science,
London School of Economics and Political Science, Houghton Street, WC2A 2AE, London, UK

Mike Oaksford (mike.oaksford@bbk.ac.uk)
Department of Psychological Sciences,
Birkbeck, University of London, Malet Street, Bloomsbury, London WC1E 7HX

Sayeh Yousefi (s.yousefi@lse.ac.uk), Nicole George (nicolelgeorge@outlook.com), Cassandra Teigen (teigencassandra@gmail.com)
Department of Psychological and Behavioural Science,
London School of Economics and Political Science, Houghton Street, WC2A 2AE, London, UK

Abstract

In this paper, we explore the Principle of Charity. This is an epistemic assumption that people should not judge people to be irrational unless they have an empirically justified account of what they are doing when they violate normative standards. Through two studies, we provide evidence in support of the principle. Study 1 suggests people believe others will arrive at the same conclusions they would themselves given the same information. Study 2 suggests that people assume others may differ in the subjective degrees of belief but that they broadly use the same (Bayesian) updating mechanism when evaluating information about other people. We believe this paper provides the first empirical test of this principle.

Keywords: Reasoning; Principle of Charity; Bayesian Argumentation; Epistemology

Introduction

Quine (1969, see also Thagard & Nisbett, 1983) described a 'principle of charity' (a phrase originally coined by Wilson, 1959)[1]. Philosophers have different versions of the principle, but we take our point of departure from Quine's use. According to the principle, people should avoid attributing irrationality to others' statements, and thus, people should not interpret the utterances of others as contradictory or absurd *unless* there is evidence to assume otherwise (Quine, 1969). For example, if a person says, 'Napoleon betrayed his fellow animals', it is reasonable to assume 'Napoleon' denotes the chief pig in Orwell's *Animal Farm* rather than the French emperor. The latter would entail absurdities, whereas the former is plausible and relevant. In line with the Principle of Charity, the listener should, by default, assume the former.

The principle further extends to reasoning. Given the same evidence, people should assume others would use evidence in the same way as they would unless given reason to believe otherwise. Of course, this is possible sometimes. An expert statistician might reasonably assume that they would make more qualified inferences from empirical data compared to a layperson. Further, it may be reasonable to assume people hold different views that may influence how they process information, such as disagreeing on which information sources are credible. The Principle of Charity underpins how people should interact with others. It suggests that we should believe that other people are equally capable to reason and think. However, it has never been empirically tested despite its centrality to language comprehension and reasoning.

Thagard and Nisbett (1983) describe different degrees of commitment to the principle. Their third version states that people "should not judge people to be irrational unless you have an empirically justified account of what they are doing when they violate normative standards" (p. 252). This version of the principle has interesting implications for reasoning and argumentation theory. Following this, people should not treat the beliefs and attitudes of others as irrational unless there is evidence to suggest that they are. Along similar lines, Dennett (1998) argues for a principle of humanity where others will have "the propositional attitudes one supposes one would have oneself in those circumstances" (p. 343). Further, Jara-Ettinger et al. (2016) denote "naive utility calculus". i.e. the notion that people reason about others' behavior and internal mental states by implicitly assuming that agents choose goals and actions to maximize the rewards they expect to obtain. Finally, Davidson (1974) focused on peoples' subjective probabilities in trying to understand degrees of beliefs. According to these principles, people should assume that other people would update their beliefs in the same way they would themselves unless there is explicit reason to believe otherwise.

This expectation relates directly to Bayesian argumentation (Hahn & Oaksford, 2006; 2007). Bayesian reasoning operates with *degrees of belief* between 0 (highly uncertain) to 1 (certainty). People integrate their prior beliefs in a hypothesis (H) with the likelihood, Pr(e|H), that is, the likelihood of the evidence given the hypothesis. This integration yields the posterior degree of belief in the hypothesis *given* the evidence (Pr(H|e)). The principle of charity supplemented with a Bayesian approach, therefore,

[1] Davidson (1974) make use of a similar principle of charity when calling for a principle of rational accommodation in which interpretations of the utterances of others should be in a way that optimises agreement.

implies that no person should believe that 3rd persons entertain different prior beliefs and likelihoods unless given evidence to assume otherwise (as mentioned above, there may be ample reason to assume this in some real-life contexts). Technically, this interpretation and Dennett's principle of humanity extend the original principle of charity, which states that people assume that others share the same rational norms for updating their beliefs (Bayes' Theorem), not that others share the same beliefs.

In Bayesian argumentation tasks, participants are usually presented with a dialogue containing an argument structure, some evidence, and different interlocutors who provide differentially reliable sources of information. People respond with a rating of convincingness expressing a posterior degree of belief. This method has been used from a first-person perspective. For example, participants are asked, "In light of the above dialogue, how convinced are *you* now of the conclusion?" (Harris et al., 2013; Expt. 2 in Hahn et al., 2009). More commonly, the third-person perspectives are used in Bayesian studies. For example, participants are asked, "In light of the above dialogue, how convinced *should Anne* now be of the conclusion?" (Corner et al., 2011; Harris et al., 2012; Oaksford & Hahn, 2004).

In this paper, we test whether people follow the principle of charity in the context of argumentation by comparing people's evaluations of their own beliefs on encountering an argument and the beliefs they believe others should have after hearing the same argument. Without evidence that the other person has different reasoning capabilities or information, the principle of charity predicts that the two posterior responses should be the same. In Study 1, we test the principle in cases where participants have no reason to believe the other person differs from them. In Study 2, we test if people who may reasonably believe others have different subjective beliefs still assume people update in line with Bayesian principles. To our knowledge, there is no previous study that empirically investigates whether people conform to the principle of charity in argumentation or reasoning.

Study 1: Showing similarities

Bayesian argumentation studies suggest that people are sensitive to the likelihood of the information, even in logically fallacious arguments such as arguments from ignorance (Oaksford & Hahn, 2004), slippery slopes (Corner et al., 2011), the ad Hitlerum (Harris et al., 2012), and the ad hominem (Oaksford & Hahn, 2013). Study 1 makes use of these argument structures and employs strong and weak versions arguments. (e.g., a weak argument from ignorance refers to a single study that has failed to find negative consequences of a proposed policy; a strong version refers to 50 such studies). Given past findings, we hypothesise that strong versions of fallacious arguments will be more persuasive than weak ones. We do not expect any effect of argument strength on the conclusions drawn in the 1st or a 3rd person condition. Nonetheless, having two types of each argument structure provides a broader and stronger ground for testing whether people follow our version of the principle of charity.

A subsidiary goal of this experiment was to investigate two factors hypothesised to influence the perceived strength of an argument. Argument strength can be manipulated by changing the credibility of the source of an argument (Bovens and Hartmann, 2003). For example, a layperson in the street or a scientific expert. In the Bayesian approach, source credibility has been regarded as an amalgamation of epistemic expertise and trustworthiness (Hahn et al., 2009). This is in line with social psychological classifications of reliability along warmth and competence lines (see e.g. Fiske et al., 2007). The Bayesian approach to source credibility has been tested empirically on single-report arguments (Harris et al., 2015; Madsen, 2016) as well as on multiple reports for one hypothesis (Madsen et al., 2020).

Here, we explore the single-report argument. In this approach, participants can be provided with a claim (e.g., that a made-up medical product would cure a particular ailment) advocated by a source identified as a friend or an enemy (trustworthiness) and as a doctor or musician (epistemic expertise). We hypothesise that claims from an expert and trustworthy source (E_+/T_+) are more convincing than claims from a mixed source (E_+/T_- or E_-/T_+), which are more convincing than claims from an untrustworthy and inexpert source (E_-/T_-). Again, we did not expect any effect of argument strength on our hypotheses (conclusions from a 1st or a 3rd person perspective), but the distinction provides a broader and stronger test of the hypotheses.

The main purpose of Study 1 is to test the hypothesis that we should observe *no* difference between the conclusions from a 1st or a 3rd person perspective. Standard t-tests can only provide indicative rather than confirmatory evidence, as these may not *reject* but simultaneously not *prove* the null. Rouder and colleagues argue that Bayes Factor analyses can be used to test for similarities and thereby circumvent the problem of null-hypothesis significance testing (Rouder et al., 2009; Morey & Rouder, 2011, see also Kruschke, 2011). That is, "the relative evidence measure B is known as the Bayes factor" (Morey & Rouder, 2011, p. 408; see also Kass & Raftery, 1995), which indicates "…the relative strength of evidence for two theories" (Dienes, 2014)[2]

Rather than testing a point hypothesis, interval null hypothesis testing allows for slight deviations across a Gaussian distribution by determining an effect size in standard deviations and testing the likelihood of the evidence falling within this boundary. In Morey and Rouder's terminology, the epistemic entailment following the principle of charity is a null hypothesis rather than a nil hypothesis. A nil hypothesis refers to "…the point hypothesis that the parameter is identically 0." A null hypothesis is more general,

[2] On a more general point, Kruschke (2013) provides evidence to support the claim that Bayesian estimation *generally* out-performs standardised t-tests.

such that "…it may be restricted to a nil hypothesis or may allow for values that deviate slightly from the nil" (both quotes, Morey & Rouder, 2011, p. 406). As with confidence intervals, this allows predictions of the null to deviate within an a priori defined and delineated effect size[3]. We use this statistical approach to calculate the likelihood of the null hypothesis being true when comparing the posterior ratings of convincingness. Thus, we test the principle of charity using standard t-tests as well as via a Bayes factor with an equivalence region. This approach allows for a positive test of the likelihood of the principle rather than simply not rejecting it given a non-significant t-test (see results section).

Method

Participants: 250 participants recruited from Mechanical Turk. 14 dropped out before completion, 2 provided the wrong validation code at the end of the experiment, and 2 did not fill out the entire experiment. These were excluded from the analyses, leaving 232 participants.

Design: First, participants read and evaluated the six argument fallacies. This was a 2 (perspective: 1st vs. 3rd person) × 2 (argument strength: Strong vs. Weak) between-subjects design. Subsequently, participants read single reports from more or less credible sources. The design was a 2 × 2 × 2 mixed design, with trustworthiness (High vs. Low) and expertise (High vs. Low) as within-subjects factors and perspective (1st vs. 3rd person) as a between-subjects factor. Given the exclusions above, in the first part of the experiment, there were 60 participants in the weak/3rd-person condition, 54 in the strong/3rd-person condition, 56 for the weak/1st-person condition, and 62 for the strong/1st-person) condition. In the second part of the experiment, all participants stayed in the perspective group they were in the first part of the experiment but then responded to all four reports with the trustworthiness and expertise manipulation.

We used a between-subjects design for the perspective manipulation because we wanted to avoid two countervailing tendencies in within-subjects designs. First, if asked to perform virtually the same task twice, there is a tendency to respond the second time differently, especially given only a minor variation. Second, this may be opposed by a tendency not to notice the change and make the same response. Consequently, in a within-subjects design, we may have found no differences between first and third-person responses simply as a result of these countervailing tendencies in responding. By using a between-subjects design, we are relying on participants in this population responding similarly to these arguments. If there are no differences here, then this is unlikely to be an artefact of the design. The experiment was designed using *Qualtrics* software and analysed using *SPSS 20.0.0* (for standard t-tests) and *R 3.1* (for Bayes factor analysis).

Materials and procedure: Six fallacies and four types of source credibility were tested. The fallacious structures used were predominantly chosen from previous literature. The study includes the argument from ignorance (Oaksford & Hahn, 2004), the slippery slope (Corner et al., 2011), the ad hominem (Oaksford & Hahn, 2013), and the ad Hitlerum (Harris et al., 2012). Also, two previously untested fallacies (Nirvana and the argument from silence) were included to increase the number of dialogues. The order of the fallacies was fully randomised.

Source credibility dialogues were taken from Harris et al. (2015). They describe appeals to the testimonies of others concerning the effectiveness of a made-up medical product. Following Harris and colleagues, sources were described in terms of trustworthiness (high: friend; low: enemy) and epistemic authority (high: doctor; low: musician).

Participants first read the six dialogues with the argument fallacies. Following the fallacies, participants read four dialogues with appeals to the testimony of others (source credibility). Participants were asked: "In light of the above dialogue, how convinced [are you now/ should A now be] of the conclusion proposed". Following the principle of charity, we predicted no difference in posterior degrees of belief for 1st vs. 3rd person perspectives, as participants were given no explicit reason to believe that the recipient of the argument in the dialogue had access to different information.

After each dialogue, the participant rated their degree of convincingness on a scale from 0-100 where 0 represented complete disbelief in the idea proposed in the dialogue and 100 represented complete belief in the idea. Excluding the participants mentioned above, this left 60 participants for weak (3rd-person), 54 for strong (3rd-person), 56 for weak (1st-person), and 62 for strong (1st-person) as well as 115 participants (source credibility, 3rd-person) and 117 participants (source credibility, 1st-person).

Results

Fallacies As participants have no specific reason to assume interlocutors are differentially able to reason, the principle of charity predicts that participants in the different perspective conditions (1st and 3rd-person) should yield the same posteriors. As the principle predicts support for the null hypothesis, we performed a Bayes Factor analysis and paired-samples t-tests for each fallacy dialogue.

For the Bayes Factor, we use Morey's software for *R* (version 0.9.2+) to calculate a Bayes factor with equivalence region for a two-sampled paired t-test. We set the effect size to 0.25 (a relatively conservative measurement) with a default Cauchy (r = 0.707) distribution for the priors. The Bayes Factor describes the likelihood of the null hypothesis being true for both weak and strong versions of the argument. That is, that participants who see arguments in 1st and 3rd-person format respond identically in line with the Principle of

[3] For the current calculations of Bayes factor with equivalence region, the effect size is set to 0.25 SD, as this can be considered a small effect size.

Charity. To evaluate the outcome of the analyses, Kass and Raftery (1995) describe the factors from 0-2 as barely worth mentioning, 2-6 as a positive indication for the hypothesis, 6-10 as strong evidence for the hypothesis and >10 as very strong evidence. In this case, the Bayes Factor indicates the ratio of the probability of the data given the null hypothesis and the probability of the data given the alternative hypothesis. In line with the Principle of Charity, the null hypothesis posits that we should observe no difference between groups.

In line with expectations, paired-sample t-tests show that no dialogue yielded significantly different responses when comparing 1st and 3rd-person conditions (for each argument, p-values were between .079 and .874). These results that the null hypothesis could not be rejected suggest that participants reached the same conclusions from a 1st or a 3rd person perspective regardless of the structure and strength of the argument.

Complementing this analysis, the Bayes Factor analysis also supports the conclusion, as all dialogues are in favour of the null hypothesis (BF between 2.01, the weak argument from silence, to 9.68, the strong argument from ignorance). Using Kass and Raftery's descriptions, these results provide either 'positive' or 'strong' evidence in favour of the Principle of Charity. As such, both the frequentist t-test and the Bayes Factor analysis support the Principle of Charity.

Each dialogue had between 54 and 60 responses (the number of participants in each condition). As all dialogues yielded responses in favour of the null hypothesis, we performed a further paired-sample Bayes Factor t-test with responses from all 1st and 3rd person dialogues collapsed over argument type. We found very strong evidence in favour of the null hypothesis, as means for 3rd-person evaluations were 47.00 and means for 1st-person evaluations were 46.13 (BF: 26.83). Responses from the argument fallacies show no difference between conditions and thereby provide strong (or very strong) evidence in favour of the principle of charity.

As a further test, we manipulated argument types as strong or weak and compared these in one-way ANOVAs. These tests provide mixed results for the effectiveness of the strength manipulations. We observed significant differences for the argument from silence and nirvana fallacy, borderline significant effects for the argument from ignorance, and no significant differences for the remaining arguments. As discussed later, however, this is not surprising, as the strength manipulation was tentative at best for the current design.

Source credibility As with argument fallacies, the principle of charity predicts that no difference should occur in posterior ratings of convincingness between question types given similar priors and likelihood ratios. As before, we performed a Bayes Factor analysis for paired-samples for each source credibility dialogue. Table 3 shows that the dialogues yielded 'strong' or 'very strong' evidence in favour of the null, supporting the principle of charity.

We collapsed across the four source credibility conditions for each question type and performed a Bayes Factor analysis for all responses. The data for source credibility yielded a Bayes Factor of 22.77, with means for 3rd-person evaluations were 48.07 and means for 1st-person evaluations were 49.01. This suggests that participants, in line with the Principle of Charity, treated reports similarly in both perspective conditions A Bayes Factor of this magnitude provides 'very strong' support for the principle of charity (Kass & Raftery, 1995). Consequently, results from the argument fallacy and source credibility dialogues provide 'positive', 'strong' or 'very strong' evidence in favour of the Principle of Charity. No dialogue failed to provide support for the principle, as the weakest Bayes Factor was 2.01 in favour of the null.

Harris et al., (2015) show that participants are sensitive to appeals to authority such that highly credible sources are more persuasive than less credible sources. As we observed no differences in question types, we collapsed responses from 1st and 3rd person source credibility dialogues. In line with previous findings, a paired-sample t-test showed that sources with low trustworthiness and expertise were less convincing than sources with high trustworthiness and low expertise (t(231) = 1.993, p=0.047) and sources with low trustworthiness and high expertise (t(231) = 2.516, p = 0.013). There was no significant difference between a source with high trustworthiness and low expertise compared with sources with low trustworthiness and high expertise (t(231) = .792, p= 0.429). Sources with high trustworthiness and high expertise were always the most convincing (compared to high trustworthiness and low expertise, t(231) = 3.802; compared to low trustworthiness and high expertise, t(231) = 3.712; and compared to low trustworthiness and low expertise, t(231) = 3.712; all p's < 0.001).

Study 2: Showing differences

Study 1 provides tentative support for the principle of charity, as we see no differences in observed posteriors for 1st-person and 3rd-person conditions. This holds true for both argument fallacies as well as source credibility. However, according to the principle of charity, people should also be able to believe that others will arrive at different conclusions if they have reason to believe other people operate from different points of view. For example, if a person believes COVID-19 is dangerous and that the vaccine is safe and effective, they may have reasonably decided to stay inside and vaccinate during the pandemic. However, another person may earnestly think that COVID-19 is not dangerous and that the vaccine has not been adequately tested (whether it is true or not is irrelevant to this case, as behaviour here rests on *subjective* degrees of belief). It is plausible to infer different behaviours due to the differences in subjective degrees of beliefs about the world (see Madsen et al., 2023 for a Bayesian model on vaccination hesitancy). According to the Principle of Charity, a person may believe that another person would reasonably behave in a different way or use information differently if they entertain a different view of the world.

Method

To explore whether people believe other people make similar inferences given their respective subjective beliefs, we use the Harris et al. (2015) model. We choose to base the design on the model for two reasons. First, we use it in Study 1. As such, we do not have to re-introduce the approach. Second, it provides an explicit reasoning mechanism that we can use to evaluate whether participants appear to make use of different inference assumptions for themselves than they do for people that may disagree on subjective degrees of beliefs in the credibility of sources of information.

Participants: 100 participants were recruited from Prolific. As the political figures were Rishi Sunak and Keir Starmer, participants had to be UK citizens and have English as their first language.

Materials and procedure: We use a between-subjects design where participants are asked to evaluate two scenarios regarding Rishi Sunak (Prime Minster of the UK and leader of the Conservative Party) and Keir Starmer (Leader of the Opposition in the UK and leader of the Labour Party). First, participants are asked how they rate the political expertise and trustworthiness of each candidate. Then, participants rate how expert and trustworthy a staunch supporter of each political figure would think the politician is.

To clarify the task of rating politicians, we provided the following definitions of 'political efficiency' (expertise): "politician's capacity to develop policies that are relevant and appropriate for their purpose regardless of whether you have the parliamentary power to implement those policies" and trustworthiness: "politicians who are honestly trying to develop policies that the politician believes are relevant and appropriate for their purpose regardless of whether you have the parliamentary power to implement those policies".

These definitions also clarify that we measure capacity and intent rather than parliamentary ability to carry out legislation (which would naturally differ between the Prime Minister and the Leader of the Opposition). Participants rate the trust and expertise scores from 0-100 for themselves and a hypothetical staunch supporter of each candidate (e.g. for trustworthiness 0 = the politician is completely untrustworthy and 100 = completely trustworthy). This provides the prior degree of belief for candidate perception (own and other), which can be used to model predicted posterior degrees of belief in line with Harris et al. (2015).

Having elicited the prior beliefs, participants are told of a new policy proposal for each politician. In the case of Sunak, they are told 'Hypothetically, Rishi Sunak has expressed support for a policy meant to rejuvenate the UK economy. It is a policy introduced by a Conservative MP and with strong backing from the party and Sunak himself'. They then have to rate how appropriate they believe the policy would be given the endorsement from Sunak/Starmer. This provides the posterior rating for their own perception. They are then asked to evaluate how appropriate a staunch supporter of the politician would believe the policy to be, which provides the posterior rating for the imagined other. This provides grounds for testing whether participants believe others would update with similar mechanisms.

Results

We expected overall ratings of trustworthiness and expertise from individual participants to be lower than their estimates of staunch political supporters of either politician. This is because we expect the general population to be less keen on either politician than the subset of the population that happens to strongly support each politician. We find support for this in the data. Participants rated their own perception of Sunak's expertise as 0.31 and trustworthiness as 0.24 while they believed supporters of Sunak would rate his expertise as 0.77 and his trustworthiness as 0.76. We find a similar pattern with Starmer. Participants' own expertise and trustworthiness rating are 0.48 and 0.44 respectively while supporters are rated as 0.72 and 0.78 respectively. Using an independent t-test, we find that all comparisons between own and supporter ratings are significantly different ($p < 0.001$ for all; t-values vary between 6.913 and 16.571). This is encouraging, as the differences allow us to test whether people believe that supporters would update similarly given the differences that participants assume exist between their own views and those of the political supporters.

To test this, we implement the Harris et al. (2015) source credibility model[4]. According to this model, people should update their subjective degree of belief on the back of their perception of the expertise and trustworthiness of the source. As can be gleaned from Figure 1, participants seem to update beliefs for themselves and the supporter in a similar manner.

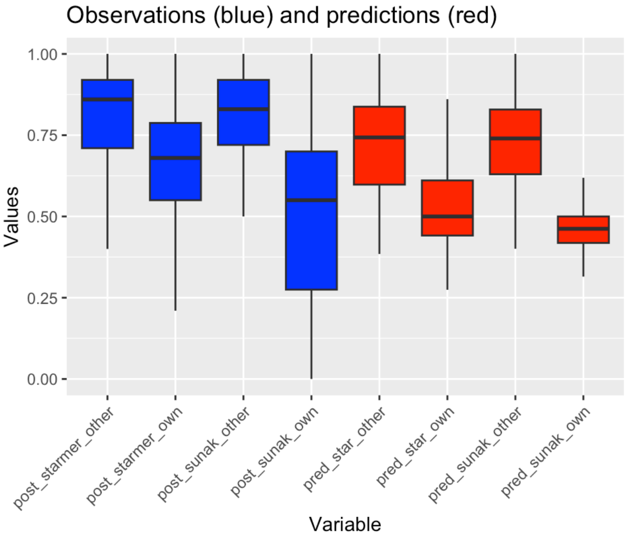

Fig. 1: Bayesian predictions (red) and observed posteriors (blue) for each group

Specifically, the Bayesian model predicts that participants' own posterior degree of belief for Sunak should be 0.47. The

[4] We use the same conditional probability table to calculate expected posteriors as Harris et al., (2015, p. 9). We use the priors for expertise and trustworthiness to calculate expected posteriors degrees of belief. To estimate predictions for the participants' own beliefs, we use their stated expertise and trustworthiness ratings for each politician. To estimate predictions for supporters, we use the participants' estimated priors for supporters of each candidate.

observed value is 0.51. Similarly, participants' own posterior for Starmer is predicted to be 0.53 and the observed is 0.65. For supporters, the model predicts that participants should believe the posterior is 0.72 for Sunak and 0.73 for Starmer while observed posteriors are 0.81 for Sunak and 0.8 for Starmer. For all four categories, then, we see a similar pattern where the Bayesian model captures the general posterior, but slightly undershoots for all conditions.

Study 2 indicates that people can imagine what a person who happens to find Sunak or Starmer expert and trustworthy should believe, even if this differs from their own subjective beliefs about the candidates. Further, it indicates that people may believe others would reason in a similar way, but would arrive at different conclusions due to their different subjective degrees of belief. This is in line with the principle of charity, as the observed differences appear to stem from reasonable disagreements on who is and who is not a credible source.

Discussion and concluding remarks

To our knowledge, this paper presents the first empirical evidence for two complementary approaches. Firstly, the principle of charity that people should not assume others to possess irrational beliefs or making irrational inferences unless they have specific evidence. It follows that the first-person perspective is *not* an epistemologically unique view and that the people should not assume that they are cleverer, better informed, more rational, etc. unless given specific reasons to believe so. Secondly, consistent with the Principle, Bayes' theorem predicts that given similar prior beliefs and likelihood estimations, interlocutors should reach the same posterior degree of belief regardless of epistemic positioning, argument strength and source credibility. Both studies support the intuition of the Principle of Charity.

Study 1 supports the idea that people conform to the Principle of Charity, with Bayes Factors between 'positive' evidence (2.01) and 'very strong' evidence (13.59). When we collapsed the results over the two sets of dialogues, the results provided very strong support. The analyses yielded Bayes factors of 26.83 (fallacies) and 22.77 (source credibility). These results provided evidence for the assumption that the type of fallacies and source credibility does not matter when eliciting posterior degrees of beliefs unless the participant has reason to assume differences. If the 1st person perspective does not have a unique position, it has potential consequences for psychological theories of the self and the other.

Study 2 provides further evidence for the Principle, as participants appear to believe people who differ from their own priors would use similar updating mechanisms (captured by Bayesian updating) to their own. In this study, we show that people may reasonably believe that others disagree with them on fundamental ratings (expertise and trustworthiness), but that they would nonetheless update reasonably *given* their subjective position. This is encouraging, as tentatively suggests that people may believe others to be fundamentally reasonable in how they treat information, which would, in turn, encourage deliberation and discussion. Indeed, if people believed that others arrive at conclusions due to flaws in reasoning, communication would be less desirable. Studies 1-2 provide initial support for the Principle of Charity.

Although the experiments support the Principle of Charity, we recognise that they are conducted in a highly controlled environment where participants viewed arguments on a screen, with background information suppressed, and no losses were at stake for expressing beliefs. Consequently, these result cannot be extrapolated to claim that we would not observe differences between conclusions from a 1st or a 3rd person perspective in real life. Differences in prior beliefs and conditional probabilities are commonplace between people (e.g. an evolutionary biologist might reasonably assume different prior beliefs and estimations of argument strength between herself and a creationist when considering evidence in favour of evolution theory). If we observe different degrees of beliefs, Bayes' theorem predicts that these should yield different posterior degrees of belief in the conclusion. Evidence from Bayesian argumentation experiments shows this pattern. In all, the Principle of Charity may be an underpinning principle of how humans approach other humans. But frequently it is reasonable to assume differences in beliefs, and consequently, we could expect differences in the degree of belief in the conclusions drawn.

Recent approaches to the psychology of the self suggests that the self does not have a stable core, but rather is an emergent property of variables such as the immediate environment, interactions with others, personal memories, socio-cultural background, framing and so on (e.g. Hood, 2012). Similarly, this has been suggested to apply to the ontogenesis of probabilistic estimations (Madsen, 2014). The Principle of Charity, in this framework, suggests that, ceteris paribus, emergent probabilities should not differ unless given reason to believe otherwise. However, if the self and probabilistic estimations emerge as a product of the immediate phenomenal placement in the world (e.g., context, framing, etc.), we should expect subtly (or potentially not so subtly) different probabilistic estimations to emerge given variation in the key variables defining the self. The current findings provide an interesting perspective on emergent rationality, as participants did not assume differences in epistemic capabilities or access for other interlocutors in a situation where there was no reason to assume otherwise. These findings are in line with theories of emergent properties of the self, as there *should* be no difference between the rational capabilities of the self and the other unless given reason to assume otherwise.

In conclusion, the studies provide empirical support for the Principle of Charity. As an underpinning Principle, it has potential theoretical and analytical entailments for how we conceive the self, the mechanisms of rationality, and how people approach other people. We seem compelled to believe others to be as rational as we are – unless we have evidence to believe otherwise.

References

Bovens, L., & Hartmann, S. (2003). *Bayesian epistemology*, Oxford: Oxford University Press.

Chaiken, S. & Maheswaran, D. (1994) Heuristic Processing Can Bias Systematic Processing: Effects of Source Credibility, Argument Ambiguity, and Task Importance on Attitude Judgement, *Journal of Personality and Social Psychology* 66 (3), 460-473

Corner, A., Hahn, U. & Oaksford, M. (2011) The psychological mechanisms of the slippery slope argument, *Journal of Memory and Language 64,* 133-152

Dennett, D. (1998) *The Intentional Stance*, MIT Press

Davidson, D. (1974) Belief and the basis of meaning. *Synthese* **27,** 309–323

Dienes, Z. (2014) Using Bayes to get the most out of non-significant results*, Frontier in Psychology* 5, 1-17

Fiske, S. T., Cuddy, A. J. & Glick, P. (2007) Universal dimensions of social cognition: warmth and competence, *Trends in Cognitive Sciences* 11, 77-83

Grice, P: (1989) *Studies in the Way of Words*, Harvard, MA: Harvard University Press

Hahn, U., Harris, A. J. L., & Corner, A. (2009). Argument content and argument source: An exploration. *Informal Logic 29,* 337-367

Hahn, U. & Oaksford, M. (2006) A Bayesian Approach to Informal Reasoning Fallacies. *Synthese* 152*,* 207-23

Hahn, U., & Oaksford, M. (2007) The rationality of informal argumentation: A Bayesian approach to reasoning fallacies, *Psychological Review 114*, 704-732

Harris, A. J. L., Corner, A. & Hahn, U. (2013) James is polite and punctual (and useless): A Bayesian formalisation of faint praise, *Thinking & Reasoning* 19 (3-4), 414-429

Harris, A., Hahn, U., Madsen, J. K. & Hsu, A. (2015) The Appeal to Expert Opinion: Quantitative support for a Bayesian Network Approach, *Cognitive Science* 39 (7), 1-38

Harris, A., Hsu, A. & Madsen, J. K. (2012) Because Hitler did it! Quantitative tests of Bayesian argumentation using *Ad Hominem*, *Thinking & Reasoning* 18 (3), 311-343

Hood, B. (2012) *The Self Illusion: Why there is no 'you' inside your head*, Constable & Robinson Ltd: London, UK

Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in cognitive sciences*, *20*(8), 589-604.

Kass, R. & Raftery, A. (1995) Bayes factors, *Journal of the American Statistical Association* 90, 773-79

Koralus, P. & Mascarenhas, S. (2013) *The Erotetic Theory of Reasoning*, Philosophical Perspectives 27, 312-365

Kruschke, J. K. (2011) Bayesian assessment of null values via parameter estimation and model comparison, *Perspectives on Psychological Science* 6, 299–312

Kruschke, J. K. (2013) Bayesian Estimation Supersedes the *t* Test, *Journal of Experimental Psychology: General* 142 (2), 573-603

Luria, A. R. (1976). *Cognitive Development its Cultural and Social Foundations*. Harvard University Press, Cambridge, MA.

Madsen, J. K. (2014) Approaching Bayesian subjectivity from a temporal perspective, *Cybernetics and Human Knowing* 21 (1/2), 98-112

Madsen, J. K. (2016) Trump supported it?! A Bayesian source credibility model applied to appeals to specific American presidential candidates' opinions, Papafragou, A., Grodner, D., Mirman, D., & Trueswell, J.C. (Eds.) *Proceedings of the 38th Annual Conference of the Cognitive Science Society*, Austin, TX: Cognitive Science Society*, 165-170*

Madsen, J. K., Cordier, D., Pilditch, T. D., & Zagala, H. (2023) Estimating attitudes toward vaccination: A Bayesian framework, *Proceedings of the 45th Annual Conference of the Cognitive Science Society*

Madsen, J. K., Hahn, U., & Pilditch, T. D. (2020). The impact of partial source dependence on belief and reliability revision. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 46* (9), 1795–1805.

Mascarenhas, S. & Koralus, P. (2016) Illusory inferences with quantifiers, *Thinking & Reasoning* 23 (1), 33-48

Mercier, H. (2011). On the universality of argumentative reasoning. *Journal of Cognition and Culture*, *11*, 85-113.

Mercier, H. & Sperber, D. (2011) Why do humans reason? Arguments for an argumentative theory*, Behavioral and Brain Sciences* 34, 57-111

Morey, R. D. & Rouder, J. N. (2011) Bayes Factor Approaches for Testing Interval Null Hypotheses*, Psychological Methods* 16 (4), 406-419

Oaksford, M. & Chater, N. (2007) *Bayesian Rationality: The probabilistic approach to human reasoning*. Oxford, UK: Oxford University Press.

Oaksford, M. & Hahn, U. (2004) A Bayesian approach to the argument from ignorance, *Canadian Journal of Experimental Psychology 58*, 75-85

Oaksford, M., & Hahn, U. (2013) Why are we convinced by the ad hominem argument? Bayesian source reliability and pragma-dialectical discussion rules, in F. Zenker (Ed.), *Bayesian argumentation* (pp. 39-58), Dordrecht, The Netherlands: Springer.

Perelman, C. & Olbrechts-Tyteca, L. (1969) *The New Rhetoric: A Treatise on Argumentation*, Notre Dame: University of Notre Dame Press

Petty, R. E. & Cacioppo, J. T. (1984) Source Factors and the Elaboration Likelihood Model of Persuasion*, Advances in Consumer Research* 11, 668-672

Quine, W. V. O. (1969) *Ontological Relativity and Other Essays*, Columbia University Press

Rouder, J. N., Speckman, P. L., Sun. D., Morey, R. D. & Iverson, G. (2009) Bayesian t-tests for accepting or rejecting the null hypothesis*, Psychonomic Bulleting Review* 16, 225-237

Sperber, D. & Wilson, D. (1995) *Relevance: Communication and Cognition*, 2nd edition, Blackwell Publishing

Thagard, P. & Nisbett, R. E. (1983) Rationality and Charity, *Philosophy of Science* 50 (2), 250-267

Wilson, N. L. (1959) Substances without substrata, *The Review of Metaphysics* 12 (4), 521-539