UNIVERSITY of CALIFORNIA PRESS | Collabra: Psychology

Barbosa, S., & Jiménez-Leal, W. (2022). Different Algorithmic Models Underlie Virtue and Vice Attributions. *Collabra: Psychology*, *8*(1). https://doi.org/10.1525/collabra.38032

Cognitive Psychology

# Different Algorithmic Models Underlie Virtue and Vice Attributions

Sergio Barbosa[1] [a], William Jiménez-Leal[2]

[1] Escuela de Medicina y Ciencias de la Salud, Universidad del Rosario, Bogotá, Colombia, [2] Departamento de Psicología, Universidad de los Andes, Bogotá, Colombia

Keywords: Moral Judgment, Character Judgment, Trait Attribution, Bayesian Models, Algorithmic Models

https://doi.org/10.1525/collabra.38032

Collabra: Psychology

Vol. 8, Issue 1, 2022

One of the most pressing questions in social psychology is how people update character attributions about other people considering novel information. A possible way to tackle this question is to algorithmically model trait attribution updating and confront it to how people actually update character attributions. Here, we present, parameterize, and empirically test several Bayesian and averaging models of character-based moral judgment over multiple pieces of morally relevant or distractor information. Taken as a whole, results from two experiments suggest that virtue and vice attributions follow different algorithms. Depending on the structure of received information virtue and vice attributions can follow differently weighted Bayesian algorithms or average-based models. We discuss these results in light of both classic findings in moral psychology and cognitive sciences in general.

## 1. Introduction

Recently, a call for the formalization of cognitive theories of morality has been made (Crockett, 2013, 2016a, 2016b). This has been echoed by numerous studies testing cognitive models of different aspects of morality such as emotion and theory of mind (Chris L. Baker et al., 2005, 2008, 2009, 2011; Baker & Tennenbaum, 2014; Kleiman-Weiner et al., 2017; Kleiman-Weiner & Levine, 2015; Ong et al., 2019; Pöppel & Kopp, 2019; Ullman et al., 2009), common sense morality (Hagmayer & Osman, 2012; Kim et al., 2018; Kleiman-Weiner et al., 2017; Yu et al., 2019), group decision-making (Khalvati et al., 2019), implicit moral judgment (Cameron et al., 2017) and and its explicit counterpart (Brand, 2015; Bretz & Sun, 2018; Cao et al., 2019; Cushman, 2013) among others.

A special place among these processes is occupied by moral judgement updating in face of new information (Monroe & Malle, 2018; Okten et al., 2019; Siegel et al., 2018). Monroe and Malle (2018) show that blame judgments are updated roughly equally when novel information makes blame judgments stronger or lesser. However, we are interested here in a particular kind of moral judgement different from blame attributions (Malle, 2020): character-based attributions. Contrary to Monroe and Malle (2018), Siegel et al (2018) show that virtue and vice attributions are not equally volatile. Their results show that vice attributions are easier to change with new information whereas virtue attributions are comparatively more fixed and relatively impervious to new information. Also, broad character attributions such as "good person" are more easily updated than narrower character attributions like "honest" or "aggressive" (Okten et al., 2019). Taken together these results suggest that moral judgment is generally liable to updating according to new information but may differ according to what specific type of moral judgment is considered. Situational judgments such as blame or praise updating relatively easily while character attribution updating could differ according to the type (i.e. broad vs to narrow traits) of attribution. Although processes relevant to moral judgement have been the object of cognitive modelling (Cameron et al., 2017; Hagmayer & Osman, 2012; Kim et al., 2018; Kleiman-Weiner et al., 2017; Siegel et al., 2018; Yu et al., 2019), there is a comparative lack of models for character based moral judgment updating when contrasted with the rich landscape of behavioural findings in this area. Consequently, we propose and test a number of algorithmic models of how moral judgment aimed at an agent's moral character, both virtuous and vicious, changes according to new information.

Character-based moral judgment is a type of moral judgment aimed at an agent's character rather than to her actions or other features of moral reasoning (Malle, 2020). This type of moral judgment has called for a growing number of studies (Fleeson et al., 2014; Hartley et al., 2016; Helzer & Critcher, 2015; Pizarro & Tannenbaum, 2012; E. L. Uhlmann et al., 2013, 2014, 2015). People are willing to say that an agent is morally virtuous even though she acted

a Correspondence concerning this article should be addressed to Sergio Barbosa, Universidad del Rosario. E-mail: sergio.barbosad@urosario.edu.co

wrongly or vice versa, which suggests a qualitative difference between moral judgment about an action and about character (E. L. Uhlmann et al., 2013, 2014; E. L. Uhlmann & Zhu, 2014). Character is perceived to be more important than personal preferences, physical aspect or warmth and competence attributions in person and identity perception (Goodwin et al., 2014, 2015; Newman et al., 2014; Strohminger et al., 2017; Strohminger & Nichols, 2014). Coupled with this perceived importance, the folk concept of moral character is complex and distinguishes several types of morally relevant traits (Landy et al., 2016; Melnikoff & Bailey, 2018; Piazza et al., 2014) and recognizes their interaction with qualitatively different social roles (Barbosa & Jiménez-Leal, 2020). However, despite growing attention to moral character there is surprisingly little research on how these attributions change with new information and how this process could be modelled.

## 1.1. Cognitive models of character-based moral judgment

Cognitively, character-based moral judgment can be modelled as a case of uncertain inference where an unobservable characteristic of the target agent (i.e. his virtuous or vicious character, hypothesized to be relatively stable over time and across situations) is inferred based on directly observable, noisy and sequentially available pieces of information (i.e. observed behaviours such as whether the agent steals a quarter that someone just dropped in front of them, whether they freely share their own knowledge about a subject…). Any piece of information about an agent's character is not directly diagnostic of virtuous or vicious character for several reasons. First, successfully *appearing* to possess a virtuous character without actually incurring in its costs would result in an individual having willing cooperating partners without actually incurring in the cost of cooperation (Batson et al., 1999; Batson & Thompson, 2001) which makes it highly desirable to *feign* being virtuous. Second, any and all behaviours can be heavily influenced by situational factors (Doris, 2005; Harman, 2003, 2009) which makes any single behaviour a highly noisy signal. Therefore, to accurately assess an agent's character several independent observations have to be observed and aggregated somehow to arrive at a relatively stable and hopefully accurate attribution (Fleeson, 2004; Fleeson et al., 2014). Finally, any behaviour can be diagnostic for several traits of character. For instance, being faithful to one's partner might be diagnostic of the virtues of honesty and loyalty but might also be diagnostic of the vice of cowardice if one does not actually try to cheat out of the fear of being caught even if one has the ardent desire to. Hence, an inference procedure is necessary to make sense of these disparate, noisy and sequential pieces of information about a specific agent (Alves & Mata, 2019). To address this necessity, several algorithmic models of trait attribution have been proposed. These models rely on simple arithmetical additions or weighted averages of behaviours values to attribute character traits to a target agent. In this view, character attribution amounts to averaging each behaviour's diagnostic value for the considered trait to reach a final attribution. In turn, this attribution is updated with every new relevant behaviour. These algorithmic models follow classic primacy and Recency effects in long term memory literature (Baddeley & Hitch, 1993).

### *1.1.1. Bayesian models of character-based moral judgments attribution*

Previous trait attribution models have been proposed for general (i.e. both moral and non-moral) trait attribution rather than for the special case of moral trait attribution. These models are based on averaging the value of every observed behaviour for the target trait and attributing the trait based on how high this average is. To account for Recency and primacy effects on memory, these models overweight the first and/or last received information (N. H. Anderson, 1961; N. H. Anderson & Barrios, 1961; Birnbaum, 1972, 1973). Since the pressure and difficulty to appear moral is different from the pressure to appear to possess any other non-moral trait it is doubtful that a single class of models can account for both moral and non-moral attributions. Indeed, as mentioned, an agent's social success heavily hinges on being *perceived* as being willing to cooperate (i.e. being considered virtuous) instead of not (i.e. being considered vicious) but without actually incurring in the costs of cooperating. Thus, agents are particularly motivated to *appear* moral without actually *being* moral (Batson et al., 1999; Batson & Thompson, 2001). Consequently, accurately inferring moral traits among noisy, situation-dependent behaviours is both an especially difficult and especially important task in a way that does not apply to non-moral traits such as those relating to warmth and competence. Therefore, algorithms used to attribute *moral* traits may favour accuracy instead of frugality while algorithms attributing *non-moral* traits might favour frugality over accuracy since misattributing non-moral traits (i.e. competence or warmth) implies a smaller risk than misattributing moral character. Consequently, cognitively complex but accurate Bayesian models appear to be ideal candidates for moral character attributions (Siegel et al., 2018) while cognitively simpler, average-based models might be used for the less important task of attributing non-moral traits.

Research on moral judgment suggests that people are sensitive to the perceived frequency of behaviour, taking less frequent behaviours as more diagnostic of an agent's character than more frequent behaviours (Bear & Knobe, 2017; Brand & Oaksford, 2015; Gray & Keeney, 2015; Monroe & Malle, 2018; Shenhav & Greene, 2010). Also, moral character judgment is not a one-off attribution but takes into account the agent's prior actions by influencing attributed mental states such as intention to cause harm (Kliemann et al., 2008; Mende-Siedlecki, Baron, et al., 2013; Mende-Siedlecki, Cai, et al., 2013; Mende-Siedlecki & Todorov, 2016). Considering the base rate probabilities of specific behaviours and prior attributions about the same target agent is consistent with a Bayesian framework of character-based moral judgment. Hence, following the budding literature of Bayesian models of social and causal cognition (Chris L. Baker et al., 2005, 2011; Baker & Tennen-

baum, 2014; Doya, 2007; Fenton et al., 2013; Griffiths et al., 2008; Jacobs & Kruschke, 2011; Khalvati et al., 2019; Moutoussis et al., 2014; Pöppel & Kopp, 2019; Ullman et al., 2009), we propose several Bayesian Belief Updating models of trait based moral judgment.

Our proposed Bayesian models assume that trait attribution amounts to the computation of the posterior probability of the target agent possessing a target trait given that he carried out a number of observed behaviours (*P(Tr|B)*). Trait attribution requires using Bayes' rule to determine this posterior probability based on prior probabilities of target behaviours (*P(B)*) and traits (*P(Tr)*) and their conditional probability (*P(B|Tr)*). Thus, the neutral Bayesian model directly follows Bayes' rule (EQ 1). Here, we assume that attributions are continuous variables.

$$EQ1:\ P\left(Tr|B\right) = \frac{\left[P\left(B|Tr\right) * P(Tr)\right]}{P(B)}$$

This Bayesian "neutral" model does not distinguish between Virtue and Vice attribution and follows research suggesting that blame and praise judgments are updated similarly (Monroe & Malle, 2019).

Multiple studies highlight motivational influences that can determine moral judgment (Ditto et al., 2009). Motivation in moral trait attribution could be understood in at least three senses. First, given evolutionary pressures that heavily penalized failing to identify a vicious person, people could be "intuitive prosecutors" actively looking for incriminating evidence against other agents and thus readily attributing vicious character while seldom attributing virtuous character (Alicke, 2001). Compatible results in negativity bias (Rozin & Royzman, 2001) suggest that people more easily pay attention to, memorize and recall negative information compared to neutral or positive information. Also, it appears that updating impressions in face of negative information is more difficult than updating in face of positive information (Kappes & Sharot, 2019) and that blame judgments are more extreme than praise judgments (Guglielmo & Malle, 2019) which is compatible with the "intuitive prosecutor" view. Algorithmically, this view could be modelled by "prosecutor" models overestimating the probability of a person possessing a vice but not the probability of them possessing a virtue.

Second, people could be motivated to quickly determine a target agent's character, regardless of whether it is positive or negative, in order to avoid opportunity costs. Research showing that moral judgments are done extremely quickly (Decety & Cacioppo, 2012) and that proto-moral considerations are present very early in human development (Hamlin, 2013; Surian et al., 2018) are compatible with the idea that humans quickly and intuitively form an opinion on agent's character. Algorithmically, this could be modelled by "quick" models that equally overestimate the probabilities of people possessing either virtues or vices, thereby arriving at quicker decisions with minimal information. These models would simply accelerate character attributions by overestimating the probabilities of both virtue and vice attributions.

Third, adaptively we could be more attuned to attribute virtues due to the potentially higher costs of failing to recognize a virtuous, cooperative partner and that virtue attributions are less liable to change according to new information (Siegel et al., 2018). Algorithmically, this motivation could be modelled by "optimist" models overestimating the probabilities of a person possessing a virtue but not overestimating vice probabilities.

To capture these alternative motivations, we fitted three type of Bayesian models reflecting differences in the way information could be assessed: *prosecutor*, *quick* and *optimist* Bayesian models. Specifically, *prosecutor* Bayesian models overestimate high probabilities exclusively for vice attributions and reflect negativity bias in memory, attention and moral cognition. On the contrary, *optimist* models overestimate high probabilities for virtue attribution exclusively, reflecting motivation to attribute virtuous character not to miss out on the benefits of cooperation. Finally, *quick* models overestimate high probabilities for both virtue and vice attributions reflecting motivation to arrive at quick conclusions about a target agent's virtuous and vicious character.

Overestimation of high probabilities for Prosecutor, Optimist and Quick models was achieved by applying an S-shaped transformation (Prelec, 1998) to the neutral Bayes model's posterior probability estimations. First, we fitted the Bayesian neutral model as described above. After fitting that model, we modified its posterior probabilities using the S-shaped transformation presented in EQ2. To reflect true functional form of EQ2 rather than arbitrary values for $\alpha$ and $\beta$ we optimized each free parameter for each participant using the optim() function in base R (R Core Team, 2021). This procedure allows us to estimate an optimized value of both $\alpha$ and $\beta$ for each participant based on their own observed data thereby testing functional form of each model rather than arbitrary parameter values. This procedure was applied to Bayesian model in all conditions to create Quick model, in the Virtue condition, but not on all others to create the Optimist model and in the Vice condition but not all others to create the Prosecutor model. A similar procedure was followed for competing models, namely, Recency, Primacy and U models (see below. All supporting information is in the OSF page of the project).

$$EQ\ 2: t(p) = \exp\left(-\beta \times \left(-\log(p)\right)^{\alpha}\right)$$

### *1.1.2. Competing models of trait attribution*

Our main hypothesis is that the way moral trait attribution changes according to new information follows a Bayesian algorithm operationalized above. To contrast this hypothesis, we fitted four, competing, averaged-based models and a random model. All average-based models used a similar function, different weighted averages of target behaviours prior probabilities to attribute traits. Following classic literature on trait attribution and long-term memory literature (N. H. Anderson, 1961; N. H. Anderson & Barrios, 1961; Baddeley & Hitch, 1993; Birnbaum, 1972, 1973). We computed four types of averaging models. A Primacy model overweighs the first perceived behaviour, Recency model overweighs the last one, a U model unites both Recency and Primacy models by overweighting both the first and last perceived behaviours and an Average model

equally weighs them all. Recency, Primacy and U models follow the same rationale as Bayesian models and assign optimized weights to the first and/or last presented behaviour. As for Bayesian models described above, free parameters were optimized for each participant based on their observed data using the optim() function in base R (R Core Team, 2021). In the case of competing models, free parameters correspond to the weight, *w,* assigned to each presented information. Hence, for these models trait attribution, $\alpha$, corresponds to a weighted average of all received informations overweighting the first information (Primacy model, EQ 3), the last information (Recency model, EQ 4) or both the first and last received information (U model, EQ 5). In EQs 3 through 5 $\alpha_n$ is the trait attribution at iteration n, parameters *a* through *h* are the values corresponding to presented behaviours (see pilot study below) and *w* is the optimized weight assigned to each overweighed information. *w* parameter is the only free parameter and is optimized according to each participant's observed data.

$$EQ\ 3: Primacy\ model.\ \alpha_n = \sum_{1:n} a.w + b + c \ldots h$$

$$EQ\ 4: Recency\ model.\ \alpha_n = \sum_{1:n} a + b + c \ldots h.w$$

$$EQ\ 5: U\ model.\ \alpha_n = \sum_{1:n} a.w + b + c \ldots h.w$$

All Bayesian and competing models were confronted to empirical data to determine which model best reflects how character-based moral judgment changes with new information. Model selection was carried out using fit and generalizability indexes as well as Bayes Factors for each model against the null model (BF10) and against the model with the best fit and generalizability (BF01).

## 2. Pilot Study: Prior Determination

In order to properly parameterize Bayesian and competing models, a relevant set of priors for each behaviour and trait must be computed. Study 1 aimed to determine these in a student sample. We used a set of 116 behaviours associated with virtuous character, vicious character or an amoral trait (being boring). All behaviours and traits were chosen based on previous validation for a student sample (reference redacted for peer review) and will be used in Studies 1 and 2 (Barbosa & Jiménez-Leal, 2020).[1] To ensure out-of-sample prediction we ran separate prior determination studies for Studies 1 and 2. They followed exactly the same methodology, so they will only be described once.

### 2.1. Methods

#### *2.1.1. Participants*

For the model parametrization of study 1 we recruited 54 participants (20 women, 2 did not identify with these genders/ would rather not say, $M_{age}$ = 20.56, $SD_{Age}$ = 2.6) to complete an online survey. Model parametrization for Study 2 included 57 participants (36 females, 6 did not identify with these genders/ would rather not say, $M_{age}$ = 21.28, $SD_{Age}$ =3.7). All participants were recruited through social media affiliated with Universidad de los Andes (Bogotá - Colombia).

#### *2.1.2. Design, materials, and procedure*

In order to obtain relevant priors for each behaviour and target trait and their conditional probability, participants were presented with a random set of 98 behaviours and were asked to provide probabilities of that behaviour happening at all (i.e. *P*(B) ) and of it happening given that the agent is or is not a virtuous/ vicious/ boring person (i.e. *P*(B|Tr) and *P*(B|¬Tr) ). Also, participants provided base rate probabilities for any person possessing or not possessing a virtuous vicious and boring character (i.e. P(Tr) and P(¬Tr)).

All questions followed the same format, asking how many people out of 100 performed the target behaviour or possessed the target trait (i.e. Out of 100 people, how many [(don't)possess target trait/ perform target behaviour]. Conditional probabilities followed a similar structure (i.e. Out of 100 (non)virtuous/ (non)vicious/ (non)boring persons, how many [perform target behaviour]?). All responses were averaged and adjusted to fit a 0 to 1 scale to be used in model parametrization. Participants used a non-numbered slider anchored in the middle to give out their estimations.

### 2.2. Results

This procedure resulted in a comprehensive list of prior probabilities for all behaviours and traits (i.e. P(B), P(¬B), P(Tr) and P(¬Tr) respectively) as well as conditional probabilities of all behaviours given that the agent (does not) possess the target trait (i.e. P(B|T) and P(B|¬T) respectively). These will be used in studies 1 and 2 described below.

## 3. Study 1

Study 1 aimed to directly test described algorithmic models of character-based moral judgment to empirical data.

### 3.1. Methods

#### *3.1.1. Participants*

We recruited 187 participants (99 women, 3 did not identify with either male or female/ would rather not say, age = 20.08, *SD* = 1.96) randomly assigned to one of three experimental conditions. Experimental conditions were defined

1 See behaviour and trait determination procedures, final materials, data analysis scripts and raw data for all studies in https://osf.io/xsv3e/?view_only=07ef8777795b4990855c02722624d690 .

by the type of character attribution elicited: Virtue (i.e. being a good person), Vice (i.e. being a bad person), or non-moral (i.e. being a boring person).

### *3.1.2. Design, materials, and procedure*

Through an online survey, we presented participants with a random set of seven pieces of information about the same fictitious target agent. Randomization of the presented items was done following a two-step procedure (see fig 1). First, to ensure participants had equal chances to receive positive, negative and neutral information, we randomized which information they received, namely target behaviours (i.e. virtuous behaviour in the virtue condition or vicious behaviour in the vice condition) or distractor (i.e. boring behaviours in both conditions) information. Next, we randomized which specific behaviour was presented to participants out of the 26 virtuous behaviours, 18 vicious behaviours and 20 amoral behaviours. This randomization procedure presents several advantages. First, it ensures that all participants had identical chances to receive target or distractor information. Second, it allowed us to limit potential ceiling effects due to excessively congruent information.[2] Finally, it yields a unique set of behaviours for each participant which allows us to examine proposed models independently of the specific presented behaviours, ensuring better out-of-sample prediction. After each behaviour participants were asked to provide character-based moral judgment about the target agent considering all available information about them using a 0 to 100 slider (0 being "The agent is not a good/ bad/ boring person at all" and 100 being "The agent is totally a good/ bad/ boring person"). This procedure was carried out 7 times, once for each presented behaviour (see fig 1). In order to ensure identical base rates across conditions we presented all participants in all conditions with the same morally neutral fictitious target agent (Andrew, a university student who likes football and cooking) prior to all information presentation. Since information in T0 was identical for all participants and conditions, attributions based on this information were excluded from data analysis.

### *3.1.3. Model parameterization*

For every participant's unique set of randomly presented pieces of information we ran all described algorithmic models simulating participant responses. These responses were compared to empirical data for each participant to determine which model better reflected how participant's trait attributions changed with every new piece of information.

As described above, with every presented behaviour Bayesian models updated their trait attribution following EQ 1. After computing the posterior probability considering the new information, an S-shaped transformation was applied to the computed posterior for *quick*, *prosecutor* and *optimist* models (see EQ 2). This process was iterated for every presented behaviour. We optimized $\alpha$ and $\beta$ free parameters for each participant on each iteration based on their observed data. Similarly, every competing model (Average, Recency, Primacy and U models) computed weighted averages of all available behaviours for every new behaviour, assigned weight for each participant and each iteration was also optimized based on their observed data.

## 3.2. Results

We computed a series of linear regressions where each algorithmic model predicted observed data, controlling for condition and iteration. Next, we computed fit (RMSE) and generalizability (BIC, AIC and ICOMP) indexes for each linear model across conditions. Overall, the better the fit and generalizability indexes, the more likely it is that the candidate model reflects the cognitive operations underlying character based moral judgment. Fit index (RMSE) varies from 0 to 1 and reflects the unexplained error not captured by the model. A RMSE value closer to 0 reflects a better fit.

However, when considering models with complex functional forms or multiple parameters, fit indexes risk overfitting, that is, capturing variance that is not explained by the underlying function but by statistical noise. To address this limitation literature suggests generalizability indexes (AIC, BIC and ICOMP) (Pitt et al., 2002; Pitt & Myung, 2002) because they reflect how well a given model generalizes to unknown data sets reflecting the same cognitive process irrespective of the complexity of considered models by penalizing fit indexes according to the model's functional complexity (e.g. linear vs quadratic models) and number of free parameters, thereby reducing the chance of overfitting and maximizing the chance of estimating the "true" underlying function. Unlike fit indexes, generalizability indexes follow an arbitrary scale unique to every dataset which make it difficult to interpret them across data sets. However, they do allow comparison to models fitted from the same data set where model with the smallest generalizability value are preferred.

However, there are some reasons not to take generalizability indexes at face value. First, generalizability indexes follow an arbitrary, data set-specific scale which troubles comparison between different models, even if they were fitted using the same data set. Indeed, the theoretical importance of, say, a 0.5 difference in AIC between two candidate models is highly dependent on considered data set and cannot directly be interpreted across them. Therefore, we computed Bayes Factors (BF10) for each model compared

2 Target or distractor behaviours in the virtue and vice conditions were determined through pilot study results. Behaviours whose marginal probability ($P(B|Tr)$ with virtue or vice was below 0.1 were considered distractors. Behaviours whose prior probability for virtue or vice was above 0.1 were considered target for virtue or vice attribution. No behaviour was strongly associated with both virtues and vices simultaneously.

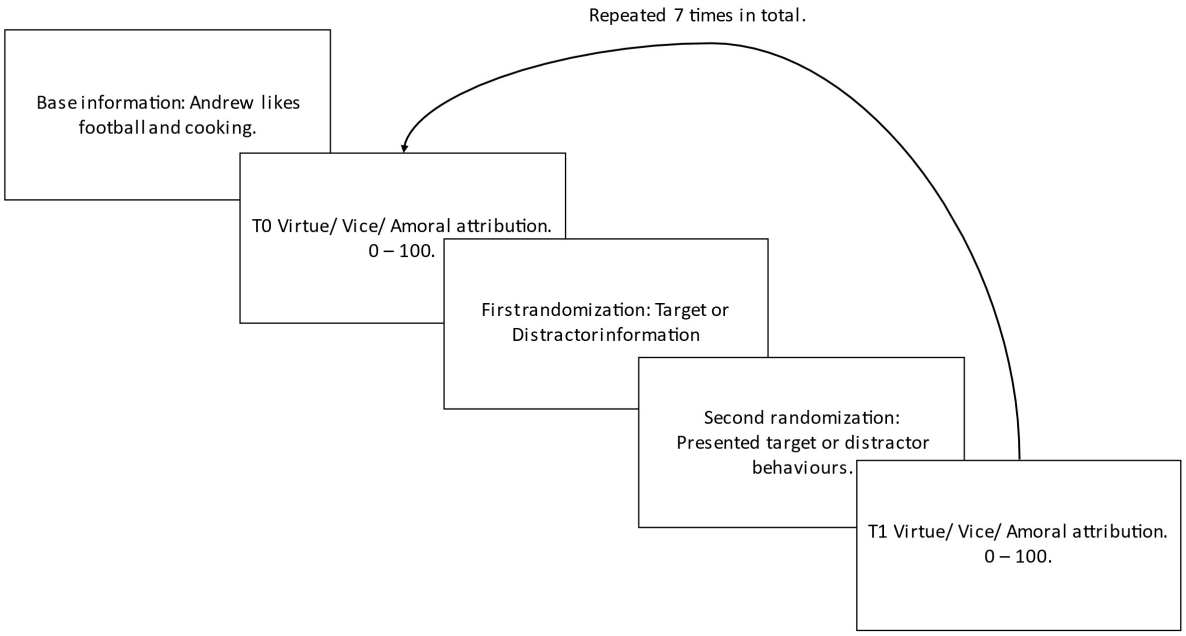


**Figure 1. Task description Study 1.**

**Table 1. Summary of fit and generalizability indexes, BF10 and BF01 across conditions for study 1.**

| Model | AIC | BIC | ICOMP | RMSE | BF10 | BF01 |
|---|---|---|---|---|---|---|
| Bayesian | **757.548** | **814.495** | **745.631** | **0.320** | **120.331** | 1 |
| Bayesian Optimist | 778.389 | 835.336 | 770.721 | 0.323 | 0.004 | 0.00003 |
| Bayesian Prosecutor | 777.801 | 834.748 | 768.388 | 0.323 | 0.005 | 0.00004 |
| Bayesian Quick | 778.101 | 835.048 | 1929.433 | 0.323 | 0.004 | 0.00003 |
| Mean | 779.328 | 836.275 | 769.710 | 0.323 | 0.002 | 0.00002 |
| Primacy | 761.688 | 818.635 | 746.131 | 0.321 | 15.188 | **0.12622** |
| Recency | 776.544 | 833.491 | 760.992 | 0.323 | 1.000 | 0.00831 |
| U | 770.722 | 827.669 | 756.297 | 0.322 | 0.009 | 0.00008 |
| Random | 772.306 | 824.076 | | 0.323 | 0.166 | 0.00138 |

Note: Preferred model according to each index in bold.

to the random model, taken to be a null model, to reflect the likelihood of the considered model given observed data compared to the likelihood of the random model given observed data (Jarosz & Wiley, 2014; Raftery, 1995) (see table 1). BF 10 offers a numeric index reflecting the difference on the weight of the evidence in favour of a target model H1 compared to a null model H0. BF 10 varies between 0 and +∞ with BF10 below 1 reflecting support in favour of the null model (here the random model) whereas BF10 above 1 reflects evidence in favour of the alternative hypothesis. The larger the value is the stronger is evidence in favour of the alternative. Conventionally, values above 150 are considered very strong or decisive evidence and values under 5 are considered weak or anecdotal evidence (Jarosz & Wiley, 2014).

Overall results seem to strongly favour the Bayesian model. Indeed, all generalizability indexes as well as BF10 directly point to this model (*AIC* = 757.542; *BIC* = 814.495; *ICOMP* = 745.630; BF10 = 9.58). To directly compare how much more credible the chosen model is compared to all other fitted, theoretically interesting models we computed a second Bayes Factor (BF01) taking the Bayesian model as the null and comparing each model to it. For readability's sake we computed a BF01 where values below 1 reflect support in favour of the null hypothesis (here Bayesian model) and values above 1 reflect evidence in favour of the alternative (here, each other model). This procedure allows us to quantify how much more credible the Bayesian model is compared to all other theoretically relevant models instead of only comparing it to the theoretically irrelevant Random model. Note that while both BF10 and BF01 compare how much more likely is one model to another, they are *not* inverses of one another. Indeed, here BF10 compares how much more likely is each model *compared to the random model,* while BF01 compares how much less likely is each model *compared to the best model* as de-

**Table 2. Summary of fit and generalizability indexes, BF10 and BF01 for the Virtue condition.**

| Model | AIC | BIC | ICOMP | RMSE | BF10 | BF01 |
|---|---|---|---|---|---|---|
| Bayesian | 282.232 | 319.034 | 270.836 | 0.326 | 0.00001 | 4.24x10^-13 |
| Bayesian Optimist | 256.160 | 292.962 | 1344.979 | 0.317 | 6.19238 | 1.94 x10^ -7 |
| Bayesian Prosecutor | 310.969 | 347.771 | 299.573 | 0.337 | 0 | 2.44 x10^ -19 |
| Bayesian Quick | 265.434 | 302.235 | 1354.250 | 0.320 | 0.06000 | 1.88 x10^ -9 |
| Mean | **225.254** | **262.055** | **215.865** | **0.306** | **31855089** | 1 |
| Primacy | 280.922 | 317.724 | 268.473 | 0.326 | 0.00003 | 8.16 x10^ -13 |
| Recency | 274.282 | 311.083 | 261.804 | 0.324 | 0.00072 | 2.26 x10^ -11 |
| U | 249.337 | 286.138 | 240.026 | 0.315 | 187.77571 | **5.89 x10^ -6** |
| Random | 263.896 | 296.608 | | 0.320 | 1 | 3.14 x10^ -8 |

Note: Preferred model according to each index in bold.

termined by BF10, here the Bayesian model. Hence, BF10 and BF01 indexes offer different and complementary results pertaining to how credible different models are compared to both a theoretically irrelevant baseline (i.e. BF10 comparing each model to the Random model) and to a theoretically relevant best model (i.e. BF01 comparing each model to the best model, Bayesian model). Results suggest that the next best model, the Primacy model, is approximately eight times less likely than the Bayesian model (BF01 = 0.126). Both BF10 and BF01imply positive evidence in favour of the Bayesian model (Jarosz & Wiley, 2014).

However, since there is reason to believe that virtue and vice attribution may follow the different cognitive algorithms (Alicke, 2001; Batson et al., 1999; Batson & Thompson, 2001; Kappes & Sharot, 2019; Rozin & Royzman, 2001; Siegel et al., 2018) we ran similar analyses distinguishing virtue vice and amoral conditions. If the same algorithms (Bayesian model) underlies both virtue, vice and non-moral attributions we expect these analyses to show that this model is preferable in both conditions and according to all considered indexes.

Therefore, we fitted a series of linear regression models for virtue, vice and non-moral conditions separately and computed similar fit and generalizability indexes as well as Bayes factors (see Tables 2 through 4 for the Virtue, Vice and Amoral conditions respectively). Results in the Virtue condition appear inconsistent with general results. According to fit and generalizability indexes Mean model is preferable (*AIC* = 225.254; *BIC* = 262.055; *ICOMP* = 215.865; *BF10* = 3x10^7). BF10 index also suggest vast evidence in favour of the Mean model compared to the random model. The second-best model is the U model but BF01 still suggests definite evidence in favour of the Mean model (*BF01* = 5.89 x 10^-6) (see Table 2).

The Vice condition shows a different pattern of results, the *U* model has better generalizability and fit indexes, corroborated by definite evidence from BF10 (*AIC =215.98*; *BIC* = 251.39; *ICOMP* = 203.02; *RMSE* = 0.314; *BF10* = 6362). Moreover, second-best model, Bayesian Quick model, is more than 33 times less likely than U model (*BF01* = 0.03) which constitutes strong evidence in favour of the U model (see Table 3).

Finally, in the non-moral condition, Primacy model is preferred according to all indexes (*AIC* = 240.19; *BIC* = 277.94; *ICOMP* = 226.77; *RMSE* = 0.3035; *BF10* = 1.44*10^11). BF01 evidence suggests that Primacy model is clearly more likely than its closest model, the Bayesian Prosecutor model (*BF01* = 5.86x10^-3) which constitutes clear evidence in favour of Primacy model (see Table 4).

Finally, we plotted optimized parameters for all preferred models (see fig 2). Since overall results point to the untransformed Bayesian model and the preferred model for Virtue condition was another untransformed model, the Mean model, those are not shown.

### *3.2.1. Discussion*

Study 1 aimed to test several Bayesian and competing algorithmic models of character-based moral judgment. Overall results decisively support a Bayesian neutral model as preferable. However, distinguishing Virtue, Vice and Amoral attributions paints a different picture with Mean model preferred for Virtue attributions, U model for Vice attributions and Primacy models for Amoral attributions. While contrary to our predictions, *a minima* this pattern of results suggests that Virtue and Vice attributions follow different algorithms. Best fitting models when taking into account different attributions separately are average-based and thus compatible with classic studies in person perception and memory (N. H. Anderson, 1961; N. H. Anderson & Barrios, 1961; Baddeley & Hitch, 1993; Birnbaum, 1972, 1973). However, they do follow slightly different algorithms, overweighting different bits of information. On the contrary, overall results suggest a Bayesian neutral model is preferred. This puzzling pattern might result from one of the key limitations of Study 1.

Study 1 only used consistent (i.e. vicious behaviours in the Vice condition and virtuous behaviours in the Virtue condition) or morally irrelevant behaviours (i.e. behaviours that were not relevant for either virtue or vice attributions). This is a serious limitation in two senses. First, the trade-off between accuracy and costs/benefits that character judgment might involve cannot be fully modelled only with consistent information. Hence, it is possible that offered information is too easily interpreted and does not require

**Table 3. Summary of fit and generalizability indexes, BF10 and BF01 for the Vice condition..**

| Model | AIC | BIC | ICOMP | RMSE | BF10 | BF01 |
|---|---|---|---|---|---|---|
| Bayesian | 223.33 | 258.74 | 213.40 | 0.317 | 161 | 0.02533 |
| Bayesian Optimist | 223.70 | 259.11 | 213.76 | 0.318 | 134 | 0.02107 |
| Bayesian Prosecutor | 237.75 | 273.16 | 1124.55 | 0.324 | 0.1 | 0.00002 |
| Bayesian Quick | 222.93 | 258.34 | 1109.73 | 0.317 | 196 | **0.03093** |
| Mean | 259.35 | 294.77 | 250.21 | 0.333 | 0.0 | 0.00000 |
| Primacy | 260.58 | 295.99 | 247.19 | 0.334 | 0.0 | 0.00000 |
| Recency | 234.77 | 270.19 | 221.39 | 0.322 | 0.5 | 0.00008 |
| U | **215.98** | **251.39** | **203.02** | **0.314** | **6362** | 1 |
| Random | 237.43 | 268.91 | | 0.324 | 1 | 0.00016 |

Note: Preferred model according to each index in bold.

**Table 4. Summary of fit and generalizability indexes, BF10 and BF01 for the Amoral condition.**

| Model | AIC | BIC | ICOMP | RMSE | BF10 | BF01 |
|---|---|---|---|---|---|---|
| Bayesian | 270.92 | 308.67 | 271.17 | 0.3132 | 2.97E+04 | 2.12 x10^ -07 |
| Bayesian Optimist | 319.91 | 357.66 | 320.15 | 0.3293 | 6.86E-07 | 4.90 x10^ -18 |
| Bayesian Prosecutor | 250.47 | 288.22 | 250.72 | 0.3068 | 8.20E+08 | **5.86 x10^ -03** |
| Bayesian Quick | 312.33 | 345.89 | | 0.3274 | 2.46E-04 | 1.76 x10^ -15 |
| Mean | 311.80 | 349.55 | 311.96 | 0.3266 | 3.94E-05 | 2.82 x10^ -16 |
| Primacy | **240.19** | **277.94** | **226.77** | **0.3035** | **1.40E+11** | 1 |
| Recency | 286.65 | 324.40 | 273.23 | 0.3183 | 1.14E+01 | 8.17 x10^ -11 |
| U | 319.87 | 357.62 | 307.66 | 0.3293 | 6.99E-07 | 4.99 x10^ -18 |
| Random | 295.72 | 329.27 | | 0.3219 | 1 | 7.14 x10^ -12 |

Note: Preferred model according to each index in bold.

a cognitively costly Bayesian model. Consequently, a more fruitful context in which to test Bayesian models might be in a more incongruent environment which offers more opaque information and thus may require costlier, Bayesian models.

In a similar vein, Study 1 does not closely reflect naturally occurring human interactions since agents often exhibit character-inconsistent behaviour (e.g. an apparently honest person evading taxes or an apparently dishonest person giving back a lost wallet with all money and documents intact) (Doris & Stich, 2007; Harman, 2003; Okten & Moskowitz, 2020). The adequacy of any cognitive model clearly depends on both its objective and how adapted the system is to its environmental structure (J. R. Anderson, 1990; Gigerenzer, 2019). Results presented in Study 1 correspond to models optimized to an environment that is simplified compared to what people face in their daily lives and, therefore, arguably favour cognitively simpler models. Thus, Study 2, aims to provide a more externally valid test of our models by examining whether an alternative, more realistic, description of the environment (i.e. providing both consistent and inconsistent data) results in an alternative model selection. Therefore, Study 2 will more closely mimic daily human life by directly replicating Study 1 but offering both consistent and inconsistent behaviours by the target agent.[3]

## 4. Study 2

Study 2 closely follows Study 1, model parametrization, priors, conditions and moral judgments are identical to Study 1 described above. The main difference is that in Study 2, participants randomly received both congruent (i.e. virtuous behaviours in the Virtue condition or vicious behaviours in the Vice condition) and incongruent (i.e. vicious behaviours in the Virtue condition or virtuous behaviours in the Vice condition) as well as distractor behaviours (i.e. behaviours that were not strongly associated with neither virtuous nor vicious character) as opposed to only receiving congruent or distractor behaviours in Study 1. Also, we did not include a control condition where participants

3 While we believe the more incongruent information offered study 2 is a better reflection of naturalistic informational contexts in which Bayesian models could be used to make sense of agent's behaviour, this study was *not* designed nor run prior to these conclusions. Study 2 was designed and ran after data from study 1 was collected and analysed. Hence, we cannot claim complete pre-registration of study 2.

attributed a non-moral trait in Study 2. See Figure 3 for a description of the task.

### 4.1. Participants, design, materials and procedure

In Study 2 we recruited 149 participants (63 females, 2 did not identify with these genders/ would rather not say, $M_{age}$ = 21.48, $SD_{Age}$ = 7.6). Contrary to with Study 1, participants were randomly assigned to one of only two conditions: Virtue condition where participants attributed overall virtue and Vice condition where participant attributed overall vice. We did not collect an Amoral. Participants received eight behaviours associated to an agent. After each information participants were instructed to judge the target agent's character traits after each information considering all available information about the target agent. Traits were attributed using a 100-point slider ranging from 0 corresponding to "The agent is not a good/ bad/ boring person at all" and 100 corresponding to "The agent is totally a good/ bad/ boring person".

### 4.2. Model parametrization

Models in Study 2 are identical to those used in Study 1. We fitted all 8 models as specified above with no differences in implementation between studies.

### 4.3. Results

As with Study 1, we ran several linear regression models where predicted values predicted observed data controlling by the experimental condition and iteration. Based on these linear regressions we computed fit (*RMSE*) and generalizability (*BIC*; *AIC*, *ICOMP*) indexes as well as Bayes Factors (*BF10* and *BF01*) across all conditions (see Table 5) and for Virtue and Vice condition separately (see Tables 7 and 8 respectively). Similar to study 1, overall results favour a Bayesian model, here, the Bayesian Prosecutor model (*AIC* = 721.79; *BIC* = 777.71; *ICOMP* = 708.15; *RMSE* = 0.3245; *BF10* = 4209). Comparison with the second-best model, Mean model, suggests moderate evidence in favour of the Bayesian prosecutor model compared (*BF01* = 0.13) (see Table 5).

As with Study 1, we computed the same analysis for Virtue and Vice conditions separately to determine the existence of idiosyncratic patterns for virtue and vice attributions. For the Virtue condition, the Bayesian Optimist model was preferred (*AIC* = 445.22; *BIC* = 491.01; *ICOMP* = 430.83). Comparison with the second-best model, U model, suggest definite evidence in favour of the Bayesian Optimist model (*BF01* = 0.0144) (see Table 6).

Finally, For the Vice condition, the U model exhibits the best indexes (*AIC* = 268.34; *BIC* = 309.91; *ICOMP* = 255.49; *RMSE* = 0.3148; *BF10* = 3.37*10^15). Finally, BF01 indexes suggest that the differences in likelihood between u model and the next-best model, Bayesian Prosecutor is only anecdotal (*BF01* = 0.642) (see Table 7).

As for study 1, we plotted parameters for preferred models in all conditions (see fig 4). Overall results point to Bayesian Prosecutor model, while results in the Virtue condition point to the Bayesian Optimist model. While conceptually different, these models both result from an S-transformation of Bayesian neutral model only in different circumstances: only for Vice attributions for the Prosecutor model and only for Virtue attributions for the Optimist model. Hence, their corresponding plots are very similar.

### 4.4. Discussion

The objective of Study 2 was to examine the proposed model's behaviour under a pattern of data that resembles more closely what people face in their daily lives. We operationalized this by showing participants both consistent and inconsistent information about the target agent. In this more ecologically valid situation, results paint a contrary picture to Study 1, whereby Virtue attributions follow cognitively costly Bayesian Optimist model, instead of a cognitively cheaper Mean model in Study 1. However, Vice attributions in Study 2 follow the same cognitively cheap U as in study 1. General discussion aims to reconcile these apparently contradictory results.

## 5. General Discussion

Literature in moral judgement has given a growing importance to different types of moral judgment (Malle et al., 2014; Malle, 2020). Here we are especially interested in character-based moral judgments, taken as a kind of moral judgment that is aimed at personality traits that are consistently linked with morally relevant actions and that serve as basis to predict behaviour (Fleeson et al., 2014; Hartley et al., 2016; Helzer & Critcher, 2015; Pizarro & Tannenbaum, 2012; E. L. Uhlmann et al., 2013, 2014, 2015) and how these character attributions change according to new information.

Our main hypothesis was that virtue and vice attributions would follow a cognitively costly, yet accurate Bayesian Belief Updating Algorithm instead of cognitively cheaper, averaging models. Literature suggests that moral judgment is a motivated process (Ditto et al, 2009). We fitted three types of Bayesian models per participant to reflect qualitatively different motivations: Prosecutor, Quick and Optimist models. Building on literature ranging from on negativity bias (Rozin & Royzman, 2001) and moral reasoning (Alicke, 2001) Prosecutor model overestimates high probabilities for Vice attributions but not for Virtue attributions reflecting a motivation to confirm previous bad character attributions. Quick model on the other hand, reflects findings showing that moral judgment is a rapid, highly adaptive process (Decety & Cacioppo, 2012) and present early in human development (Hamlin, 2013; Surian et al., 2018) by overestimating high probabilities in both Virtue and Vice attributions, thereby arriving at quick conclusions with little information. Finally, Optimist model overestimates high probabilities for virtue attributions only, reflecting that the cost of missing out on cooperation with a virtuous person could outweigh the risk of mislabelling a non-virtuous person which would explain people being less willing to revise virtue compared to vice attributions (Siegel et al., 2018). We also fit 4 competing, average-based mod-

**Table 5. Summary of fit and generalizability indexes, BF10 and BF01 across conditions in Study 2.**

| Model | AIC | BIC | ICOMP | RMSE | BF10 | BF01 |
|---|---|---|---|---|---|---|
| Bayesian | 741.30 | 797.21 | 725.60 | 0.3272 | 0.245 | 5.81x10^-05 |
| Bayesian Optimist | 744.88 | 800.79 | 730.10 | 0.3277 | 0.041 | 9.71 x10^ -06 |
| Bayesian Prosecutor | **721.79** | **777.71** | **708.15** | **0.3245** | **4209.406** | 1 |
| Bayesian Quick | 742.28 | 798.20 | 1882.49 | 0.3273 | 0.150 | 3.56 x10^ -05 |
| Mean | 725.83 | 781.74 | 712.43 | 0.3251 | 559.802 | **0.13** |
| Primacy | 736.53 | 792.45 | 720.37 | 0.3265 | 2.656 | 6.31 x10^ -04 |
| Recency | 744.50 | 800.41 | 728.34 | 0.3276 | 0.049 | 1.17 x10^ -05 |
| U | 739.53 | 795.45 | 723.79 | 0.3269 | 0.593 | 1.41 x10^ -04 |
| Random | 743.57 | 794.40 | | 0.3278 | 1 | 2.38 x10^ -04 |

Note: Preferred model according to each index in bold.

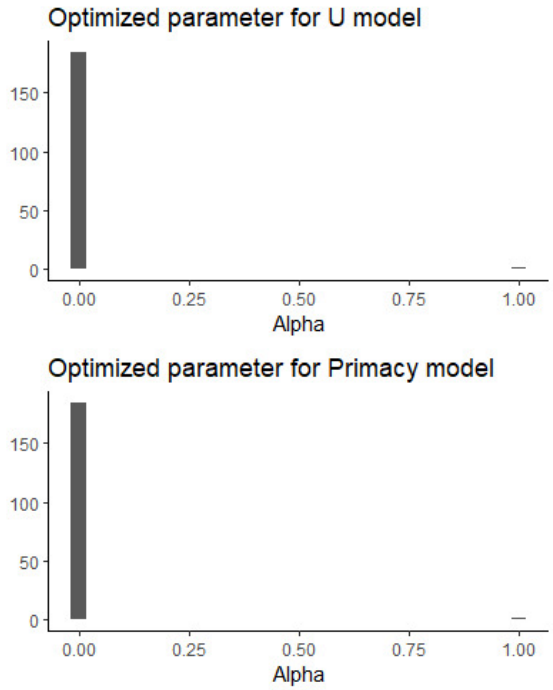


**Figure 2. Optimized parameters for preferred models: U model in the Vice condition and Primacy model in the Amoral condition.**

els. Mean model deals trait attribution based on the average of all perceived behaviours about a target agent. Primacy, Recency and U models respectively overweight the first, last and both first and last received behaviours respectively (N. H. Anderson, 1961; N. H. Anderson & Barrios, 1961; Baddeley & Hitch, 1993; Birnbaum, 1972, 1973)

Surprisingly, Bayesian models were not consistently preferred across studies 1 and 2 nor across virtue or vice attributions (see table 8). Interestingly, the key difference between studies 1 and 2 was that Study 1 only included congruent and morally irrelevant information whereas Study 2 was more externally valid including congruent, incongruent and morally irrelevant information. In this sense, Study 2 is a better test of proposed models our hypothesis. Overall, our results suggest that attributions of virtue or vice do not follow a single, fixed algorithm for either virtue or vice attributions (Guglielmo & Malle, 2019; Mikhail, 2007; Siegel et al., 2018) but rather that people can flexibly change these algorithms depending on how congruent available information is to ensure that attributions are done optimally, that is, as accurate as possible but also with as little cognitive expense as possible. We interpret results as showing that people use Bayesian algorithms as their base attribution mechanism while flexibly changing between them or overestimation of different bits of information according to the attribution at hand (Virtue, Vice or Amoral) and how congruent or incongruent is received information. Specifically, in highly congruent situations (i.e. Study 1) people prefer cognitively cheaper, average-based algorithms for all attributions. However, in incongruent situations (i.e. Study 2) people prefer a cognitively costly Bayesian algorithm for the highly noisy task of accurately identifying Virtue, whereas they also default to cognitively cheaper, U model for the comparatively simpler task of identifying Vice. Also, the fact that across both studies results are much more clear-cut in virtue or vice separately compared to general results as shown by larger BF10 values in separate conditions compared to overall results also hints at virtue and vice attributions following qualitatively different algorithms that are not completely captured by averaging across both conditions.

The trade-off between accuracy, speed and costs is represented differentially for vice and virtue depending on the environment structure, here modelled by congruence of received information. Our results suggest that a fast and frugal, average-based heuristic might be at work when both task and available information are highly consistent and easily interpreted while more cognitively costly algorithms are at play when information is more nuanced, specially, when dealing with positively identifying virtuous agents, rather than simply avoiding vicious ones. This points to several possible algorithms underlying different types of moral judgments instead of a single, unified algorithm dealing with most instances of character attributions (see also Okten & Moskowitz, 2020).

The issue then comes down to which combination of task and informational congruence better represents the formal structure of the environment where these functions perform and whether there are reasons to believe that pres-

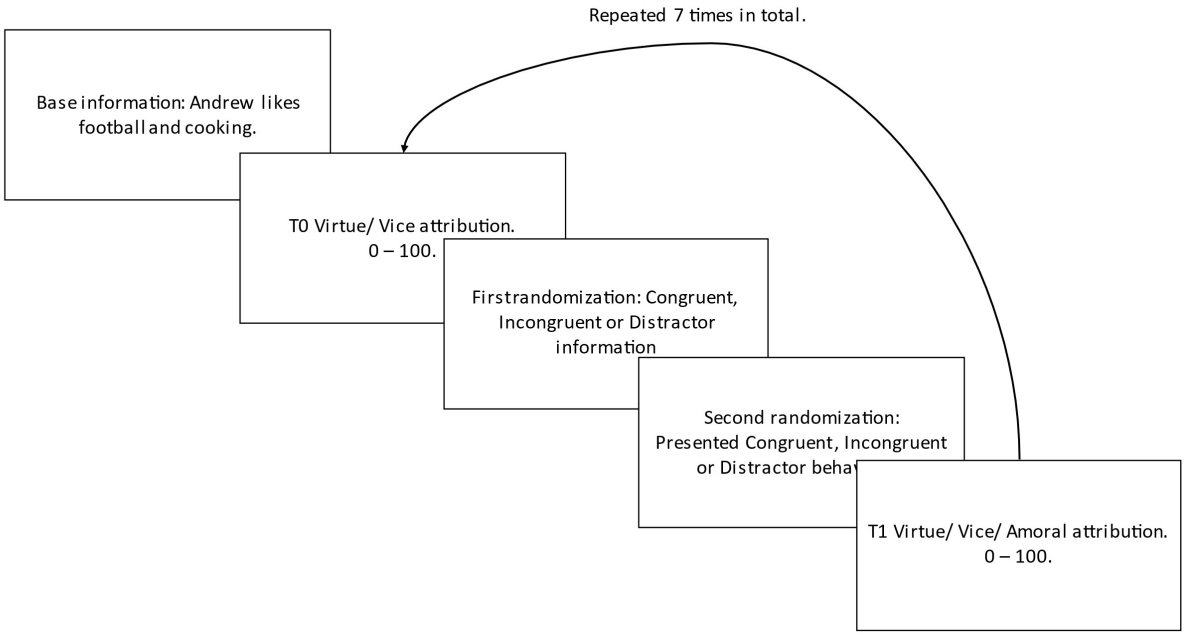


**Figure 3. Task description Study 2.**

**Table 6. Summary of fit, generalizability, BF10 and BF01 for the Virtue condition.**

| Model | AIC | BIC | ICOMP | RMSE | BF10 | BF01 |
|---|---|---|---|---|---|---|
| Bayesian | 464.75 | 510.54 | 450.35 | 0.3295 | 2.39 x10^-13 | 5.75x10^-05 |
| Bayesian Optimist | **445.22** | **491.01** | **430.83** | **0.3251** | **4.15 x10^ -09** | 1 |
| Bayesian Prosecutor | 467.05 | 512.85 | 1476.24 | 0.3301 | 7.54 x10^ -14 | 1.81 x10^ -05 |
| Bayesian Quick | 460.10 | 505.89 | 1469.29 | 0.3285 | 2.44 x10^ -12 | 0.00058713 |
| Mean | 465.68 | 511.47 | 453.33 | 0.3297 | 1.50 x10^ -13 | 3.61 x10^ -05 |
| Primacy | 447.70 | 493.49 | 432.79 | 0.3257 | 1.20 x10^ -09 | 0.28977432 |
| Recency | 453.69 | 499.49 | 438.79 | 0.3270 | 6.00 x10^ -11 | **0.01445203** |
| U | 481.98 | 527.77 | 467.22 | 0.3335 | 4.33 x10^ -17 | 1.04 x10^ -08 |
| Random | 411.20 | 452.42 | | 0.3179 | 1 | 240482531 |

Note: Preferred model according to each index in bold.

**Table 7. Summary of fit, generalizability, BF10 and BF01 indexes for the Vice condition.**

| Model | AIC | BIC | ICOMP | RMSE | BF10 | BF01 |
|---|---|---|---|---|---|---|
| Bayesian | 283.35 | 324.92 | 271.81 | 0.3198 | 1.85x10^12 | 0.00055073 |
| Bayesian Optimist | 314.25 | 355.82 | 1559.19 | 0.3305 | 361355.783 | 1.0703E-10 |
| Bayesian Prosecutor | 269.23 | 310.79 | 257.69 | 0.3151 | 2.1679x10^15 | **0.64209925** |
| Bayesian Quick | 294.54 | 336.11 | 1539.44 | 0.3236 | 6914836847 | 2.0481E-06 |
| Mean | 272.71 | 314.28 | 260.13 | 0.3162 | 3.79x10^14 | 0.11249025 |
| Primacy | 301.17 | 342.74 | 286.61 | 0.3259 | 250298584 | 7.4135E-08 |
| Recency | 303.53 | 345.10 | 288.97 | 0.3267 | 77240839.9 | 2.2878E-08 |
| U | **268.34** | **309.91** | **255.49** | **0.3148** | **3.37x10^15** | 1 |
| Random | 344.01 | 381.42 | | 0.3418 | 1 | 2.9619E-16 |

Note: Preferred model according to each index in bold.

**Table 8. Best and Second-best model for each Study and Condition.**

| | | Best model | BF10 | Second-best model | BF01 |
|---|---|---|---|---|---|
| Study 1 | | | | | |
| | General results | Bayesian | 120 | Primacy | 0.12 |
| | Virtue | Mean | 3*10^7 | U | 5.89x10^-6 |
| | Vice | U | 6362 | Bayesian Quick | 0.031 |
| | Non-moral | Primacy | 1.4 *10^11 | Bayesian Prosecutor | 5.86x10^-3 |
| Study 2 | | | | | |
| | General results | Bayesian Prosecutor | 4209 | Primacy | 0.13 |
| | Virtue | Bayesian Optimist | 4.15x10^-9 | Recency | 0.014 |
| | Vice | U | 3.37*10^15 | Bayesian Prosecutor | 0.642 |

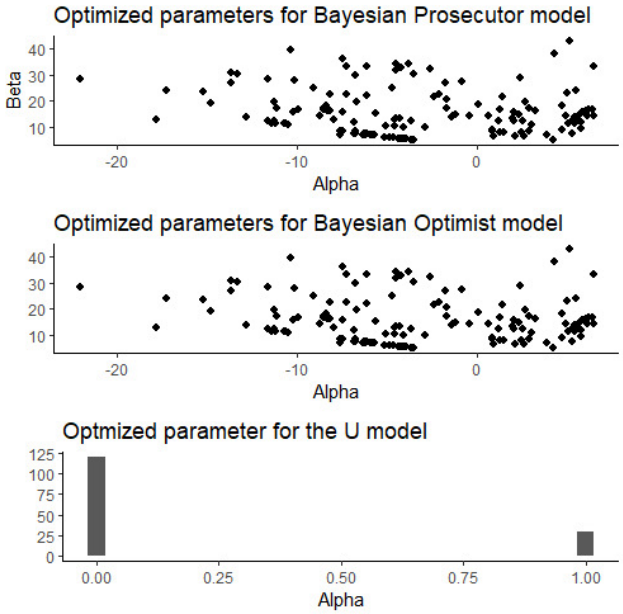


**Figure 4. Optimized parameters for the preferred models: Bayesian Prosecutor model in the overall results, Bayesian Optimist model for the Virtue condition and U model for the Vice condition.**

sures in one or the other direction better justify a particular kind of model as having normative pre-eminence. Future research should naturalistically measure and experimentally manipulate behaviour congruency to empirically tests this explanation. Also, since this interpretation stems from an unforeseen pattern of results, we explicitly encourage independent replication and pre-registration to offer stronger evidence in favour of our hypothesis.

*A minima*, our results support a particular conclusion: both studies offer direct and solid evidence that virtue and vice attributions do not follow identic algorithms but rather that people flexibly manipulate both the kind of algorithm (i.e. Bayesian or average-based) and the overestimation of different bits of information, possibly depending on how congruent presented information is. This conclusion is compatible with recent research proposing that blame and praise attributions, serving different adaptive purposes, are not symmetric processes but rather follow from distinct cognitive mechanisms (R. A. Anderson et al., 2020). These conclusions might be applied to recent models purporting to model the complete cognitive architecture of social prediction (Tamir & Thornton, 2018).

Future research should explore the link between blame and praise, as attributions about a person's actions in a given context, as opposed to more permanent, character attributions. In what circumstances or based on which information do blame and praise attributions can cross over and become more permanent attributions? How do situational variables affecting blame and praise attributions can have an effect on character attributions and the algorithms underlying them? Future research should look into these questions while continuing with nascent work on formalization of moral decision-making. Specifically, there is no consensus on the nature of the algorithms underlying virtue and vice attributions which makes it premature to interpret person-level parameters. However, once established, efforts ought to be made to determine person-level parameters or even algorithms and their consequences for social cognition.

## Acknowledgments

Authors would like to thank A.E. Monroe and G.P.D. Ingram for comments on a very early version of this article. L.F. Talero for invaluable help in data collection as well as several anonymous reviewers whose comments helped shape the final version of this manuscript.

## Funding

This paper was funded by the Colombian Ministry for Science and Technology - Minciencias through the national doctoral training scholarships 727 (2015).

## Competing Interests

The authors have no competing interests to declare.

## Author Contributions

Conceptualization, Writing- Original draft preparation: SB. Methodology, Data curation, Visualization, Investigation, Writing- Reviewing and Editing: SB & WJL.

## Data Accessibility Statement

All materials (including materials, raw data and code) available at: https://osf.io/xsv3e/?view_only=07ef8777795b4990855c02722624d690 .

Submitted: January 26, 2022 PDT, Accepted: August 28, 2022 PDT

## References


Alicke, M. D. (2001). Culpable control and the psychology of blame. *Psychological Bulletin*, *126*(4), 556–574. https://doi.org/10.1037/0033-2909.126.4.556

Alves, H., & Mata, A. (2019). The redundancy in cumulative information and how it biases impressions. *Journal of Personality and Social Psychology*, *117*(6), 1035–1060. https://doi.org/10.1037/pspa0000169

Anderson, J. R. (1990). *The adaptive character of thought*. Psychology Press. https://doi.org/10.4324/9780203771730

Anderson, N. H. (1961). Averaging versus adding as a stimulus-combination rule in impression formation. *Journal of Experimental Psychology*, *70*(4), 394–400. https://doi.org/10.1037/h0022280

Anderson, N. H., & Barrios, A. A. (1961). Primacy effects in personality impression formation. *The Journal of Abnormal and Social Psychology*, *63*(2), 346–350. https://doi.org/10.1037/h0046719

Anderson, R. A., Crockett, M. J., & Pizarro, D. A. (2020). A Theory of Moral Praise. *Trends in Cognitive Sciences*, *24*(9), 694–703. https://doi.org/10.1016/j.tics.2020.06.008

Baddeley, A. D., & Hitch, G. (1993). The Recency effect: Implicit learning with explicit retrieval? *Memory & Cognition*, *21*(2), 146–155. https://doi.org/10.3758/bf03202726

Baker, Chris L., Goodman, N. D., & Tenenbaum, J. B. (2008). Theory-based Social Goal Inference. *Proceedings of the 30th Annual Conference of the Cognitive Science Society*.

Baker, Chris L., Saxe, R. R., & Tenenbaum, J. B. (2011). Bayesian Theory of Mind: Modelling Joint Belief-Desire Attribution. *Proceedings of the Cognitive Science Society*, *6*.

Baker, Chris L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329–349. https://doi.org/10.1016/j.cognition.2009.07.005

Baker, Chris L., Tenenbaum, J. B., & Saxe, R. R. (2005). Bayesian models of human action understanding. In *Advances in Neural Information Processing Systems 18* (p. 8).

Baker, C.L., & Tennenbaum, J. B. (2014). Modeling Human Plan Recognition using Bayesian Theory of Mind. In G. Sukthankar, R. P. Goldman, C. Geib, D. Pynadath, & H. Bui (Eds.), *Plan, activity and intent recognition: Theory and practice*.

Barbosa, S., & Jiménez-Leal, W. (2020). Virtues disunited and the folk psychology of character. *Philosophical Psychology*, *33*(3), 332–350. https://doi.org/10.1080/09515089.2020.1719396

Batson, C. D., & Thompson, E. R. (2001). Why Don't Moral People Act Morally? Motivational Considerations. *Current Directions in Psychological Science*, *10*(2), 54–57. https://doi.org/10.1111/1467-8721.00114

Batson, C. D., Thompson, E. R., Seuferling, G., Whitney, H., & Strongman, J. A. (1999). Moral hypocrisy: Appearing moral to oneself without being so. *Journal of Personality and Social Psychology*, *77*(3), 525–537. https://doi.org/10.1037/0022-3514.77.3.525

Bear, A., & Knobe, J. (2017). Normality: Part descriptive, part prescriptive. *Cognition*, *167*, 25–37. https://doi.org/10.1016/j.cognition.2016.10.024

Birnbaum, M. H. (1972). Morality judgments: Tests of an averaging model. *Journal of Experimental Psychology*, *93*(1), 35–42. https://doi.org/10.1037/h0032589

Birnbaum, M. H. (1973). Morality judgment: Test of an averaging model with differential weights. *Journal of Experimental Psychology*, *99*(3), 395–399. https://doi.org/10.1037/h0035216

Brand, C. M. (2015). The Effect of Probability Anchors on Moral Decision Making. *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*, 5.

Brand, C. M., & Oaksford, M. (2015). The Effect of Probability Anchors on Moral Decision Making. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.), *37th Annual Meeting of the Cognitive Science Society*. Cognitive Science Society.

Bretz, S., & Sun, R. (2018). Two Models of Moral Judgment. *Cognitive Science*, *42*(S1), 4–37. https://doi.org/10.1111/cogs.12517

Cameron, C. D., Payne, B. K., Sinnott-Armstrong, W., Scheffer, J. A., & Inzlicht, M. (2017). Implicit moral evaluations: A multinomial modeling approach. *Cognition*, *158*, 224–241. https://doi.org/10.1016/j.cognition.2016.10.013

Cao, J., Kleiman-Weiner, M., & Banaji, M. R. (2019). People Make the Same Bayesian Judgment They Criticize in Others. *Psychological Science*, *30*(1), 20–31. https://doi.org/10.1177/0956797618805750

Crockett, M. J. (2013). Models of morality. *Trends in Cognitive Sciences*, *17*(8), 363–366. https://doi.org/10.1016/j.tics.2013.06.005

Crockett, M. J. (2016a). Computational modelling of moral decisions. In *Sydney symposium of social psychology. The social psychology of morality* (pp. 71–90). Routledge/Taylor & Francis Group. https://doi.org/10.4324/9781315644189-5

Crockett, M. J. (2016b). How Formal Models Can Illuminate Mechanisms of Moral Judgment and Decision Making. *Current Directions in Psychological Science*, *25*(2), 85–90. https://doi.org/10.1177/0963721415624012

Cushman, F. (2013). Action, Outcome, and Value: A Dual-System Framework for Morality. *Personality and Social Psychology Review*, *17*(3), 273–292. https://doi.org/10.1177/1088868313495594

Decety, J., & Cacioppo, S. (2012). The speed of morality: A high-density electrical neuroimaging study. *Journal of Neurophysiology*, *108*(11), 3068–3072. https://doi.org/10.1152/jn.00473.2012

Doris, J. M. (2002). *Lack of character : Personality and moral behavior*. Cambridge University Press. https://doi.org/10.1017/cbo9781139878364

Doris, J. M., & Stich, S. P. (2007). *As a Matter of Fact: Empirical Perspectives on Ethics*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199234769.003.0005

Doya, K. (Ed.). (2007). *Bayesian brain: Probabilistic approaches to neural coding*. MIT Press.

Fenton, N., Neil, M., & Lagnado, D. A. (2013). A General Structure for Legal Arguments About Evidence Using Bayesian Networks. *Cognitive Science*, *37*(1), 61–102. https://doi.org/10.1111/cogs.12004

Fleeson, W. (2004). Moving Personality Beyond the Person-Situation Debate: The Challenge and the Opportunity of Within-Person Variability. *Current Directions in Psychological Science*, *13*(2), 83–87. https://doi.org/10.1111/j.0963-7214.2004.00280.x

Fleeson, W., Furr, R. M., Jayawickreme, E., Meindl, P., & Helzer, E. G. (2014). Character: The Prospects for a Personality-Based Perspective on Morality. *Social and Personality Psychology Compass*, *8*(4), 178–191. https://doi.org/10.1111/spc3.12094

Gigerenzer, G. (2019). How to Explain Behavior? *Topics in Cognitive Science*, *12*(4), 1363–1381. https://doi.org/10.1111/tops.12480

Goodwin, G. P., Piazza, J., & Rozin, P. (2014). Moral character predominates in person perception and evaluation. *Journal of Personality and Social Psychology*, *106*(1), 148–168. https://doi.org/10.1037/a0034726

Goodwin, G. P., Piazza, J., & Rozin, P. (2015). Understanding the Importance and Perceived Structure of Moral Character. In C. B. Miller, R. M. Furr, A. Knobel, & W. Fleeson (Eds.), *Character* (pp. 100–126). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780190204600.003.0005

Gray, K., & Keeney, J. E. (2015). Impure or Just Weird? Scenario Sampling Bias Raises Questions About the Foundation of Morality. *Social Psychological and Personality Science*, *6*(8), 859–868. https://doi.org/10.1177/1948550615592241

Griffiths, T. L., Kemp, C., & Tenenbaum, J. B. (2008). Bayesian models of cognition. *BAYESIAN MODELS*, *49*.

Guglielmo, S., & Malle, B. F. (2019). Asymmetric morality: Blame is more differentiated and more extreme than praise. *PLOS ONE*, *14*(3), e0213544. https://doi.org/10.1371/journal.pone.0213544

Hagmayer, Y., & Osman, M. (2012). From colliding billiard balls to colluding desperate housewives: Causal Bayes nets as rational models of everyday causal reasoning. *Synthese*, *189*(S1), 17–28. https://doi.org/10.1007/s11229-012-0162-3

Hamlin, J. K. (2013). Moral Judgment and Action in Preverbal Infants and Toddlers: Evidence for an Innate Moral Core. *Current Directions in Psychological Science*, *22*(3), 186–193. https://doi.org/10.1177/0963721412470687

Harman, G. (2003). No Character or Personality. *Business Ethics Quarterly*, *13*(1), 87–94. https://doi.org/10.5840/beq20031316

Harman, G. (2009). Skepticism about Character Traits. *The Journal of Ethics*, *13*(2–3), 235–242. https://doi.org/10.1007/s10892-009-9050-6

Hartley, A. G., Furr, R. M., Helzer, E. G., Jayawickreme, E., Velasquez, K. R., & Fleeson, W. (2016). Morality's Centrality to Liking, Respecting, and Understanding Others. *Social Psychological and Personality Science*, *7*(7), 648–657. https://doi.org/10.1177/1948550616655359

Helzer, E. G., & Critcher, C. R. (2015). What do we evaluate when we evaluate moral character? In *Atlas of Moral Psychology* (p. 21). Guilford University Press.

Jacobs, R. A., & Kruschke, J. K. (2011). Bayesian learning theory applied to human cognition. *WIREs Cognitive Science*, *2*(1), 8–21. https://doi.org/10.1002/wcs.80

Jarosz, A. F., & Wiley, J. (2014). What Are the Odds? A Practical Guide to Computing and Reporting Bayes Factors. *The Journal of Problem Solving*, *7*(1). https://doi.org/10.7771/1932-6246.1167

Kappes, A., & Sharot, T. (2019). The automatic nature of motivated belief updating. *Behavioural Public Policy*, *3*(1), 87–103. https://doi.org/10.1017/bpp.2017.11

Khalvati, K., Park, S. A., Mirbagheri, S., Philippe, R., Sestito, M., Dreher, J.-C., & Rao, R. P. N. (2019). Modeling other minds: Bayesian inference explains human choices in group decision-making. *Science Advances*, *5*(11), 8783. https://doi.org/10.1126/sciadv.aax8783

Kim, R., Kleiman-Weiner, M., Abeliuk, A., Awad, E., Dsouza, S., Tenenbaum, J. B., & Rahwan, I. (2018). A Computational Model of Commonsense Moral Decision Making. *ArXiv*, 1801.04346 [Cs]. https://doi.org/10.1145/3278721.3278770

Kleiman-Weiner, M., & Levine, S. (2015). Inference of Intention and Permissibility in Moral Decision Making. *Proceedings of the 37th Annual Conference of the Cognitive Science Society*, 6.

Kleiman-Weiner, M., Saxe, R., & Tenenbaum, J. B. (2017). Learning a commonsense moral theory. *Cognition*, *167*, 107–123. https://doi.org/10.1016/j.cognition.2017.03.005

Kliemann, D., Young, L., Scholz, J., & Saxe, R. (2008). The influence of prior record on moral judgment. *Neuropsychologia*, *46*(12), 2949–2957. https://doi.org/10.1016/j.neuropsychologia.2008.06.010

Landy, J. F., Piazza, J., & Goodwin, G. P. (2016). When It's Bad to Be Friendly and Smart: The Desirability of Sociability and Competence Depends on Morality. *Personality and Social Psychology Bulletin*, *42*(9), 1272–1290. https://doi.org/10.1177/0146167216655984

Malle, B. F. (2020). Moral Judgments. *Annual Review of Psychology*, *72*(1), 293–318. https://doi.org/10.1146/annurev-psych-072220-104358

Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014). A Theory of Blame. *Psychological Inquiry*, *25*(2), 147–186. https://doi.org/10.1080/1047840x.2014.877340

Melnikoff, D. E., & Bailey, A. H. (2018). Preferences for moral vs. Immoral traits in others are conditional. *Proceedings of the National Academy of Sciences*, *115*, 592–600.

Mende-Siedlecki, P., Baron, S. G., & Todorov, A. (2013). Diagnostic Value Underlies Asymmetric Updating of Impressions in the Morality and Ability Domains. *Journal of Neuroscience*, *33*(50), 19406–19415. https://doi.org/10.1523/jneurosci.2334-13.2013

Mende-Siedlecki, P., Cai, Y., & Todorov, A. (2013). The neural dynamics of updating person impressions. *Social Cognitive and Affective Neuroscience*, *8*(6), 623–631. https://doi.org/10.1093/scan/nss040

Mende-Siedlecki, P., & Todorov, A. (2016). Neural dissociations between meaningful and mere inconsistency in impression updating. *Social Cognitive and Affective Neuroscience*, *11*(9), 1489–1500. https://doi.org/10.1093/scan/nsw058

Mikhail, J. (2007). Universal moral grammar: Theory, evidence and the future. *Trends in Cognitive Sciences*, *11*(4), 143–152. https://doi.org/10.1016/j.tics.2006.12.007

Monroe, A. E., & Malle, B. F. (2018). People systematically update moral judgments of blame. *Journal of Personality and Social Psychology*, *116*(2), 215–236. https://doi.org/10.1037/pspa0000137

Monroe, A. E., & Malle, B. F. (2019). People systematically update moral judgments of blame. *Journal of Personality and Social Psychology*, *116*(2), 215–236. https://doi.org/10.1037/pspa0000137

Moutoussis, M., Fearon, P., El-Deredy, W., Dolan, R. J., & Friston, K. J. (2014). Bayesian inferences about the self (and others): A review. *Consciousness and Cognition*, *25*, 67–76. https://doi.org/10.1016/j.concog.2014.01.009

Newman, G. E., Bloom, P., & Knobe, J. (2014). Value Judgments and the True Self. *Personality and Social Psychology Bulletin*, *40*(2), 203–216. https://doi.org/10.1177/0146167213508791

Okten, I. O., & Moskowitz, G. B. (2020). Easy to Make, Hard to Revise: Updating Spontaneous Trait Inferences in the Presence of Trait-Inconsistent Information. *Social Cognition*, *38*(6), 571–625. https://doi.org/10.1521/soco.2020.38.6.571

Okten, I. O., Schneid, E. D., & Moskowitz, G. B. (2019). On the updating of spontaneous impressions. *Journal of Personality and Social Psychology*, *117*(1), 1–25. https://doi.org/10.1037/pspa0000156

Ong, D. C., Zaki, J., & Goodman, N. D. (2019). Computational Models of Emotion Inference in Theory of Mind: A Review and Roadmap. *Topics in Cognitive Science*, *11*(2), 338–357. https://doi.org/10.1111/tops.12371

Piazza, J., Goodwin, G. P., Rozin, P., & Royzman, E. B. (2014). When a Virtue is Not a Virtue: Conditional Virtues in Moral Evaluation. *Social Cognition*, *32*(6), 528–558. https://doi.org/10.1521/soco.2014.32.6.528

Pitt, M. A., & Myung, I. J. (2002). When a good fit can be bad. *Trends in Cognitive Sciences*, *6*(10), 421–425. https://doi.org/10.1016/s1364-6613(02)01964-2

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review*, *109*(3), 472–491. https://doi.org/10.1037/0033-295x.109.3.472

Pizarro, D. A., & Tannenbaum, D. (2012). Bringing character back: How the motivation to evaluate character influences judgments of moral blame. In M. Mikulincer & P. R. Shaver (Eds.), *The social psychology of morality: Exploring the causes of good and evil.* (pp. 91–108). American Psychological Association. https://doi.org/10.1037/13091-005

Pöppel, J., & Kopp, S. (2019). *Satisficing Mentalizing: Bayesian Models of Theory of Mind Reasoning in Scenarios with Different Uncertainties*. ArXiv:1909.10419 [Cs]. http://arxiv.org/abs/1909.10419

Prelec, D. (1998). The Probability Weighting Function. *Econometrica*, *66*(3), 497. https://doi.org/10.2307/2998573

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Raftery, A. E. (1995). Bayesian Model Selection in Social Research. *Sociological Methodology*, *25*, 111. https://doi.org/10.2307/271063

Rozin, P., & Royzman, E. B. (2001). Negativity Bias, Negativity Dominance, and Contagion. *Personality and Social Psychology Review*, *5*(4), 296–320. https://doi.org/10.1207/s15327957pspr0504_2

Shenhav, A., & Greene, J. D. (2010). Moral Judgments Recruit Domain-General Valuation Mechanisms to Integrate Representations of Probability and Magnitude. *Neuron*, *67*(4), 667–677. https://doi.org/10.1016/j.neuron.2010.07.020

Siegel, J. Z., Mathys, C., Rutledge, R. B., & Crockett, M. J. (2018). Beliefs about bad people are volatile. *Nature Human Behaviour*, *2*(10), 750–756. https://doi.org/10.1038/s41562-018-0425-1

Strohminger, N., Knobe, J., & Newman, G. (2017). The True Self: A Psychological Concept Distinct From the Self. *Perspectives on Psychological Science*, *12*(4), 551–560. https://doi.org/10.1177/1745691616689495

Strohminger, N., & Nichols, S. (2014). The essential moral self. *Cognition*, *131*(1), 159–171. https://doi.org/10.1016/j.cognition.2013.12.005

Surian, L., Ueno, M., Itakura, S., & Meristo, M. (2018). Do Infants Attribute Moral Traits? Fourteen-Month-Olds' Expectations of Fairness Are Affected by Agents' Antisocial Actions. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.01649

Tamir, D. I., & Thornton, M. A. (2018). Modelling the Predictive Social Mind. *Trends in Cognitive Sciences*, *22*(3), 201–212. https://doi.org/10.1016/j.tics.2017.12.005

Uhlmann, E. L., Pizarro, D. A., & Diermeier, D. (2015). A Person-Centered Approach to Moral Judgment. *Perspectives on Psychological Science*, *10*(1), 72–81. https://doi.org/10.1177/1745691614556679

Uhlmann, E. L., & Zhu, L. [Lei]. (2014). Acts, Persons, and Intuitions: Person-Centered Cues and Gut Reactions to Harmless Transgressions. *Social Psychological and Personality Science*, *5*(3), 279–285. https://doi.org/10.1177/1948550613497238

Uhlmann, E. L., Zhu, L. [Lei], & Diermeier, D. (2014). When actions speak volumes: The role of inferences about moral character in outrage over racial bigotry: Person-centered moral judgments. *European Journal of Social Psychology*, *44*(1), 23–29. https://doi.org/10.1002/ejsp.1987

Uhlmann, E. L., Zhu, L. (Lei), & Tannenbaum, D. (2013). When it takes a bad person to do the right thing. *Cognition*, *126*(2), 326–334. https://doi.org/10.1016/j.cognition.2012.10.005

Uhlmann, E., Pizarro, D. a, Tannenbaum, D., & Ditto, P. H. (2009). The motivated use of moral principles. *Judgment and Decision Making*, *4*(607), 479–491.

Ullman, T. D., Baker, C. L., Macindoe, O., Evans, O., Goodman, N. D., & Tenenbaum, J. B. (2009). Help or Hinder: Bayesian Models of Social Goal Inference. *Advances in Neural Information Processing Systems*, *22*, 9.

Yu, H., Siegel, J. Z., & Crockett, M. J. (2019). Modelling Morality in 3-D: Decision-Making, Judgment, and Inference. *Topics in Cognitive Science*, *11*(2), 409–432. https://doi.org/10.1111/tops.12382

# Supplementary Materials

## Peer Review History

Download: https://collabra.scholasticahq.com/article/38032-different-algorithmic-models-underlie-virtue-and-vice-attributions/attachment/98509.docx?auth_token=BA5pWngkm0sh-9tvUVBQ

## Raw Data Study 2

Download: https://collabra.scholasticahq.com/article/38032-different-algorithmic-models-underlie-virtue-and-vice-attributions/attachment/98510.xlsx?auth_token=BA5pWngkm0sh-9tvUVBQ

## Raw Data Study 1

Download: https://collabra.scholasticahq.com/article/38032-different-algorithmic-models-underlie-virtue-and-vice-attributions/attachment/98511.xlsx?auth_token=BA5pWngkm0sh-9tvUVBQ