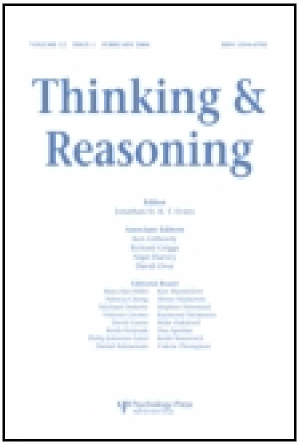


Thinking & Reasoning
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/ptar20

Because Hitler did it! Quantitative tests of Bayesian argumentation using ad hominem

Adam J. L. Harris [a], Anne S. Hsu [a] & Jens K. Madsen [a]
[a] Department of Cognitive, Perceptual and Brain Sciences, University College London, London, UK

Available online: 11 Jun 2012

THINKING & REASONING, 2012, 18 (3), 311–343

Psychology Press
Taylor & Francis Group

Because Hitler did it! Quantitative tests of Bayesian argumentation using *ad hominem*

Adam J. L. Harris, Anne S. Hsu, and Jens K. Madsen

Department of Cognitive, Perceptual and Brain Sciences, University College London, London, UK

Bayesian probability has recently been proposed as a normative theory of argumentation. In this article, we provide a Bayesian formalisation of the *ad Hitlerum* argument, as a special case of the *ad hominem* argument. Across three experiments, we demonstrate that people's evaluation of the argument is sensitive to probabilistic factors deemed relevant on a Bayesian formalisation. Moreover, we provide the first parameter-free quantitative evidence in favour of the Bayesian approach to argumentation. Quantitative Bayesian prescriptions were derived from participants' stated subjective probabilities (Experiments 1 and 2), as well as from frequency information explicitly provided in the experiment (Experiment 3). Participants' stated evaluations of the convincingness of the argument were well matched to these prescriptions.

Keywords: Argumentation; Reasoning; Bayesian probability; Fallacies.

Adolf Hitler is one of the most infamous characters in history. Responsible for atrocities such as the murder of approximately 6 million Jewish people in the Holocaust (see, e.g., Gigliotti & Lang, 2005), his name has become synonymous with evil. Consequently, Hitler has since been used as an argumentative tool to argue against propositions.

Correspondence should be addressed to Adam J. L. Harris, Department of Cognitive, Perceptual and Brain Sciences, University College London, London WC1H 0AP, UK. E-mail: adam.harris@ucl.ac.uk

The order of authorship is alphabetical. All authors contributed equally to this work. We thank the Nick Chater, David Lagnado, and David Shanks lab groups for useful discussions.

http://www.psypress.com/tar http://dx.doi.org/10.1080/13546783.2012.670753

Strauss (1953, pp. 42–43) coined the term ‘‘*ad Hitlerum*’’ to refer to the use of Hitler in argumentation: ‘‘Unfortunately, it does not go without saying that in our examination we must avoid the fallacy ... *ad Hitlerum*. A view is not refuted by the fact that it happens to have been shared by Hitler.’’ The status of the a*d Hitlerum* as an argumentation *fallacy* is implied in this quote, and stems from the traditional normative standard of argumentation, logic. As Strauss points out, the fact that Hitler shared a view does not *necessarily* refute that view (for example, using that argument against vegetarianism seems ridiculous); that is, it is not a deductively valid argument. However, there do exist a number of domains for which Hitler's endorsement may well be considered good evidence against a viewpoint (for example, in the domain of human rights). The recent Bayesian approach to argumentation (Hahn & Oaksford, 2006a, 2007a) provides a normative framework within which an argument can be understood as potentially providing some evidence in favour of a position, without the necessity of deductive validity in the manner of binary formal logic. On the Bayesian approach, argument strength is continuous and probabilistic. The use of a probabilistic framework enables us to distinguish between strong and weak instantiations of the same argument form, and thus to understand why the same argument form might be seen as weak in one context, but strong in another.

This article begins with a discussion of the various forms of the *ad Hitlerum*, before highlighting its critical components from a Bayesian perspective. Consequently, we suggest that Hitler can be invoked as part of a number of argument forms in order to argue against a proposition. In this article, however, we focus on perhaps the simplest instantiation of the *ad Hitlerum* argument. We do, however, propose that the probabilistic components that underlie this particular instantiation can be generalised to many other cases where Hitler is used in argumentation. We subsequently present three empirical studies, which demonstrate that people's understanding of the *ad Hitlerum* is well predicted by the relevant Bayesian parameters. Moreover, we extend evidence in favour of the rationality of participants' reactions to argumentative dialogue by testing the degree to which participants' understanding of an argument is in line with parameter-free quantitative predictions derived from the Bayesian framework. By grounding the parameters of the model in data (elicited from participants, or specified in the materials), we provide the most direct quantitative test of the degree to which people can be characterised as ‘‘Bayesian arguers’’. We thus provide an extension to work demonstrating (a) that people are sensitive to the relevant Bayesian parameters (e.g., Hahn, Harris, & Corner, 2009; Hahn & Oaksford, 2007a; Oaksford & Hahn, 2004), and (b) that the Bayesian model can retrospectively be fit to empirical data (Hahn & Oaksford, 2007a).

THE *AD HITLERUM*

Hitler has been used as an argumentative device ever since the Second World War. Strauss coined the *ad Hitlerum* name in 1953, but there is no reason to suppose that the argument was not present in colloquial expression prior to that date. Its usage continues today. For instance, a Penn State Trustee compared Reagan's speech to Young Americans for Freedom to Hitler's speeches to the Hitler Youth (Moser, 2006); anti-smoking campaigns in Germany have been linked to the Nazis' attempts to ban smoking in public places (Schneider & Glantz, 2008); and on his radio show, Rush Limbaugh noted that "Adolf Hitler, like Barack Obama, ruled by dictate" (McGlynn, 2010). Indeed, the pervasiveness of the argument is so dominant that Mike Godwin felt it necessary to "phrase" the truism that is the Godwin Law: "As an online discussion grows longer, the probability of a comparison involving Nazis or Hitler approaches one" (Godwin, 1994). The *ad Hitlerum* therefore remains prevalent in today's society. However, it can be seen that the argument can take on different forms. Below we outline those forms and categorise them within the existing catalogue of argument fallacies.

THE *AD HITLERUM* AS A SLIPPERY SLOPE ARGUMENT

Not all references to Hitler are made within the same argument form. Strauss (1953) explicitly considers the *ad Hitlerum* to be a specific case of the *reductio ad absurdum* argument. The *reductio ad absurdum* argues in favour of a proposition by deriving an absurdity from the denial of that proposition (Rescher, 2005). For logicians, the *ad absurdum* derives a self-contradiction from the denial of that proposition, typically via a logical derivation (Rescher, 2005). It is unclear to us, however, that the *ad Hitlerum* should be categorised as an instance of the *ad absurdum* (at least as that argument is defined by logicians). Rather, we suggest that Strauss, not an argumentation scholar himself, was considering an intuitive interpretation of what "*ad absurdum*" means. We therefore suggest that Strauss himself actually perceived the *ad Hitlerum* to be a specific instance of a slippery slope argument (SSA). SSAs have been somewhat resistant to a stable definition, but Corner, Hahn, and Oaksford (2011, p. 135) define them as comprising the following four components:

- An initial proposal (A).
- An undesirable outcome (C).
- The belief that allowing (A) will lead to a re-evaluation of (C) in the future.
- The rejection of (A) based on this belief.

Hitler can be used in a SSA, with the recognition of an implicit premise. For example, “You should not adopt policy X because Hitler did, [and look where that led . . .]”. The implication being that the adoption of Policy X would make the more negative aspects of Hitler’s dictatorship (e.g., the Holocaust, WWII) (C) more likely, thus resulting in the rejection of Policy X (A). From a Bayesian perspective, those factors that make SSAs differentially strong and weak have been explicated and supported in Corner et al. (2011).

THE *AD HITLERUM* AS *AD HOMINEM*

To refer to someone as “Hitler” or “a Nazi” is a direct attack upon that person. Within an argumentative discourse, this is an example of an abusive *ad hominem* argument. By comparing an individual or their plans to Hitler, the character of the individual is attacked, and thus it is suggested that the proposition they advance should be disregarded (see e.g., Walton, 1995, 2008b on the *ad hominem* fallacy). The *ad hominem* has classically been viewed as an argument fallacy from both a logical (e.g., Copi & Cohen, 1994) and pragma-dialectic (e.g., Van Eemeren & Grootendorst, 2004) perspective. From a logical perspective, the *ad hominem* is considered to be a fallacy of relevance—the characteristics of the individual advancing an argument are not necessarily relevant to the acceptability of that argument. From a pragma-dialectic perspective, the *ad hominem* is considered fallacious because an attack on an opponent’s character attempts to prevent the opponent from advancing standpoints and, as such, is not a valid defence of one’s standpoint (see Van Eemeren, Garssen, & Meuffels, 2009, 2012 this issue, for a detailed pragma-dialectic account of the *argumentum ad hominem*).

Although not logically valid evidence, an understanding of the characteristics of an argumentation opponent can provide relevant information in ascertaining the truth of the proposition under consideration. For example, once an individual’s credibility has been questioned, one is no longer able to have absolute confidence in facts reported by that individual (on source reliability see also, from a Bayesian perspective, Bovens & Hartmann, 2003; Corner, Harris, & Hahn, 2010; Hahn et al., 2009; Schum, 1981; and from a non-Bayesian perspective, e.g., Birnbaum & Stegner, 1979; Walton, 2008a).[1] Similarly,

[1] Note that Van Eemeren et al. (2012 this issue) recognise the non-fallaciousness of critiquing an argument opponent’s credentials, but only in reaction to an appeal to authority where that opponent refers to themself as the expert. The pragma-dialectic account does not, however, provide a measure for quantifying the strength of these arguments other than maintaining them to be either acceptable or non-acceptable.

if an individual *is* similar to Hitler in their moral values this would seem to provide *some* evidence against following their advice in certain contexts (specifically, moral contexts). Because certain of Hitler's moral judgements were clearly bad, if an individual with similar moral values to Hitler judges that an action is morally right, one should have less confidence in that judgement than if the individual were not similar to Hitler in this regard. Bovens and Hartmann (2003) and Hahn et al. (2009) both argue that Bayesian probability provides an appropriate normative framework within which to investigate source reliability generally, and the present article provides further support for this notion.

Copi and Cohen (1994) extend the *ad hominem* classification to arguments in which "a conclusion or its proponent are condemned simply because the view defended is defended also by persons widely believed to be of bad character" (p. 123). A simple *ad Hitlerum* argument clearly takes this argumentation form: "Policy *X* is not a good idea because Hitler adopted the same policy." Here, the policy is being condemned because Hitler (whom reasoners will presumably consider to be of bad character) also adopted it. Thus, this is an *ad hominem* argument because Hitler's bad character is being used to condemn a policy he proposed. Some researchers might consider that this is too broad an extension of the *ad hominem* argument because it does not constitute a direct attack against the proponent of the present argument.

In the General Discussion we consider the possibility that this example of the *ad Hitlerum* might be equally well classified as an example of an appeal to (negative) authority. However, in terms of the existing literature, Copi and Cohen's conceptualisation provides the best classification for the *ad Hitlerum* arguments considered in the current article. We therefore consider them to be examples of the *argumentum ad hominem*.

In the empirical studies that follow, we consider *ad Hitlerum* arguments that most closely resemble this latter structure. Specifically, the arguments used in our experimental materials take a dialogue form in which one individual argues for the likely "badness" of a proposition by stating that Hitler entertained the same proposition. We note, however, that the aspect of the *ad Hitlerum* argument form that we focus on is critical for the evaluation of all forms of the argument, including as an SSA and more generic forms of the abusive *ad hominem*. Essentially, our probabilistic formulation determines the likelihood that a proposal is bad given that Hitler shared it. This is relevant for all forms of the *ad Hitlerum*, as they all necessitate that Hitler is viewed as evidence against the goodness of a proposal.

BAYESIAN ARGUMENTATION

We will investigate the *ad Hitlerum* within the Bayesian framework of argumentation (Hahn & Oaksford, 2006a, 2007a).[2] Central to the Bayesian approach is the recognition that the amount to which people believe a given proposition is a matter of *degree*. That is, propositions do not have truth values of 0 or 1, but rather an individual's degree of belief in a particular proposition can be understood as a probability between 0 and 1 (see also e.g., Evans & Over, 2004; Howson & Urbach, 1996; Oaksford & Chater, 1998, 2007).

Before commencing an argumentation dialogue, an individual will hold a prior belief in a given proposition, or hypothesis (h). This belief takes the form of a probability, and is termed the prior, $P(h)$. An argumentation dialogue consists of two parties providing evidence in order to try and convince their argumentation partner of the truth of their position. Thus the discussants within an argumentation may each provide and receive evidence in support of, or contrary to, the hypothesis under consideration. The normative procedure by which individuals should update their degree of belief in a hypothesis h upon receipt of an item of evidence e is given by Bayes' Theorem:

$$P(h \mid e) = \frac{P(h)P(e \mid h)}{P(e)} \tag{1}$$

where $P(h \mid e)$ represents the posterior degree of belief that a hypothesis h is true after having received some evidence, e. $P(e \mid h)$ represents the probability of receiving the evidence e if the hypothesis is true. $P(e)$ is the probability of the evidence occurring regardless of the truth or falsity of the hypothesis, and can be calculated from $P(h)$, $P(e \mid h)$, and $P(e \mid \neg h)$—the probability of receiving the evidence if, in fact, the hypothesis is *not* true ($\neg h$). These probabilities are considered to be subjective degrees of belief. Consequently, the Bayesian framework requires beliefs to be coherent (consistent with the axioms of probability), but does not require correspondence (i.e., matching) between these beliefs and real-world probabilities. The Bayesian approach stipulates that, upon receiving evidence, people should update their probabilistic degrees of belief in a hypothesis in accordance with the prescriptions of Bayes' Theorem.

Empirical evidence has demonstrated that people are sensitive to the relevant probabilistic features of an argument across a variety of

[2] Korb (2004) also suggested Bayesian probability as a normative framework within which to investigate argumentation.

argumentation fallacies, including: the argument from ignorance (Hahn, Oaksford & Bayindir, 2005; Oaksford & Hahn, 2004), slippery slope arguments (Corner et al., 2011), circular arguments (Hahn & Oaksford, 2007a), and more prototypical examples of *ad hominem,* where the focus has been on the role of the prior (Oaksford & Hahn, in press). More generally, people's treatment of source expertise appears consistent with Bayesian prescriptions (Hahn et al., 2009). In addition, Corner and Hahn (2009) demonstrated that the evaluation of arguments pertaining to current scientific issues of considerable import were also in line with Bayesian prescriptions. In this article, we extend the results cited above in two ways. Firstly, we consider an additional argument type (a particular instantiation of the *ad hominem*). Secondly, our experimental design enables us to test the degree to which participants' evaluations are consistent with the *quantitative* prescriptions of Bayes' Theorem, where these prescriptions are derived from conditional probabilities estimated by participants or presented in the scenario. Precise, a priori, predictions can be calculated using Bayes' Theorem *if* one has information about the relevant conditional probabilities, $P(e \mid h)$ and $P(e \mid \neg h)$, and the prior, $P(h)$. In Experiments 1 and 2, we collect data from participants pertaining to their subjective conditional probabilities, allowing us, assuming a specified prior, to make precise predictions as to how convincing they should find the argument. In Experiment 3, we again collect these conditional probabilities, but also present a design in which we attempt to define these conditional probabilities such that there is an objective probabilistic prediction against which participants' ratings can be evaluated (see also Griffiths & Tenenbaum, 2006; Harris & Hahn, 2009). It is prudent to note here that, although standard, subjective Bayesianism makes no correspondence prescriptions, using frequency information from our environment (especially in the absence of other relevant information) is desirable and relevant if we wish to use evidence to maximise the accuracy of our beliefs, as we would in any situation in which an action was to be based on these beliefs (e.g., vote in favour of, or against, the proposed policy).[3]

FORMALISING THE *AD HITLERUM*

For the type of argument investigated in this paper, our hypothesis concerns the goodness of a proposal given that Hitler endorsed it. From Bayes' Theorem, we can prescribe normative predictions for how good a proposal *should* be perceived to be given that Hitler had endorsed it. In this work we make the simplifying assumption that proposals are either good or not good (i.e., "bad"). For illustration purposes, we use the example proposal of a

[3] From a philosophical perspective, on the epistemic norm of accuracy see Leitgeb and Pettigrew (2010a, 2010b).

policy previously implemented by Hitler. Here we ask participants how likely they expect it is that a policy is good after they hear the claim that Hitler had implemented such a policy in the past. We define participants' posterior probability that the policy is good, given the association with Hitler as the posterior probability, $P(good\,|\,Hitler)$. Bayes' Theorem (Equation 1) therefore defines $P(good\,|\,Hitler)$ as:

$$P(good\,|\,Hitler) = \frac{P(good)P(Hitler\,|\,good)}{P(Hitler)} \tag{2}$$

from which the denominator can be expanded to:

$$P(good\,|\,Hitler) = \frac{P(good)P(Hitler\,|\,good)}{P(good)P(Hitler\,|\,good) + P(\neg good)P(Hitler\,|\,\neg good)} \tag{3}$$

where $P(Hitler\,|\,good)$ and $P(Hitler\,|\,\neg good)$ are the probabilities that Hitler had implemented the policy given that it was a *good* and *not good* policy respectively. For simplicity, in explaining the experimental design and results of Experiments 1, 2, and 3 we refer to these two conditional probabilities collectively as the *likelihood probabilities*. Oaksford and Hahn (in press) have analysed the role of the prior in a Bayesian analysis of *ad hominem* arguments, but here we assume that the prior degree of belief (before any evidence is given) is that the policy is equally likely to be bad as it is to be good, that is $P(\neg good) = P(good) = .5$. We can then rewrite Equation 2 as:

$$P(good\,|\,Hitler) = \frac{1}{\frac{P(Hitler\,|\,\neg good)}{P(Hitler\,|\,good)} + 1} \tag{4}$$

This shows that the probability of a policy being good only depends on the likelihood ratio:

$$\frac{P(Hitler\,|\,good)}{P(Hitler\,|\,\neg good)}. \tag{5}$$

Where the likelihood ratio is less than 1, that is $P(good\,|\,Hitler) < P(good)$, the policy will be perceived as less good once one learns that it was previously implemented by Hitler. Where the likelihood ratio is greater than 1, the opposite will apply. In the case of the *ad Hitlerum* we assume that the former condition is more often consistent with people's beliefs.

In the experiments that follow, we not only require participants to rate $P(\mathit{good} \mid \mathit{Hitler})$, but subsequently require them to provide ratings of the likelihood probabilities, $P(\mathit{Hitler} \mid \mathit{good})$ and $P(\mathit{Hitler} \mid \mathit{good})$. By setting up the experimental materials in such a way that the prior degree of belief, $P(\mathit{good})$, can be assumed to equal .5, we can calculate a Bayesian posterior for each argument for each participant, and compare this with participants' explicit ratings of $P(\mathit{good} \mid \mathit{Hitler})$ in order to ascertain the degree of quantitative fit between participants' argument ratings and the Bayesian prescriptions.

OVERVIEW OF THE CURRENT EXPERIMENTS

Experiment 1 provides participants with simple *ad Hitlerum* arguments in opposition to five different propositions. We test whether participants are sensitive to changes in the topic of the argumentation, and whether these sensitivities can be predicted on the basis of the conditional probabilities they provide pertaining to the likelihood probabilities of Hitler as an item of evidence. If participants are truly sensitive to the important rational considerations of the likelihood probabilities, they should revise their opinion of an argument's convincingness when provided with additional information referring to these conditional probabilities. This is the main question addressed in Experiment 2. Experiment 3 provides background information about the likelihood probabilities, so as to create an objective rational prediction for the convincingness of these arguments. Because we are unable to control participants' background knowledge about Hitler, the argument used in Experiment 3 concerns a fictional alien leader, "Zhang", but the form of the argument is identical to the *ad Hitlerums* we use in Experiments 1 and 2. In addition, Experiment 3 explores a greater range of explicit likelihood probabilities and also examines arguments which are both for and against a proposal (*ad Hitlerum* arguments are only given against the proposal). In the work that follows, some probabilities are explicitly provided to participants in the experimental materials, while others are provided by participants. For clarity we term the former, *objective* probabilities and the latter, *subjective* probabilities, which we denote with subscripts $_{subj}$ and $_{obj}$ respectively.[4]

[4] The "objective probabilities" we present to participants were, in fact, frequencies. Whether frequencies represent objective probabilities is a philosophical question that has generated considerable debate (for recent coverage see Pettigrew, in press). Here this information is provided by, and directly observable to, the experimenter and hence for simplicity we refer to them as objective probabilities, so as to maintain consistency in subsequent notation.

EXPERIMENT 1

We presented participants with a series of dialogues concerning the likely "goodness" of propositions for various topics. All dialogues contained the same structure, and featured two speakers, *A* and *B*. In the dialogue *B* argues to *A* that a particular proposition is a bad idea because Hitler had endorsed it. We then asked participants what *A*'s opinion should be of the proposition after hearing the argument. A participant's judgement of what *A*'s opinion should be is represented as the posterior probability $P_{subj}(good\ (idea) \mid Hitler\ (endorsed\ it))$, hereafter written as $P_{subj}(good \mid Hitler)$. In order to evaluate the degree of rationality of participants' assessments, we assessed $P_{subj}(Hitler\ (endorsed\ it) \mid good\ (idea))$ and $P_{subj}(Hitler\ (endorsed\ it) \mid bad\ (idea))$, hereafter written as $P_{subj}(\mathrm{Hitler} \mid \mathrm{good})$ and $P_{subj}(\mathrm{Hitler} \mid \mathrm{bad})$, which we collectively refer to as likelihood probabilities. From a participant's likelihood probabilities, Equation 3 can be used to make a quantitative prediction for what the participant's judgement of *A*'s opinion $P_{subj,\ predict}(good \mid Hitler)$ should be.[5] Here the subscript $_{subj,predict}$ is used to indicate a prediction based on participants' subjective responses. This can be compared with $P_{subj}(good \mid Hitler)$ in order to assess how well participants' responses can be described by the rational Bayesian framework.

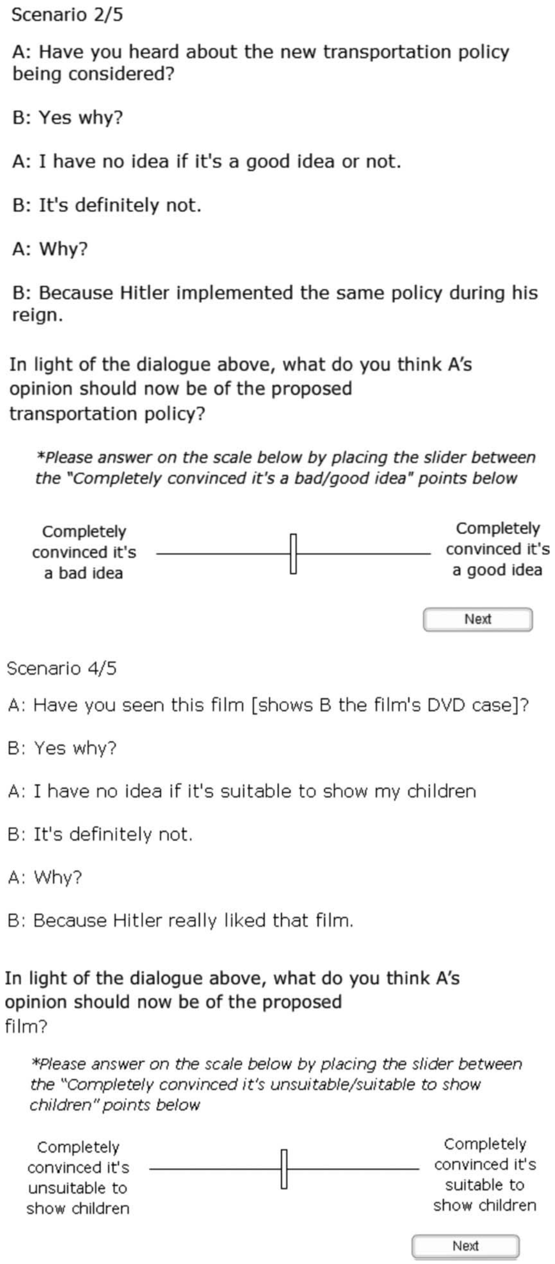
Method

Participants. After excluding participants who were under 18 years old (in line with departmental ethical guidelines),[6] 61 participants (42 female) volunteered for the experiment, which was advertised on *http://psych.hanover.edu/research/exponnet.html*, a site for recruiting volunteers to participate in web-based experiments. The age range was from 18 to 60 years (median = 22 years) and participants predominantly reported being from North America ($N = 40$), with the next largest group ($N = 9$) being from Africa. Participants received no remuneration for completing this short experiment.

Design. A within-participants design was employed. There were five different argument topics, presented in the form of a dialogue between *A* and *B* (see Figure 1). Each dialogue had the same structure. The five different topics were: a transportation policy, an economic policy, a religious

[5] Of course, the "goodness" or "badness" of a proposal concerns a value judgement rather than a factual statement. We are therefore taking these terms in their vernacular meaning as being tied to society. That is, whether people in general would perceive something as good or bad.

[6] This check was performed for all experiments reported.



Scenario 2/5

A: Have you heard about the new transportation policy being considered?

B: Yes why?

A: I have no idea if it's a good idea or not.

B: It's definitely not.

A: Why?

B: Because Hitler implemented the same policy during his reign.

In light of the dialogue above, what do you think A's opinion should now be of the proposed transportation policy?

Please answer on the scale below by placing the slider between the "Completely convinced it's a bad/good idea" points below

Completely convinced it's a bad idea
Completely convinced it's a good idea
Next

Scenario 4/5

A: Have you seen this film [shows B the film's DVD case]?

B: Yes why?

A: I have no idea if it's suitable to show my children

B: It's definitely not.

A: Why?

B: Because Hitler really liked that film.

In light of the dialogue above, what do you think A's opinion should now be of the proposed film?

Please answer on the scale below by placing the slider between the "Completely convinced it's unsuitable/suitable to show children" points below

Completely convinced it's unsuitable to show children
Completely convinced it's suitable to show children
Next

Figure 1. Screenshots showing the main dependent variable in Experiment 1. The top panel shows the dialogue structure for the policy topics (transportation in this figure). The bottom panel shows the film topic.

policy, a law banning smoking in parks, and a film. In the dialogue, we enforced our assumption of neutral priors on whether the proposal is good or bad, $P(good) = P(bad) = .5$, with A stating, ‘‘I have no idea if it's [*referring to the given topic*] a good idea or not.’’ We note that there are other possible pragmatic interpretations of A's statement: For example, the phrase ‘‘no idea’’, might just reflect a certain degree of probabilistic uncertainty, without guaranteeing an uncertainty of 0.5, but here we assume $P(h) = .5$. The presentation order of the five topics was randomised across participants.

After assessing all five dialogues, participants provided likelihood probabilities for each topic, again with the order randomised across participants (independent of the first presentation order).

Materials and procedure. The experiment was programmed in Adobe Flash and run online in the participant's web browser. The five dialogues were presented sequentially with only one dialogue on the screen at a time. Having read a dialogue, participants were asked to rate the convincingness of the argument in response to the question: ‘‘In light of the dialogue above, what do you think A's opinion should now be of the proposed [transportation policy/economic policy/policy on religion/law banning smoking in parks/film]?’’. Participants made their responses by moving a slider on a scale between ‘‘Completely convinced it's a bad idea’’ and ‘‘Completely convinced it's a good idea’’ in the four policy topics, and ‘‘Completely convinced it's unsuitable to show children’’ and ‘‘Completely convinced it's suitable to show children’’ in the film topic (see Figure 1). The slider was initially positioned at the halfway point on the scale. The participant's positioning of the slider on the scale was taken to be directly proportional to the degree of belief that the proposition was good, given the knowledge that Hitler had endorsed it. This is represented as the posterior probability $P_{subj}(good \mid Hitler)$. This is the main dependent variable of interest and will be referred to in the remainder of this article as the *posterior rating*. Note that we followed Oaksford and Hahn (2004) in asking about how convinced A *should* now be. This is in line with the Bayesian theory of argumentation being a normative one (i.e., that arguments can be evaluated according to this rational standard). In this instance, the participant is taking the place of the ‘‘reasonable critic’’ in Van Eemeren and Grootendorst's (2004, p. 1) definition of argumentation: ‘‘Argumentation is a verbal, social, and rational activity aimed at convincing a reasonable critic of the acceptability of a standpoint by putting forward a constellation of propositions justifying or refuting the proposition expressed in the standpoint.’’

After seeing all five argument dialogues, participants were asked, for each topic, two questions designed to elicit their subjective beliefs about the likelihood probabilities, $P_{subj}(Hitler \mid good)$ and $P_{subj}(Hitler \mid bad)$. The five topics were presented sequentially with only the questions relevant to one topic on the screen at a time. The two questions referring to a single topic were on the screen at the same time. The first question always elicited $P_{subj}(Hitler \mid good)$ and the second question elicited $P_{subj}(Hitler \mid bad)$ for the given topic. The first questions—eliciting $P_{subj}(Hitler \mid good)$—were of the form: "Of all German transportation policies between 1925 and 1945 that Historians now recognise as being GOOD, how many do you think Hitler was responsible for?". The second questions were exactly the same, but with "GOOD" replaced with "BAD". The questions for the other political scenarios used the same format with minor word changes. The questions for the film topic read: "Of all the films [SUITABLE/UNSUITABLE] for children, what proportion do you think Hitler would have liked?". For each question, participants used the slider (between "none" and "all") to indicate their responses. We chose to elicit likelihood probabilities in frequentist form as we perceived it as the most intuitive format for participants in the contexts of these particular experiments (see, e.g., Gigerenzer, 2002). Alternative possibilities would be to directly ask for the subjective probability, or to ask for a judgement of confidence in the conditional: "If the policy was good, then Hitler was responsible for it," with which ordinal predictions could be made on the recognition that participants typically represent such conditionals as conditional probabilities (e.g., Over, Hadjichristidis, Evans, Handley, & Sloman, 2007).[7]

At the end of the experiment participants provided their age and gender before being thanked and debriefed.

Results and discussion

A one-way ANOVA demonstrated that the *ad Hitlerum* argument was viewed as differentially convincing across the five different topics, $F(3.43, 240) = 16.74$, $p < .001$, $\text{eta}_p^2 = .22$ (Greenhouse-Geisser correction for repeated measures applied). This result cannot be explained on the basis of the argument structure (which is identical across conditions), but is readily explainable from a Bayesian perspective, which takes into account the content of the argument in its evaluation. The content is accounted for because it affects the likelihood probabilities, which determine the argument's convincingness, according to the Bayesian account.

We used each participant's likelihood probabilities to calculate, for each topic and each participant, a Bayesian prediction for how convinced *A* should be that the proposal is good in light of the argument, that is $P_{subj,predict}(good \mid Hitler)$. As mentioned above, we assumed an initial prior of

[7] We thank an anonymous reviewer for suggesting conditional confidence judgements.

$P(good) = .5$. A further simplifying assumption used in this experiment and Experiment 2 is that the judgement given by a contemporary person on the five topics would be the same regardless of whether they thought the topics were to be considered/implemented between 1925 and 1945 or considered/implemented today. The same one-way ANOVA as above was conducted on the $P_{subj,predict}(good \mid Hitler)$ values and the main effect of topic was again observed, $F(3.50, 209.78) = 13.11$, $p < .001$, $\text{eta}_p^2 = .18$. Figure 2 demonstrates that the average pattern of argument convincingness across topics was well predicted by the Bayesian model. A topic-level analysis between the average value of $P_{subj,predict}(good \mid Hitler)$, and the average value of the posterior rating for each topic yielded a correlation of $r_{\text{adj}}(3) = .94$, $p < .05$.[8] This result indicates that 89% of the variance in the posterior ratings across the different topics was explained by the Bayesian model. An alternative way to analyse the data is to compute a correlation for each participant individually across their five judgements and corresponding predicted judgements and then compute the average of those correlation coefficients.[9] This individual-level analysis resulted in a mean correlation coefficient of .24. Wallsten, Budescu, Erev, and Diederich (1997) show that, where participants' judgements are subject to some degree of random error or noise, a group-level average will be closer to the true values underlying those

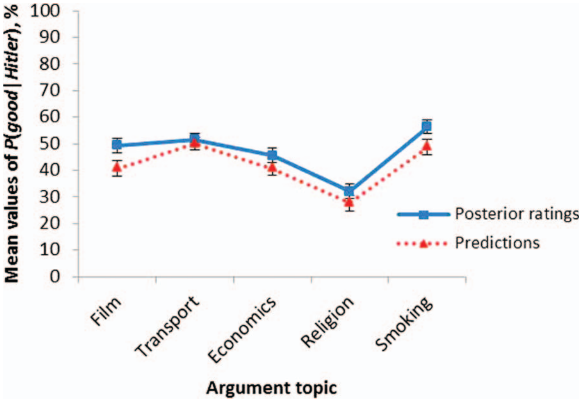


Figure 2. Posterior ratings, $P_{subj}(good \mid Hitler)$, and predicted ratings based on subjective likelihoods, $P_{subj,predict}(good \mid Hitler)$, across the five argument topics in Experiment 1. Error bars are plus and minus 1 standard error.

[8] r_{adj} is the adjusted correlation coefficient, correcting for the small number of datapoints (Howell, 1997, p. 240).

[9] Three participants could not be included in this analysis as their responses did not enable a correlation coefficient to be computed.

judgements. Hence the better model fit observed with the group-level average (Figure 2) itself provides some support for the hypothesis that participants' posterior ratings were (somewhat) noisy estimates of the Bayesian predictions.

EXPERIMENT 2

The results of Experiment 1 showed that, on average, people's posterior ratings were well predicted by their likelihood probabilities. Thus, people's responses to the argument were in line with the predictions of rational Bayesian reasoning. A further prediction of the rational Bayesian model is that if people were given objective information about the likelihood probabilities, this should be taken into account and thus should affect their posterior ratings.[10] In Experiment 2, we therefore provided participants with objective values for these likelihood probabilities. This was done by introducing a third interlocutor into the argument, who provided participants with information pertaining to these likelihood probabilities.

Method

Participants. A total of 184 participants (76 female), aged between 18 and 78 (median age 27), were recruited using Amazon Mechanical Turk (see e.g., Paolacci, Chandler, & Ipeirotis, 2010, in support of the validity of experimental data obtained via Amazon Mechanical Turk). Participants predominantly reported being from North America ($N = 72$) and Asia ($N = 70$). Each participant was paid $0.10 for completing this short experiment.

Design. A 3 × 2 (topic × likelihood probability) between-participants design was employed. We chose films and transport as two of the topics as these gave rise to the least change in participants' ratings in Experiment 1 (that is, posterior ratings were the closest to the assumed prior of .5). We also included religion, as this showed the most negative belief ratings, and is associated with the worst of all Hitler's atrocities, thus providing a strict test for the rational account. The likelihood probability variable refers to whether $P_{obj}(Hitler \mid Good)$ was high and $P_{obj}(Hitler \mid Bad)$ was low (the

[10] Although this isn't required on strict subjectivist Bayesianism, this follows given the assumption of a norm for accuracy (see also Leitgeb & Pettigrew, 2010a, 2010b), which is especially desirable in situations in which future actions may be based on these beliefs.

positive condition), or vice versa (negative condition). Low and high likelihood probabilities were .2 and .8 respectively.[11]

Materials and procedure. The initial screens of the experiment were identical to those in Experiment 1 (except that now only one topic was shown per participant). Participants were asked to make the same initial rating about the proposal's goodness as in Experiment 1. This we will refer to as *posterior rating*$_1$. Following this rating, a new screen appeared, headed with the words "Person C now joins the argument". The initial dialogue was faded but remained visible, and *C*'s contribution was presented below it in black. An example of the structure of *C*'s contribution is as follows (shown for the transportation topic):

> Expert historians agree that between 1925–1945, there were just as many GOOD policies on transportation in Germany as there were BAD.
>
> Of all German policies on transportation between 1925 and 1945 that historians now recognise as being GOOD, it is a fact that Hitler was responsible for [80%/20%] of them.
>
> Of all German policies on transportation between 1925 and 1945 that historians now recognise as being BAD, it is a fact that Hitler was responsible for [20%/80%] of them.

C mentioned the equal number of good and bad policies on the topic in Germany so as to guard against the possibility that the period 1925–1945 might have been perceived as a particularly bad or good period for policies. Such a perception would reduce the consistency between the conditional probability questions asked and those required to calculate a Bayesian posterior degree of belief (the conditional probability questions asked how likely Hitler was to have been involved in something bad or good during the period 1925–1945, whereas the posterior Bayesian question concerned how good something is likely to be in the present day).[12] On the next screen, the whole dialogue remained visible and participants were asked, "In light of this new information, what do you think *A*'s opinion should now be of the proposed policy on religion?" Thus participants were asked a second time for their responses, which we term *posterior rating*$_2$. Participants again responded using a slider, as they had for their initial judgements.

Although participants were explicitly provided with likelihood probabilities regarding how likely good and bad policies/films were to be

[11] There is no probabilistic constraint for $P(Hitler\,|\,good)$ and $P(Hitler\,|\,bad)$ to be complementary. While we only used complementary values in Experiment 2, this was not the case in Experiment 3.

[12] We note that not taking this step in Experiment 1 was a limitation of that experiment. There is, however, no reason why this should have exaggerated the fit of the Bayesian model.

implemented/liked by Hitler, participants' subjective likelihood probabilities would still have been influenced by their own subjective beliefs. Thus, at the end of the experiment, we elicited the likelihood probability judgements, $P_{subj}(Hitler \mid good)$ and $P_{subj}(Hitler \mid bad)$, from participants, as was done in Experiment 1. Here participants only provided likelihood probability judgements for the topic that they had read. All other aspects of the procedure were identical to Experiment 1.

Results and discussion

The first stage of Experiment 2 was equivalent to a between-participants replication of Experiment 1 with only three topics. However, a factorial ANOVA performed on *posterior ratings*$_1$ did not replicate Experiment 1's significant effect of topic ($F < 1$). We tentatively attribute this difference to different recruitment methods, given that these ratings should be based on participants' subjective beliefs about the relevant conditional probabilities, which can differ across people.

The main analyses of interest, however, concerned the effect of likelihood probability on judgements of *posterior ratings*$_2$, for which the population averages are displayed in Figure 3. Results showed that ratings were affected by the likelihood probability manipulation, $F(1,\ 178) = 56.50$, $p < .001$, $eta_p^2 = .24$. There was also a significant interaction between likelihood probability and topic, $F(2,\ 178) = 3.84$, $p = .04$, $eta_p^2 = .04$. This interaction

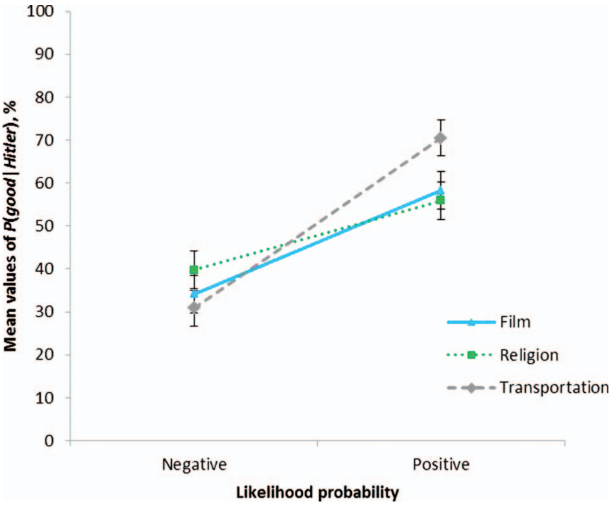


Figure 3. Mean ratings of *posterior ratings*$_2$ for the three argument topics in Experiment 2. Error bars are plus and minus 1 standard error.

is explainable from a Bayesian perspective as topics can differ in the degree to which participants would assimilate the values of $P_{obj}(Hitler \mid good)$ and $P_{obj}(Hitler \mid bad)$ offered to them by Person C in the dialogue (for example, participants might be less likely to *believe* the conditional probabilities for topics on which they already have strong opposing views). Although no effect of topic was observed in the analyses of *posterior ratings*$_1$, this is likely to be a noisy estimate of participants' true ratings (see also Vul & Pashler, 2008; Wallsten et al., 1997), and these ratings might have been more stable for some topics than for others. The significant difference observed between these three topics in the less-noisy, within-participant, Experiment 1 offers some support for this suggestion. We also acknowledge, however, that certain argument topics might be particularly difficult to fully explain within a rational framework. For example, for the topic of religion, people may be more reluctant to adopt a rational framework and to judge a policy on religion associated with Hitler as good, despite being provided with objective likelihood probabilities consistent with such a judgement.

Further support for the possibility that certain highly emotive argument topics might be less susceptible to a rational treatment than other topics comes from an analysis of the Bayesian predictions, $P_{subj,predict}(good \mid Hitler)$ (calculated as in Experiment 1). An ANOVA conducted on these values replicated the effect of the likelihood probability condition, $F(1, 178) = 114.16$, $p < .001$, $\text{eta}_p^2 = .39$, but failed to replicate the significant interaction observed above, $F(2, 178) = 1.33$, $p = .27$. This suggests that the interaction cannot be explained in terms of the subjective likelihood probabilities provided by participants.

As in Experiment 1, we wished to test the degree to which participants' posterior ratings were quantitatively predicted by the Bayesian model. Because subjective likelihood probabilities were collected after participants had made their second rating, we correlated $P_{subj,predict}(good \mid Hitler)$ with *posterior ratings*$_2$. Once again, mean ratings for each experimental condition were well predicted by the Bayesian model, $r_{\text{adj}}(4) = .95$, $p < .01$ (see Figure 4), accounting for 89% of the variance in mean responses across experimental conditions. Note that Figure 4 shows that the only topic for which the error bars of *posterior ratings*$_2$ and $P_{subj,predict}(good \mid Hitler)$ do not overlap concern the topic of religion, the topic most associated with highly emotive atrocities carried out by Hitler.

EXPERIMENT 3

Experiment 2 showed that participants' posterior ratings following an *ad Hitlerum* argument were, in general, well predicted by their subjective likelihood probabilities, in line with the predictions of the Bayesian

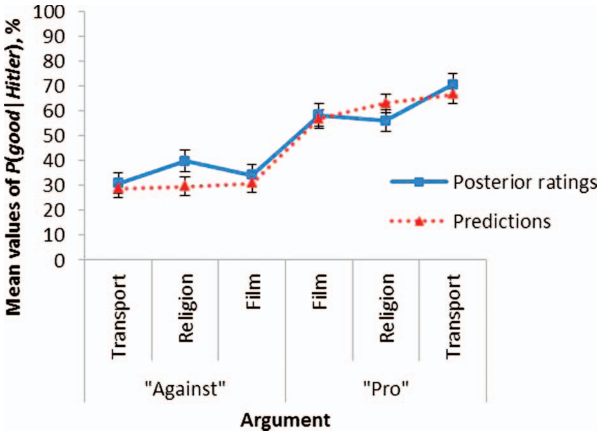


Figure 4. *Posterior ratings*$_2$ and predicted ratings, $P_{subj,predict}(good\,|\,Hitler)$, across the six experimental conditions in Experiment 2, ordered according to predicted rating value. Error bars are plus and minus 1 standard error.

framework. Experiment 3 was designed as a further exploration of this. Here, as in Experiment 2, likelihood probabilities were explicitly provided to participants. These likelihood probabilities were systematically varied across 17 different pairs of values (see Table 1), which gave rise to a range of predicted judgements regarding the goodness of the proposal. We also sought a more controlled experimental design in which there would be no previous knowledge about the likelihood probabilities, and for which we could investigate the effects of argument direction (using a reference to an individual either to attack or *support* a proposition). To achieve this, the *ad Hitlerum* argument was modified to an "*ad Zhangum*", where the scenario concerned policies on the fictional Planet Xenon, which had once been governed by Zhang. The aim was to maintain the structure of the *ad Hitlerum* argument while minimising the impact of participants' prior real-world knowledge of likelihood probabilities and allowing for exploration of both argument directions.

Method

Participants. A total of 725 participants (305 female), aged between 18 and 79 years (median = 29), were recruited via Amazon Mechanical Turk. Participants predominantly reported being from Asia ($N = 395$), with 176 from North America. They were paid $0.10 for completing this short experiment.

TABLE 1
Objective likelihood probabilities and the associated Bayesian predictions for the 17 likelihood probability conditions in Experiment 3

Condition	$P_{obj}(Zhang \mid unsuccessful)$	$P_{obj}(Zhang \mid successful)$	$P_{obj,predicted}(successful \mid Zhang)$
1	.10	.90	0.10
2	.10	.80	0.11
3	.20	.80	0.20
4	.30	.70	0.30
5	.40	.80	0.33
6	.40	.60	0.40
7	.60	.80	0.43
8	.50	.50	0.50
9	.80	.80	0.50
10	.10	.10	0.50
11	.80	.60	0.57
12	.60	.40	0.60
13	.80	.40	0.67
14	.70	.30	0.70
15	.80	.20	0.80
16	.80	.10	0.89
17	.90	.10	0.90

Design. A 17 × 2 (likelihood probability × argument direction) between-participants design was employed. Participants were randomly assigned to 1 of 34 experimental conditions. Participants were told that there had been 10 successful and 10 unsuccessful transportation policies that had been implemented on Planet Xenon. Likelihood probabilities pertaining to how likely Zhang was to have been responsible for successful and unsuccessful policies, $P_{obj}(Zhang \mid successful)$ and $P_{obj}(Zhang \mid unsuccessful)$, were provided in frequency format by showing participants the number of successful and unsuccessful transport policies (out of 10) that Zhang had been responsible for. We used 17 different combinations for $P_{obj}(Zhang \mid successful)$ and $P_{obj}(Zhang \mid unsuccessful)$ (see Table 1). These objective likelihoods, $P_{obj}(Zhang \mid successful)$ and $P_{obj}(Zhang \mid unsuccessful)$, can be used to make predictions for how successful the policy should be judged to be, $P_{obj,predicted}(successful \mid Zhang)$ (see Table 1). Note that here, unlike in Experiment 2, the likelihood probability information was provided at the very start of the experiment. Also, participants were only asked to judge the goodness of the policy, $P_{subj}(successful \mid Zhang)$ (again referred to hereafter as the *posterior rating*), once, whereas in Experiment 2 they were asked to judge this twice. Argument direction refers to how the reference to Zhang was used in the argument. In the "pro" direction condition, Zhang was used to support the argument that the proposed policy would be successful and in the "against" condition, Zhang was used to support the argument that the policy would be unsuccessful. This variable was introduced to balance the design as, unlike Hitler, Zhang is not a character with existing negative connotations, and therefore can potentially be used as positive evidence, as well as negative evidence.

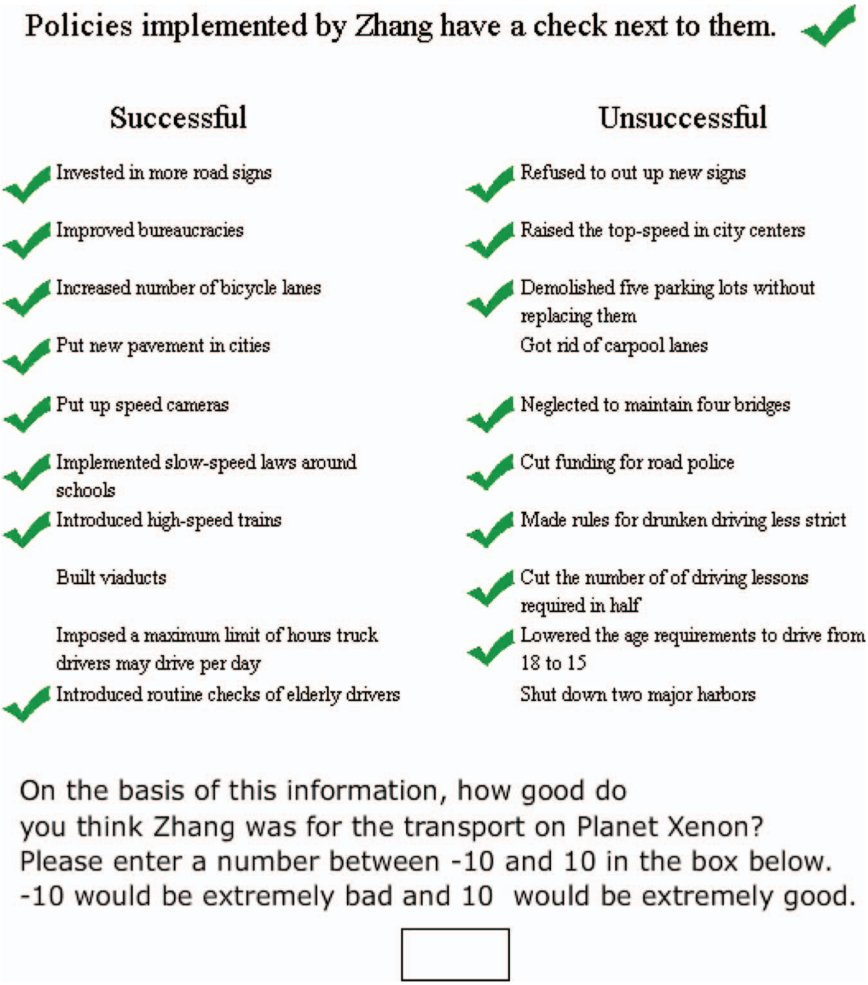
Finally, as in Experiments 1 and 2, and consistent with the Bayesian emphasis on subjective probability, predicted values were calculated from the subjective likelihoods, $P_{subj}(Zhang \mid successful)$ and $P_{subj}(Zhang \mid unsuccessful)$, which were again elicited from the participants. As before, predictions were calculated using Equation 3 and assuming $P(successful) = P(unsuccessful) = .5$.

Materials and procedure. Participants were informed that transport was new on Planet Xenon and there had consequently only been 20 transportation policies implemented to date. They were told that, of these 20 policies, 10 had been successful and 10 unsuccessful and Zhang had been responsible for some of these policies. On a subsequent screen, participants were presented with 20 transportation policies, 10 in a column labelled "Successful" and 10 in a column labelled "Unsuccessful" (see Figure 5). Those that Zhang had implemented were marked with a green check mark. Within each likelihood probability condition, the specific policies that had been implemented by Zhang were randomised across participants. To ensure that participants processed the information, and were given some initial reason for its presence on the screen, participants were asked, on the basis of this information, to indicate how good they thought Zhang had been for transport on Planet Xenon. To minimise the risk of participants simply copying this response in their subsequent posterior ratings, this response was typed as a number between –10 (extremely bad) and 10 (extremely good), instead of using a slider.

On the next screen, participants were provided with an argument in the same format as in Experiment 1. The reference to Hitler was replaced with a reference to Zhang and the argument proponents were introduced as Zeeb and Zorba, two citizens of Planet Xenon. In the "against" condition, all other aspects were identical to Experiment 1. In the "pro" condition, Zorba's (Protagonist B's) assertion that "It's definitely not (a good idea)" was replaced with "It's definitely a good idea." After providing a posterior rating, participants provided subjective likelihood probability ratings for $P_{subj}(Zhang \mid successful)$ and $P_{subj}(Zhang \mid unsuccessful)$, with the question formats following those in Experiment 1.

Results and discussion

As a manipulation check, to test whether participants had processed the likelihood probabilities provided relating to Zhang's successes and failures,



Policies implemented by Zhang have a check next to them. ✓

Successful

- ✓ Invested in more road signs
- ✓ Improved bureaucracies
- ✓ Increased number of bicycle lanes
- ✓ Put new pavement in cities
- ✓ Put up speed cameras
- ✓ Implemented slow-speed laws around schools
- ✓ Introduced high-speed trains
- Built viaducts
- Imposed a maximum limit of hours truck drivers may drive per day
- ✓ Introduced routine checks of elderly drivers

Unsuccessful

- ✓ Refused to out up new signs
- ✓ Raised the top-speed in city centers
- ✓ Demolished five parking lots without replacing them
- Got rid of carpool lanes
- ✓ Neglected to maintain four bridges
- ✓ Cut funding for road police
- ✓ Made rules for drunken driving less strict
- ✓ Cut the number of of driving lessons required in half
- ✓ Lowered the age requirements to drive from 18 to 15
- Shut down two major harbors

On the basis of this information, how good do you think Zhang was for the transport on Planet Xenon? Please enter a number between -10 and 10 in the box below. -10 would be extremely bad and 10 would be extremely good.

□

Figure 5. A screenshot showing the second screen of Experiment 3, presenting the likelihood probabilities to participants. This example corresponds to condition 9 in Table 1.

$P(Zhang \mid good)$ and $P(Zhang \mid bad)$, we first analysed participants' responses to the question, "How good do you think Zhang was for the transport on Planet Xenon?" A 17 × 2 factorial ANOVA revealed the expected main effect of likelihood probabilities, $F(16, 691) = 20.66$, $p < .001$, $\text{eta}_p^2 = .32$. No effect of argument direction was observed, $F(1, 691) = 3.00$, $p = .084$, $\text{eta}_p^2 = .004$, nor was there an interaction between the two variables ($F < 1$). This is as expected, because the two argument direction conditions are identical at this stage of the experiment. Figure 6 shows that as Zhang

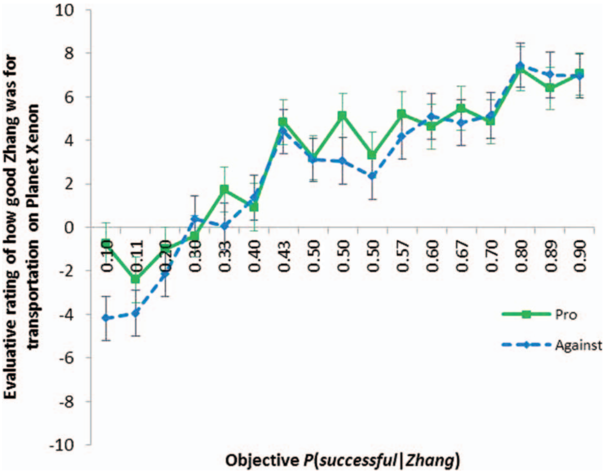


Figure 6. Evaluative perceptions of Zhang before the introduction of the argument in Experiment 3, plotted as a function of $P_{obj,predict}(successful \mid Zhang)$. Error bars are plus and minus 1 standard error.

objectively became better for transportation, so evaluative perceptions increased.

As in Experiments 1 and 2, the main dependent variable of interest was participants' subjective judgements of how good the policy seemed after reading the argument, the *posterior rating*, $P_{subj}(successful \mid Zhang)$. Figure 7 shows these ratings across experimental conditions. As expected under a rational framework, there was a main effect of likelihood probabilities on *posterior ratings*, $F(16, 691) = 5.96$, $p < .001$, $eta_p^2 = .12$, with higher ratings observed for higher values of the objective prediction, $P_{obj,predicted}$ $(successful \mid Zhang)$. There was also a main effect of argument direction, with higher ratings in the "pro" rather than "against" conditions, $F(1, 691) = 79.86$, $p < .001$, $eta_p^2 = .10$. This suggests that a pragmatic component of the argument also affected participants' evaluations of the policy following the argument. There was no interaction between likelihood probability and argument direction ($F < 1$).

As in Experiments 1 and 2, we repeated the above analyses using $P_{subj,predict}(successful \mid Zhang)$ as the dependent variable. The effect of likelihood probabilities was replicated, $F(16, 691) = 22.15$, $p < .001$, $eta_p^2 = .34$, as was the lack of a significant interaction, $F(16, 691) = 1.32$, $p = .18$. Unlike in the posterior rating data, however, there was no main effect of argument direction ($F < 1$). This suggests that a pragmatic component of the argument, responsive to the intentions of the speaker in the dialogue, also affected judgements of the proposal's goodness after receipt of the argument.

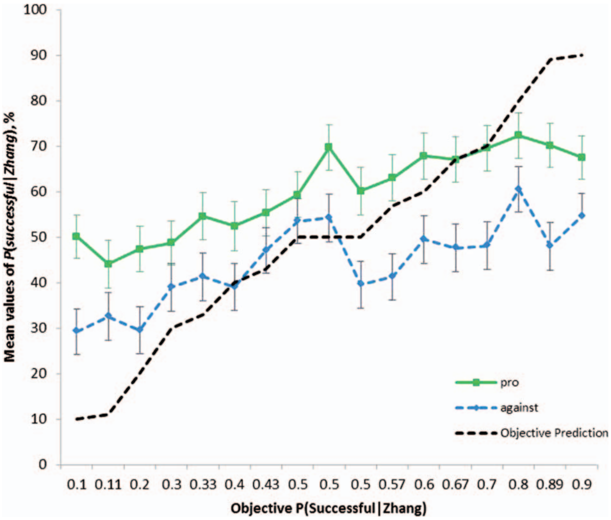


Figure 7. Mean *posterior ratings* for the "pro" and "against" argument direction conditions, plotted as a function of $P_{obj,predict}(successful \mid Zhang)$ in Experiment 3. Note that, so as to show results for every condition, the x-axis is not linear (if it were, the dotted line representing the objective predictions would be a diagonal line). Error bars are plus and minus 1 standard error.

We next sought to determine whether people's posterior ratings could be quantitatively predicted by a Bayesian model based on objective frequency data. As in Experiments 1 and 2, the prior degree of belief was assumed to be .5. In Experiments 1 and 2, we assume that participants bring into the experiment knowledge that suggests that Hitler should be invoked as negative rather than positive evidence for a particular policy (a position supported by the results of Experiment 1). For Zhang, a fictional alien, no such pre-experimental knowledge exists. Rather, at the outset of the experiment we introduce Zhang as a positive or negative individual with respect to the likely success of transportation policies he implemented (the likelihood probability manipulation). A full factorial design, manipulating both likelihood probabilities and argument direction, was required to identify the independent contributions of these two components of the argument in the analysis above. The complete factorial design of our experiment did, however, yield some conditions in which the likelihood probabilities and argument direction were inconsistent. For example, likelihood probability condition 17 could be paired with the "against" argument direction condition. This yields a situation in which participants first learn that Zhang was extremely positive for transport on Xenon, but later read an argumentation dialogue in which he is being invoked as evidence against a policy. It is somewhat unclear how participants should reconcile this inconsistency. In our quantitative analyses, we therefore focused on only those conditions in which the argument direction was consistent with the ratio of likelihood probabilities. We note, however, that the pattern and significance of the correlations reported remains unchanged when all conditions are included.

Likelihood probability conditions 1–7 predict that Zhang should negatively affect the evaluation of the proposed policy and we therefore analyse these responses from the "against" argument direction condition, whilst the converse is true for conditions 11–17 (so we analyse these responses from the "pro" condition). Conditions 8, 9, and 10 had likelihood probability ratios of 1, and we included these conditions in the analysis for both argument directions. Thus there were 10 conditions that contained "sensible" arguments for "pro" and "against" conditions. The correlation between average predictions from objective likelihoods, $P_{obj,predict}$ $(successful\,|\,Zhang)$, and average posterior ratings, $P_{subj}(successful\,|\,Zhang)$, over these "sensible" conditions was $r(18) = .88$, $p < .001$, indicating that 77% of the variance was explained by the "objective" Bayesian prescriptions (see Figure 8).

Due to the Bayesian framework's reliance on subjective probabilities, average values of $P_{subj,predict}(successful\,|\,Zhang)$ in each "sensible" condition were also compared with average posterior ratings in each "sensible" condition (see Figure 8). The resulting correlation coefficient indicated that

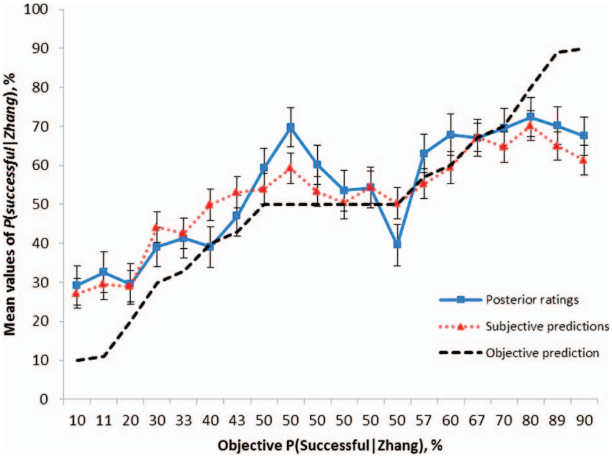


Figure 8. $P_{obj,predict}(successful\,|\,Zhang)$, $P_{subj,predict}(successful\,|\,Zhang)$, and posterior ratings, $P_{subj}(successful\,|\,Zhang)$, averaged over all sensible conditions, where direction of argument is consistent with the given likelihood-values of Zhang implementing good policies vs bad.

the Bayesian account was able to account for 85% of the variance in Experiment 3, $r(18) = .92$, $p < .001$.

GENERAL DISCUSSION

Across three experiments, we have demonstrated that people are sensitive to relevant probabilistic information in their evaluations of the convincingness of a form of the *ad hominem* argument (specifically, an *ad Hitlerum* in Experiments 1 and 2). This probabilistic information is deemed relevant on the Bayesian approach to argumentation (e.g., Hahn & Oaksford, 2007a). Moreover, in all three experiments, good quantitative fits were observed between predictions derived from participants' reported conditional probability ratings and their evaluation of the arguments (operationalised as their posterior ratings). Crucially, these model fits did not benefit from any free parameters that could be tweaked in order to enhance model fits. All the information for calculating the (subjective) predicted ratings was provided by participants, or present and consistent in the experimental materials (prior degree of belief was assumed to equal .5, with information in the experimental materials reinforcing this assumption). Finally, Experiment 3 demonstrated good model fits between a predicted value calculated from objective probabilities provided in the experiment and participants' ratings of the argument's convincingness. Once again, no free parameters were present to enhance the degree of fit between model and data.

Observing positive correlations between the Bayesian predictions and the observed data provides some support for the Bayesian account, but it is possible that a simpler model of participants' belief updating could better explain the results. One plausible simpler model would be for participants to base their judgements solely on either $P(e \mid h)$ or $P(e \mid \neg h)$, rather than integrating these conditional probabilities, as prescribed by Bayes' Theorem (Equation 1). Such a strategy would be akin to that of attribute substitution, in which participants replace one judgement with an accessible alternative judgement, such as conflating $P(A \mid B)$ and $P(B \mid A)$ (see e.g., Eddy, 1982; Kahneman & Frederick, 2002; Villejoubert & Mandel, 2002). Table 2 shows the correlation coefficients observed between participants' posterior ratings, $P_{subj}(h \mid e)$, and the Bayesian predictions, $P_{predict}(h \mid e)$, $P(e \mid h)$, and $P(e \mid \neg h)$. For Experiment 3 this was calculated from both the objective and subjective probabilities.

The first result to note is that posterior ratings of the *ad Hitlerum* arguments presented in Experiment 1 were better predicted by subjective ratings of $P(e \mid \neg h)$—i.e, $P(Hitler \mid bad)$—than by $P(e \mid h)$—$P(Hitler \mid good)$. In Experiment 2, subjecting ratings of $P(Hitler \mid bad)$ were a slightly stronger predictor of the posterior ratings than the full Bayesian model, but overall

TABLE 2
Proportion of variance in participants' posterior ratings, across experimental conditions, accounted for by the Bayesian predictions as compared with the prediction that participants substitute either P(e | h) or P(e | ¬h) as an answer for P(h | e).

	Bayesian (r^2)	*P(e | h)* (r^2)	*P(e | ¬h)* (r^2)	*Likelihood of Bayesian model vs P(e | h) model*	*Likelihood of Bayesian model vs P(e | ¬h) model*
Experiment 1	89%	55%	82%	33.8	3.4
Experiment 2	89%	85%	93%	2.5	0.26
Experiment 3 (subjective)	85%	77%	77%	71.8	71.8
Experiment 3 (objective)	77%	65%	35%	66.6	32497.9

The two right-hand columns demonstrate how much more likely the data is to have been generated by the Bayesian model, rather than each of the simpler models (Glover & Dixon, 2004).

the predictions of *P*(*Hitler* | *bad*), *P*(*Hitler* | *good*), and the full Bayesian model were fairly similar. Because the experimental materials were designed to set up a prior of .5, and participants were provided with frequentist information that was complementary for *P*(*Hitler* | *bad*) and *P*(*Hitler* | *good*), the similar model fits are expected. Notably, in Experiment 3, where the two conditional probabilities provided were no longer complementary, the Bayesian model performs much better as a predictor of the data than either *P*(*Zhang* | *successful*) or *P*(*Zhang* | *unsuccessful*). Moreover, in Experiment 3 the data were not better predicted by *P*(*Zhang* | *unsuccessful*) than by *P*(*Zhang* | *successful*). The stronger correlation with *P*(*Hitler* | *bad*) over *P*(*Hitler* | *good*) observed in Experiment 1 (and also to a small extent in Experiment 2) might be explained by the expectation that Hitler would be used as an argument *against* a proposition, based on participants' negative perceptions of Hitler. Emotional and pragmatic features of the argument might thus exert a certain bias on argument ratings, leading to an over focus on one aspect of the relevant information (as is observed in manifestations of the confirmation bias, e.g., Nickerson, 1998).

How rational were our participants?

Both Experiments 2 and 3 reported results in the analysis of the posterior ratings that were not replicated in the predicted ratings. It is likely difficult to account for all variance in argumentation solely in Bayesian terms, but we note that these particular differences in the results do not necessarily provide evidence against a Bayesian approach to argumentation as either a normative or even as a descriptive theory. Rather, the conditional probabilities entered into the Bayesian predictions might not form a full Bayesian model of the scenario. For example, in Experiment 3 we asked participants the conditional probability questions: ''Of all Xenon transportation policies that were [SUCCESSFUL/UNSUCCESSFUL], how many do you think Zhang was responsible for?'' and used these conditional probabilities to formulate Bayesian predictions. However, participants might have assigned some diagnosticity to the fact that Zorba was arguing that the policy was good in the first place; consequently participants should rationally include this evidence in their evaluation of the policy, as indeed they seem to do.

One aspect of the *ad Hitlerum,* which is amenable to a rational treatment but which was not a feature of the current research, is a consideration of the *amount* of utility associated with an individual's previous policies. In the current experiments, policies were designated as either good or bad (Experiments 1 and 2), successful or unsuccessful (Experiment 3). Such a binary classification is clearly an oversimplification of real-world situations. Indeed, a traditional, rational economic perspective (e.g., von Neumann & Morganstern, 1944) is that outcomes can be classified on a utility continuum, on which zero is neutral, and the degree of negativity or positivity of an outcome is represented by the distance from zero. Presumably, one reason for the ubiquity of the *ad Hitlerum* argument is that the severity of his bad policies (e.g., the Holocaust) greatly outweighed that of his good policies. The arguments presented in the current paper were concerned solely with the likelihood of a policy being either good or bad (as a binary construct), rather than its severity. Consequently considerations of severity were not relevant to the specific arguments that we investigated. Were the person receiving the argument attempting to make up her mind whether or not to declare her support for the proposed policy (i.e., to make a decision to take a particular action), the utilities become relevant from a decision-making perspective. However, without the necessity for a decision to be made, utilities carry no normative weight in assessing likelihoods (see also Hahn & Oaksford, 2007b, for a similar point relating to the role—or lack thereof—of the burden of proof in argumentation). That is not to say, however, that the severity of bad or good proposals might not influence likelihood judgements. Rather, the greater salience or availability of severe outcomes will likely bias people's subjective estimates of the likelihood probabilities (e.g., Bar-Hillel, Budescu, & Amar, 2008; Tversky & Kahneman, 1973) and thus, according to the Bayesian framework, their posterior probability ratings.

It is important to note that, while our work shows that, on average, people perceive these particular *ad hominem* arguments rationally, this does not mean that argument proponents usually *use* the argument rationally. However, if those who are likely to receive the arguments (a random selection of participants from the general population) tend to evaluate them against an appropriate rational standard, the overall effect of the argument will be rational. In this way, rational argument recipients would be protected against unscrupulous, fallacious argument tactics.

Appeal to authority

Throughout this article we have considered the *ad Hitlerum* to be an example of an abusive form of the *ad hominem* argument, according to the definition of Copi and Cohen (1994). Our formalisation, however, is also readily applicable to the appeal to authority. The appeal to authority "uses the opinion of a respected authority or expert on a subject as positive personal argumentation to support one's own side of the argument" (Walton, 2008b, p. 209). This definition of the appeal to authority places it as essentially the opposite argument to the *ad Hitlerum* as we have employed it in our experiments. In the case of the *ad Hitlerum*, the "authority" being referenced is a negative one, used to attack an opponent's standpoint. Neither Copi and Cohen, nor Walton, suggest that the appeal to authority is ostensibly a fallacious argument. As Copi and Cohen recognise, consulting an appropriate authority (e.g., a medical doctor) may be the only option we have available to us in a number of subject matters. The argument does, however, become fallacious when the appeal is made to "parties having no legitimate claim to authority in the matter at hand" (Copi & Cohen, 1994, p. 119), when it becomes the fallacy of *ad Verecundiam*. The Bayesian approach provides a framework within which fallacious and non-fallacious appeals to authority can be distinguished (see also Hahn, Oaksford & Harris, in press). An expert source can be considered to be one who is more likely to provide evidence in support for a position if it is true, than if it is false. That is, the likelihood ratio corresponding to evidence from an expert source, $\frac{P(e\,|\,h)}{P(e\,|\,\neg h)}$, is very high, while a non-expert source is likely to have a likelihood ratio close to 1 (see also Hahn et al., 2009; Harris, Corner & Hahn, 2009). This is analogous to the effect of $\frac{P(Hitler\,|\,good)}{P(Hitler\,|\,bad)}$ observed in the present experiments. The only difference between the *ad Hitlerum* arguments we use and appeals to authority lies in the fact that the *ad Hitlerum* is employed as a negative argument and therefore becomes stronger as the likelihood ratio approaches zero,[13] but it is still the case that the argument becomes more persuasive as the likelihood ratio diverges from 1.

More generally, and as stated elsewhere (e.g., Hahn et al., 2009; Hahn & Oaksford, 2007; Harris et al., 2009), Bayesian probability provides a rational framework within which multiple aspects of argumentation can be

[13] Note that the positive arguments in Experiment 3 are thus better classified as arguments from authority rather than *ad hominem* arguments.

understood, including the role of source characteristics. Formalisations within this framework enable normative predictions to be made across a variety of contexts, regardless of whether a given context might be viewed as sufficiently distinct to merit classification as a qualitatively distinct argument form (see e.g., Hoeken, Timmers, & Schellens, 2012 this issue) or not (see also Hahn & Oaksford, 2006b). Human performance can then subsequently be evaluated against these predictions.

Conclusion

Across three experiments, quantitative evaluations of the likely ‘‘goodness’’ of a proposal following the receipt of an *ad hominem* argument (the *ad Hitlerum* in Experiments 1 and 2, an ‘‘*ad Zhangum*’’ in Experiment 3) were well predicted by a Bayesian model. This provides the most direct quantitative evidence in support of the Bayesian framework of argumentation (Hahn & Oaksford, 2006a, 2007a) and suggests that, when compared against the appropriate normative model, people's reasoning might be more rational than has often been assumed (see also, e.g., Oaksford & Chater, 1994).

We acknowledged, however, that our model was not able to capture all the variance in people's argument evaluations. This might be a limitation of a complete rational model of argumentation, but it might equally result from our failure to include all relevant probabilistic components in our calculation of the normative posterior degrees of belief. Determining the limits of the rational model is a key area for future research. The Bayesian framework, however, provides a sensible normative model, sensitive to argument *content,* which is necessary before such research can be undertaken.

REFERENCES

Bar-Hillel, M., Budescu, D. V., & Amar, M. (2008). Predicting World Cup results: Do goals seem more likely when they pay off? *Psychonomic Bulletin and Review*, *15*, 278–283.

Birnbaum, M. H., & Stegner, S. E. (1979). Source credibility in social judgement: bias, expertise, and the judge's point of view. *Journal of Personality and Social Psychology*, *37*, 48–74.

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

Copi, I. M., & Cohen, C. (1994). *Introduction to logic* (9th ed.). New York: Macmillan.

Corner, A., & Hahn, U. (2009) Evaluating science arguments: Evidence, uncertainty, and argument strength. *Journal of Experimental Psychology: Applied*, *15*, 199–212.

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

Corner, A., Harris, A. J. L., & Hahn, U. (2010). Conservatism in belief revision and participant skepticism. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 1625–1630). Austin, TX: Cognitive Science Society.

Eddy, D. M. (1982). Probabilistic reasoning in clinical medicine: Problems and opportunities. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgement under uncertainty: Heuristics and biases* (pp. 249–267). Cambridge, UK: Cambridge University Press.

Evans, J. St. B. T., & Over, D. E. (2004). *If*. Oxford, UK: Oxford University Press.

Gigerenzer, G. (2002). *Reckoning with Risk: Learning to live with uncertainty*. London: Penguin.

Gigliotti, S., & Lang, B. (Eds.). (2005). *The Holocaust: A reader*. Oxford, UK: Blackwell.

Glover, S., & Dixon, P. (2004). Likelihood ratios: A simple and flexible statistic for empirical psychologists. *Psychonomic Bulletin & Review*, *11*, 791–806.

Godwin, M. (1994). Meme, counter-meme, *Wired* 2: 10. http://www.wired.com/wired/archive/2.10/godwin.if.html

Griffiths, T. L., & Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychological Science*, *17*, 767–773.

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. (2006a). A Bayesian approach to informal reasoning fallacies. *Synthese*, *152*, 207–223.

Hahn, U., & Oaksford, M. (2006b). A normative theory of argument strength. *Informal Logic*, *26*, 1–24.

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

Hahn, U., & Oaksford, M. (2007b). The burden of proof and its role in argumentation. *Argumentation*, *21*, 39–61.

Hahn, U., Oaksford, M., & Bayindir, H. (2005). How convinced should we be by negative evidence? In B. Bara, L. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the 27th Annual Conference of the Cognitive Science Society* (pp. 887–892). Mahwah, NJ: Lawrence Erlbaum Associates Inc.

Hahn, U., Oaksford, M., & Harris, A. J. L. (in press). Testimony and argument: A Bayesian perspective. In F. Zenker (Ed.). *Bayesian argumentation*. Dordrecht, The Netherlands: Springer.

Harris, A., Corner, A., & Hahn, U. (2009). "Damned by faint praise": A Bayesian account. In N. A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 292–297). Austin, TX: Cognitive Science Society.

Harris, A. J. L., & Hahn, U. (2009). Bayesian rationality in evaluating multiple testimonies: Incorporating the role of coherence. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*, 1366–1372.

Hoeken, H., Timmers, R., & Schellens, P. J. (2012 this issue). Arguing about desirable consequences: What constitutes a convincing argument? *Thinking & Reasoning*, *18*, 394–416.

Howell, D. C. (1997). *Statistical methods for psychology* (4th ed.). Belmont, CA: Duxbury Press.

Howson, C., & Urbach, P. (1996). *Scientific reasoning: The Bayesian approach* (2nd edition). Chicago, IL: Open Court.

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgement. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgement* (pp. 49–81). Cambridge, UK: Cambridge University Press.

Korb, K. (2004). Bayesian informal logic and fallacy. *Informal Logic*, *23*(2), 41–70.

Leitgeb, H., & Pettigrew, R. (2010a). An objective justification of Bayesianism I:Measuring inaccuracy. *Philosophy of Science*, *77*, 201–235.

Leitgeb, H., & Pettigrew, R. (2010b). An objective justification of Bayesianism II: The consequences of minimizing inaccuracy. *Philosophy of Science*, *77*, 236–272.

McGlynn, K. (2010). Wtf! 11 people who have been unfairly compared to Hitler. *Huffington Post*. http://www.huffingtonpost.com/2010/06/10/wtf-people-who-have-unfai_n_606810.html#s98403&title=Barack_Obama

Moser, S. (2006). Penn State Trustee compares Reagan to Hitler. *Accuracy in media*. http://www.academia.org/penn-state-trustee-compares-reagan-to-hitler/

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*, 175–220.

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, *101*, 608–631.

Oaksford, M., & Chater, N. (1998). *Rationality in an uncertain world: Essays on the cognitive science of human reasoning*. Hove, UK: Psychology Press.

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. (in press). Why are we convinced by the ad hominem argument? Bayesian source reliability and pragma-dialectical discussion rules. In F. Zenker (Ed.), *Bayesian argumentation*. Dordrecht, The Netherlands: Springer.

Over, D. E., Hadjichristidis, C., Evans, J. St. B. T., Handley, S. J., & Sloman, S. A. (2007). The probability of causal conditionals. *Cognitive Psychology*, *54*, 62–97.

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgement and Decision Making*, *5*, 411–419.

Pettigrew, R. (in press). Accuracy, chance, and the Principal Principle. *Philosophical Review*.

Rescher, N. (2005). Reductio ad absurdum. *Internet Encyclopedia of Philosophy*. Retrieved 2 June 2011, from http://www.iep.utm.edu/reductio/

Schneider, N. K., & Glantz, S. A. (2008) "Nicotine Nazis strike again": A brief analysis of the use of Nazi rhetoric in attacking tobacco control advocacy. *Tobacco Control*, *17*, 291–296.

Schum, D. A. (1981). Sorting out the effects of witness sensitivity and response-criterion placement upon the inferential value of testimonial evidence. *Organizational Behavior and Human Performance*, *27,* 153–196.

Strauss, L. (1953). *Natural right and history*. Chicago, IL: University of Chicago Press.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, *5*, 207–232.

Van Eemeren, F. H., Garssen, B. & Meuffels, B. (2009). *Fallacies and judgements of reasonableness: Empirical research concerning the pragma-dialectical discussion rules*. Dordrecht: Springer.

Van Eemeren, F. H., Garssen, B. & Meuffels, B. (2012 this issue). The disguised abusive *ad hominem* empirically investigated strategic manoeuvring with direct personal attacks. *Thinking & Reasoning*, *18*, 344–364.

Van Eemeren, F. H., & Grootendorst, R. (2004). *A systematic theory of argumentation: The pragma-dialectical approach*. Cambridge, UK: Cambridge University Press.

Villejoubert, G., & Mandel, D. R. (2002). The inverse fallacy: An account of deviations from Bayes's theorem and the additivity principle. *Memory and Cognition*, *30*, 171–178.

Von Neumann, J., & Morganstern, O. (1944). *Theory of games and economic behavior*. Princeton, NJ: Princeton University Press.

Vul, E., & Pashler, H. (2008). Measuring the crowd within: Probabilistic representations within individuals. *Psychological Science*, *19*, 645–647.

Wallsten, T. S., Budescu, D. V., Erev, I., & Diederich, A. (1997). Evaluating and combining subjective probability estimates. *Journal of Behavioral Decision Making*, *10*, 243–268.

Walton, D. (1995). *A pragmatic theory of fallacy*. Tuscaloosa, AL: University of Alabama Press.

Walton, D. (2008a). *Witness testimony evidence: Argumentation, artificial intelligence, and law*. Cambridge, UK: Cambridge University Press.

Walton, D. (2008b). *Informal logic: A pragmatic approach* (2nd ed.). Cambridge, UK: Cambridge University Press.