An information bottleneck view of social stereotype use

Max Taylor-Davies (m.taylor-davies@sms.ed.ac.uk)
School of Informatics, University of Edinburgh

Tadeg Quillien
Department of Psychology, University of Edinburgh

Abstract

For decades, social psychologists have wondered about the cognitive foundations of social stereotype use. Arguments have generally centred either resource constraints, framing stereotypes as 'energy-saving devices', or 'fit', framing stereotypes as tools to represent real structure in the social environment that sometimes go awry. These resource-based and fit-based accounts have typically been presented as being in opposition to one another. In this paper, we seek to show that both are compatible under an information bottleneck model of agent representation. Through a simple simulation experiment, we demonstrate how stereotype use emerges in resource-rational representations as a function of both capacity constraints and the structure of the social environment. We then use the same framework to consider a possible explanation for the outgroup homogeneity bias in terms of limited cognitive capacity.

Keywords: Social cognition; Resource rationality; Information theory; Rate-distortion theory

Introduction

Despite our best efforts, the practice of stereotyping others based on group identity remains a pervasive feature of human social cognition. Given both the prevalence and often negative consequences of social stereotyping, it seems that a unified account of why such stereotypes are so hard to resist could prove very valuable. However, there remains disagreement on the cognitive basis of social stereotype use. Two main perspectives have been advanced in the literature. The 'resource-based' account has a long history (Lippmann, 1922; Allport, 1954; Bodenhausen & Lichtenstein, 1987; Macrae et al., 1994), and holds that stereotypes are employed as 'energy-saving devices' motivated by a need to reduce information processing in complex environments. This account rests on the argument that stereotyping is a less resource-intensive process than individuation (Macrae et al., 1994), and appears to be supported by empirical research finding that stereotype usage increases under higher cognitive load (Bodenhausen & Lichtenstein, 1987; Stangor & Duan, 1991; Pratto & Bargh, 1991; Macrae et al., 1993).

Others have pushed back against the resource-based account, suggesting instead that people's usage of categorical representations is determined more by the extent to which they 'fit' social reality, in the context of the observer's particular goal(s) (Oakes & Turner, 1990; Nolan et al., 1999). Proponents of this position argue that the mere fact of a representation being categorical in nature does not inevitably entail a distortion of social reality. Following Bruner (1957) and Neisser (1987), they may also argue that it is misleading to view categorisation solely through the lens of information *loss*; rather, categories can increase the availability of the information most relevant to a decision-maker. The 'fit-based' account is perhaps best summarised by Oakes & Turner (1990) as the idea that social stereotype use operates in "functional interaction with context rather than in a contextually random manner based on purely internal, information-processing demands and limitations".

In this paper, we argue that rather than being in opposition to one another, resource-based and fit-based accounts of social stereotype use are in fact compatible under a single model. We suggest that both these positions contain an element of truth; i.e. that social stereotypes are motivated by capacity constraints, but can also serve to represent real structure in the environment, while being influenced by an observer's particular decision-making goals.

To do this, we adopt the perspective of *resource-rational analysis*, a conceptual framework which has seen growing adoption within the cognitive sciences (Lieder & Griffiths, 2020; Bhui et al., 2021; Icard, 2023), and seeks to understand cognitive systems as embodying the optimal allocation of limited computational resources to the information-processing problems faced in their environment. Within this high-level approach, there are many possible ways to formalise resource limitations. We adopt an explicitly information-theoretic perspective, modelling a limit on cognitive resources as an upper bound on the amount of information an observer is able to extract from their environment. The advantage of this framing is that it remains agnostic as to particular implementation or substrate details, permitting interpretation in terms of various constructs such as memory or inference capacity (Icard, 2018). Information theory has previously been used to model resource-rational cognition across many domains (Wei & Stocker, 2015; Sims, 2016; Binz & Schulz, 2022; Arumugam et al., 2024; Lai & Gershman, 2024; Icard & Goodman, 2015; Kinney & Lombrozo, 2024; Zaslavsky et al., 2021; Futrell, 2023; Cheyette et al., 2024)—of particular interest to the present work, these principles have recently been applied to studying both categorization (Martínez, 2024; Imel & Zaslavsky, 2024) and social representation (Taylor-Davies & Lucas, 2023). In this paper, we draw on these same principles to study the use of social stereotypes.

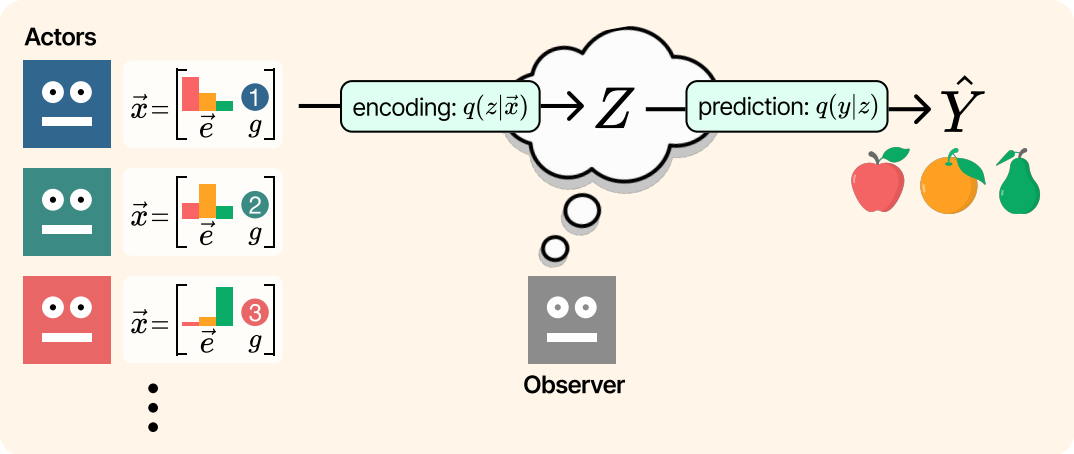

Figure 1: An illustration of our choice prediction task. The *observer* must predict future fruit choices made by a population of *actors*, given input data ($\vec{x}$) in the form of individual choice histories ($\vec{e}$) and discrete group identity labels (g). As a result of limited cognitive capacity, the observer first converts input data $\vec{X}$ into compressed representations Z via encoder $q(z|\vec{x})$, before using Z to make predictions via decoder $q(y|z)$.

We consider the general setting of a social environment where agents are organised into non-overlapping groups, which influence but do not entirely determine their individual attributes. An observer is given access to both some individuating evidence and a categorical group label for each agent in the environment, and seeks to predict their future behaviour under limited cognitive capacity [1]. In our first simulation experiment, we use this setting to argue that social stereotype use is motivated by a combination of both resource constraints and environment structure, rather than responding to either one alone. We then show that under minimal assumptions the same framework produces systematic underestimates of the variability within less-encountered groups, suggesting a role of capacity limits in the phenomenon of outgroup homogeneity bias.

Task setup

To illustrate our modelling approach, we use a simple prediction task, in which an *observer* tries to predict the choices Y made by a population of *actors* between a selection of different fruits. Each actor's fruit choices are (noisily) guided by a vector of preference weights $\vec{W}$, which are influenced by their social group identity G. The observer has access to input data $\vec{X} = [\vec{E}, G]$ for each actor, where $\vec{E}$ is some individuating evidence of the actor's fruit preferences (here a fixed-length history of previous choices). From $\vec{X}$, the observer will produce compressed representations Z via a stochastic encoding $q(z|\vec{x})$, and then use these to predict subsequent fruit choices Y by $q(y|z)$. In detail, the environment is described by the following generative model. First, for each group g we sample a mean preference vector:

$$\vec{\mu}_g \sim \mathcal{N}(\mathbf{0}, \Sigma) \tag{1}$$

Then for each agent i, we first assign them to a group (Eq. 2), and then sample their preferences according to their group mean (Eq. 3). Each agent's choice history and future choices are then distributed according to Eqs. 4 and 5:

$$G_i \sim \text{Uniform}(K) \tag{2}$$

$$\vec{W}_i \sim \mathcal{N}(\vec{\mu}_g, \rho\Sigma) \tag{3}$$

$$\vec{E}_i \sim \text{Multinomial}\big(N_E, \text{softmax}(\vec{w}_i, \beta_{\text{choice}})\big) \tag{4}$$

$$Y_i \sim \text{Categorical}\big(\text{softmax}(\vec{w}_i, \beta_{\text{choice}})\big) \tag{5}$$

where K is the number of groups, Σ is the covariance of the base preference distribution, $\vec{\mu}_g$ is the mean of the preference distribution for group g, and ρ is the ratio between intra- and intergroup variance (see also Equation 11). N_E is the number of previous fruit choices appearing in $\vec{E}$, and β_{choice} is the temperature parameter that determines how noisy actors' choices are[2]. Unless otherwise specified, we use ($K = 3$, $N_E = 15$, $\beta_{\text{choice}} = 0.25$) throughout our simulations.

Modelling framework

The observer's goal is to find a stochastic encoding $q(z|\vec{x})$ that is maximally useful for predicting the actors' future choices Y, given a particular cognitive capacity limit. Formally, this describes an information bottleneck (IB) problem (Tishby et al., 1999), a framework which is closely related to rate-distortion theory (Shannon, 1948; Berger, 2003) and has been applied in fields such as deep learning (Tishby & Zaslavsky,

[1] Since our focus is on modelling stereotype *usage*, and not stereotype *formation*, we assume the observer also knows the distribution of individual properties associated with each group—extending the framework to account for both processes is left for future work.

[2] The softmax function is a common response model for stochastic choice. Given a vector of preferences $\vec{w} = [w_1, w_2, w_3]$, the agent chooses option i with probability $p(i) \propto \exp(w_i/\beta_{\text{choice}})$.

2015; Alemi et al., 2017), neuroscience (Palmer et al., 2015; Rubin et al., 2016), computational linguistics (Mollica, 2024) and concept learning (Imel & Zaslavsky, 2024). In an IB problem, we operationalise both predictive utility and capacity via mutual information, with the optimal encoder q^* given by:

$$q^* = \underset{q}{\text{argmax}}\ I(Y;Z) \text{ subject to } I(X;Z) \leq C \tag{6}$$

I.e., given a ceiling on how much information Z can extract from X, the optimal encoding q^* is one which preserves the most information about Y. The objective in Equation (6) can be approximately solved using a variant of the Blahut-Arimoto algorithm (Blahut, 1972; Arimoto, 1972; Tishby et al., 1999), which iterates the following updates until convergence:

$$q_{t+1}(z|\vec{x}) \propto p_t(z)\exp\big(-\beta d(\vec{x},z)\big) \tag{7}$$

$$p_{t+1}(z) = \sum_{\vec{x}} q_{t+1}(z|\vec{x})p(\vec{x}) \tag{8}$$

$$q_{t+1}(y|z) = \sum_{\vec{x}} p(y|\vec{x})q_{t+1}(\vec{x}|z) \tag{9}$$

with the distortion function $d(\vec{x},z)$ taken as the Kullback-Leibler divergence between $p(y|\vec{x})$ and $q(y|z)$:

$$d(\vec{x},z) = D_{\text{KL}}\big[p(y|\vec{x})||q(y|z)\big] = \sum_y p(y|\vec{x})\log\frac{p(y|\vec{x})}{q(y|z)} \tag{10}$$

In essence, these equations can be understood as saying "iteratively update the encoder so that $q(z|\vec{x})$ is *decreased* to the extent that using z in place of $\vec{x}$ leads to prediction error". The β parameter that appears in Equation (7) controls how the optimisation procedure trades off between rate and distortion (i.e. between simplicity and informativeness), and can be viewed as a Lagrange multiplier for the capacity constraint C in Equation (6)—at low β we obtain encoding policies that extract little information from $\vec{X}$; at high β, we obtain encoders that preserve much more of the information in $\vec{X}$.

Before we proceed further, it is worth briefly clarifying how we wish to interpret the IB account. Under the model given here, social stereotyping might appear identical to any more general process of resource-motivated categorisation. Of course, this ignores the crucial difference that people can perceive *themselves* as members of social categories. For simplicity, this dimension is absent from our simplistic 'ideal observer' task design—but will doubtless have an impact, and should be considered in future work (alongside other non-epistemic motivations for stereotype use).

Simulation 1

Code to reproduce our simulations is available here.

Let us now use the task and framework outlined above to investigate how stereotyped representations emerge from the combination of social structure and capacity constraints. Given an optimal encoding q^* found for a particular capacity limit C, we can determine the extent to which q^* relies on stereotyping by computing the relative mutual information between the learned representations Z and the group labels G vs the individuating evidence $\vec{E}$. Our hypothesis is that this will be jointly determined by both the representational capacity afforded to the encoder, and the stereotype 'fit', i.e. the extent to which G is predictive of an actor's food choices. Intuitively, the extent to which group identity is predictive of an individual's choice behaviour depends on two factors: the variance in preferences *between* groups and the variance in preferences *within* groups. Group identity is most predictive of individual choice behaviour when intergroup variance is high and intragroup variance is low; it is least predictive when intergroup variance is low and intragroup high. The parameter ρ introduced in Eq. 3 corresponds to the ratio:

$$\rho = \frac{\text{intragroup variance}}{\text{intergroup variance}} \tag{11}$$

with knowledge of the actor's group identity becoming less informative as ρ increases (illustrated in Fig. 2 for a single example setting of the group means $\{\vec{\mu}_g\}$).

Our key prediction is that the amount of stereotype use by the optimal encoding q^* will increase as both C and ρ decrease. To test this, we use the Blahut-Arimoto algorithm to find optimal encoding schemes for the task environment described in Eqs. 1-5, for various values of both C and ρ. Specifically, for each value of ρ we sample 50 different instances of the task environment (corresponding to different random seeds). For each instance, we then compute q^* for a range of C values[3], using the technique of reverse deterministic annealing (Zaslavsky & Tishby, 2019) to minimise

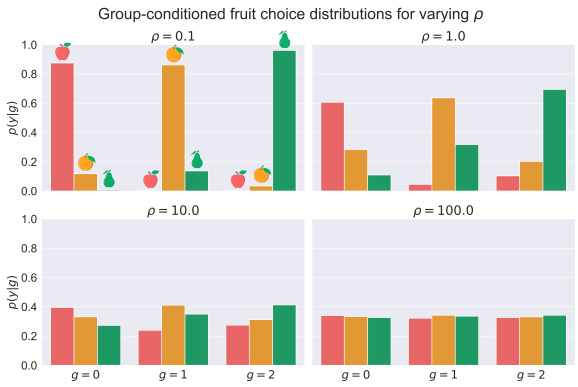

Figure 2: The effect of ρ (the ratio between intragroup and intergroup preference variation) on the group-conditioned fruit choice distribution $p(y|g) \propto \sum_{\vec{w}} p(y|\vec{w})p(\vec{w}|g)$ for one example instantiation of the task environment described in Equations 1-5 (and with $\beta_{\text{choice}} = 0.25$), showing that the utility of knowing G for predicting actors' choice behaviour decreases as we increase ρ.

[3]Practically, we actually find encoders for different values of β, and then compute the normalised mutual information $I(\vec{X};Z)/H(\vec{X})$ to determine corresponding values of $C \in [0,1]$.

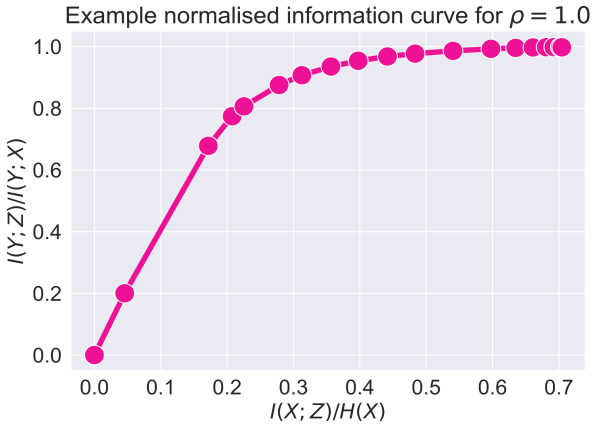

Figure 3: Normalised information curve obtained by running the Blahut-Arimoto algorithm (Equations 7-9) with different values of β for a single example instantiation of the task environment with $\rho = 1.0$.

the risk of convergence to local minima. For each learned encoding q^*, we then compute the mutual information terms $I(\vec{X};Z)$, $I(Y;Z)$, $I(G;Z)$ and $I(\vec{E};Z)$.

As a preliminary test of our experimental setup, we plot the mutual information $I(Y;Z)$ as a function of $I(\vec{X};Z)$ (the 'information curve'), for a single value of $\rho = 1.0$. If the algorithm works as expected, we should see a Pareto relationship between the two quantities—Fig. 3 shows that we do indeed find this, indicating that the optimisation procedure given in Eqs. 7-9 is converging properly.

Next, we test our key hypothesis—that the extent to which the optimal encoding q^* relies on stereotyping will depend on both the capacity of the observer (C) and the ratio between intragroup and intergroup preference variance (ρ). Fig. 4 shows how the relative amounts of information preserved by Z about the group label G and individual choice history $\vec{E}$ vary with both C and ρ. Fig. 5 offers a condensed view of the same data, showing the amount of stereotyping $I(G;Z) - I(\vec{E};Z)$ directly as a function of ρ and C.

We find that both parameters have a meaningful influence on the amount of stereotyping. Interestingly, stereotyping is maximised for $\rho \to 0$ and *moderate* C. To see why, we can consider the low-ρ case (e.g. $\rho = 0.01$). At very low capacity, the observer can extract little information about either group labels or individuating inputs. As capacity is increased, it is initially allocated mainly to extracting information about the cheaper-to-represent G (which at low ρ is usefully predictive of Y), with a comparable amount of information extracted about $\vec{E}$ only once C is high enough that there is 'spare' capacity to do so. In sum, our model predicts that stereotyping should occur when knowledge of group membership is highly predictive of behaviour (consistent with fit-based accounts), and in conditions where observers face some cognitive capacity constraints (consistent with resource-based accounts).

Simulation 2

In our first simulation experiment, we showed that the information bottleneck approach can unify the high-level perspectives of both resource-based and fit-based accounts of social stereotype use under a single model. We now investigate whether it can also capture the phenomenon of *outgroup homogeneity bias*, where people perceive members of an outgroup as being less differentiated than their ingroup (Quattrone & Jones, 1980; Ostrom & Sedikides, 1992; Judd et al., 2005). While a multitude of theories have been offered to explain this effect, we highlight just two here. An early account from Ostrom et al. (1993) focuses on the 'differential encoding' of stimulus information. They suggest that information about ingroup and outgroup members is stored using different categorical structures, based on individual person categories or stereotypical attributes, respectively. More recently, Konovalova & Le Mens (2020) have advanced a sampling-based explanation under which the outgroup homogeneity bias arises simply from the structure of the environment, without any need for differential processing. They suggest that people's perceptions of group variability are based on the sample variance over their encounters with group members—since sample variance tends to increase with sample size, the assumption that people encounter outgroup members less frequently than ingroup members is then sufficient to produce an outgroup homogeity bias.

We draw on both of these theories to argue that the outgroup homogeneity bias arises from a combination of environment structure *and* information processing—emerging as a consequence of resource-rational agent representation in environments where outgroup members are encountered less frequently than ingroup members. Like Konovalova & Le Mens (2020), we suggest that perceptions of outgroup homogeneity result from differences in sample variance—however, we consider the sample variance not over 'true' observations of group members but rather over *reconstructed* attributes from compressed representations. Importantly, this difference in sample variance is driven not by differences in sample size, but from how much *information* Z preserves about attributes of ingroup vs outgroup members. In this way, our account is one of differential encoding, like (Ostrom et al., 1993)—but with the crucial point that the encoding process is shaped directly by the structure of the social environment. To see how this follows, consider an observer with fixed capacity C, in an environment where some inputs are encountered frequently (high $p(\vec{x})$) and others are rare (low $p(\vec{x})$). If the observer is motivated to maximise the expected usefulness of their representations Z (i.e. minimise $\langle d(\vec{x},z)\rangle$), they should intuitively allocate a larger share of C to representing the more frequent inputs, resulting in coarser representations of the rarer inputs. If we apply this general argument to a social setting where the observer interacts primarily with their ingroup, we should expect them to represent outgroup members in less detail—producing perceptions that underestimate their heterogeneity.

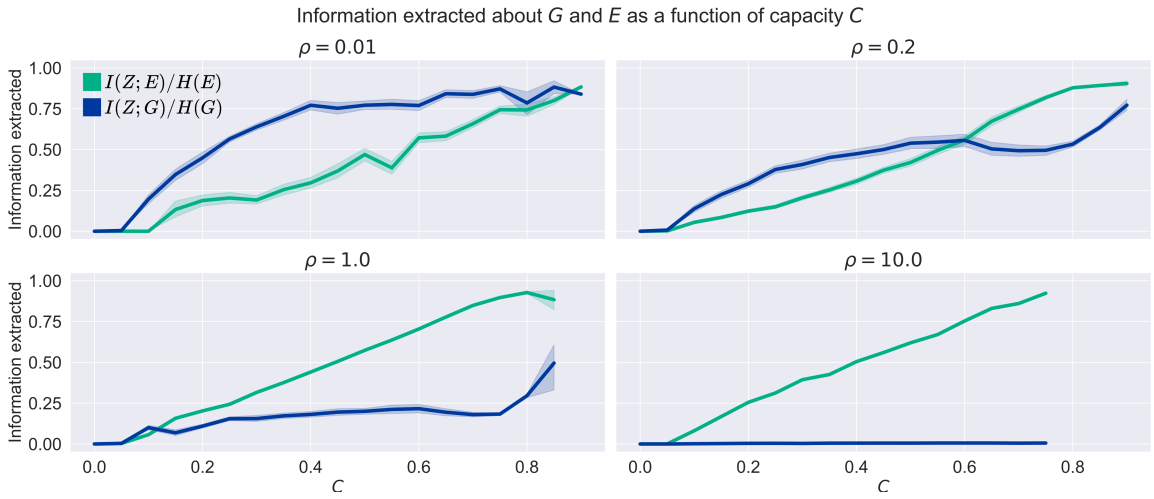

Figure 4: Information extracted by the optimal encoder about group labels G and individuating evidence E as a function of observer capacity C for selected values of the variance ratio ρ. $H(*)$ denotes the Shannon entropy of *. Shaded areas represent bootstrapped 95% confidence intervals over 50 seeds.

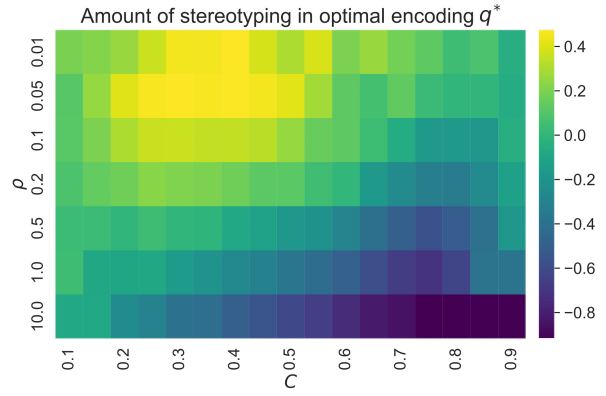

Figure 5: Relative difference in group vs individuating information extracted $I(Z;G)/H(G) - I(Z;E)/H(E)$ as a function of both cognitive capacity C and variance ratio ρ, averaged over 50 seeds. $H(*)$ denotes the Shannon entropy of *.

To test this idea, we consider the same choice prediction setting as in Simulation 1, but rather than three agent groups, we now just have two—an ingroup and an outgroup—with non-uniform probabilities $p(g)$. For different values of $p(g = \text{outgroup}) \leq p(g = \text{ingroup})$, we use the algorithm given in Eqs. 7-9 to determine the optimal encoder q^* for different levels of cognitive capacity C. Since for this setting we're not interested in the variance ratio ρ, we fix the group means at $\mu_{\text{ingroup}} = [1, 0, -1]$, $\mu_{\text{outgroup}} = [-1, 0, 1]$ and the intragroup variance at $\Sigma = I_3$ (i.e. both the ingroup and outgroup always have the same *true* variability).

For each optimised encoder q^*, we first sample for both groups an *equally-sized* set of 10^4 representative inputs $\{\vec{x}\}_g \sim p(\vec{x}|g)$, and then encode them to produce compressed representations $\{z\}_g \sim q^*(z|\vec{x})$. From each z, we then estimate the underlying latent fruit preferences $\vec{w}|z$ following the Bayes-optimal decoder

$$\hat{w} = \mathbb{E}\left[\vec{w}|z\right] = \sum_{\vec{x},\vec{w}} \vec{w} p(\vec{w}|\vec{x}) q(\vec{x}|z) \tag{12}$$

to produce $\{\hat{w}\}_g$. We then use this to estimate the group variance as $\text{Var}_{q^*}(\hat{w}|g) = \text{tr}(\Sigma_g)$ where Σ_g is the sample covariance of $\{\hat{w}\}_g$. Given a large enough sample, $\text{Var}_{q^*}(\hat{w}|g)$ tells us the variance in preferences that the observer ascribes to group g, assuming they use q^* to process information about all actors in their environment. We compute estimated variances for both groups, and then take the level of outgroup homogeneity bias in encoder q^* as

$$\text{OHB} = \frac{\text{Var}_{q^*}(\hat{w}|g = \text{ingroup})}{\text{Var}_{q^*}(\hat{w}|g = \text{outgroup})} \tag{13}$$

Fig. 6 shows this measure recorded for different values of $p(\text{outgroup})$ and encoder capacity C. We find an interesting pattern. For all $p(\text{outgroup}) < 0.5$, the observed bias initially increases sharply with C, as q^* goes from encoding both agent groups with minimal detail to encoding the ingroup with higher fidelity. As we continue to increase C, the bias decays towards as a constant level as more information is encoded about outgroup individuals. While the basic shape of the curve is the same for all $p(\text{outgroup}) < 0.5$, we see a systematic effect of $p(\text{outgroup})$ on the maximum observed bias. Finally, for $p(\text{outgroup}) = 0.5$ we see no bias emerge regardless of capacity, as we would expect. Fig. 7 shows the same results in heatmap form—we see that the outgroup homogeneity bias is maximised as we decrease $p(\text{outgroup})$ at moderately low (but not minimal) C.

Discussion

In this paper, we used the information bottleneck framework to study the phenomenon of social stereotype use in a simple

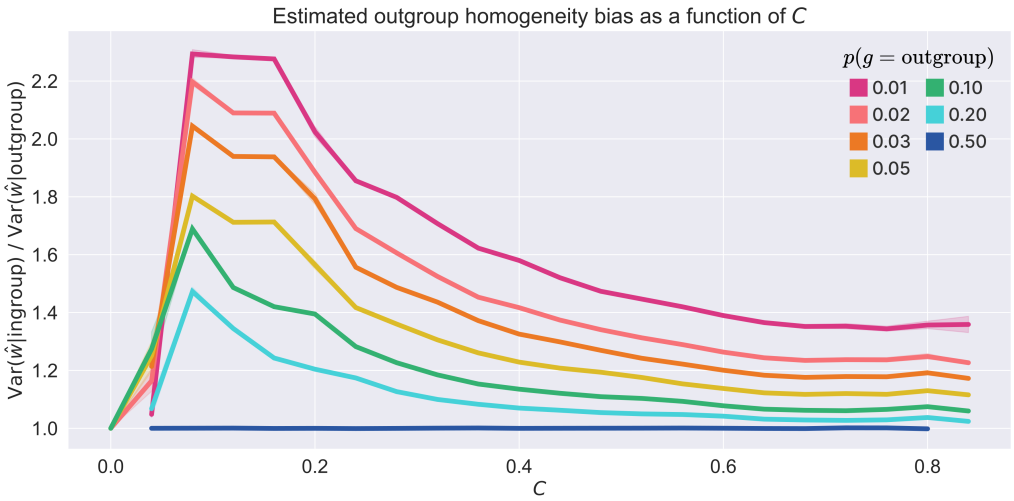

Figure 6: Outgroup homogeneity bias exhibited by the optimal encoder, $\mathrm{Var}_{q^*}(\hat{w}|g = \text{ingroup})/\mathrm{Var}_{q^*}(\hat{w}|g = \text{outgroup})$ as a function of observer capacity for different values of $p(g = \text{outgroup})$. Shaded areas represent bootstrapped 95% confidence intervals over 50 seeds.

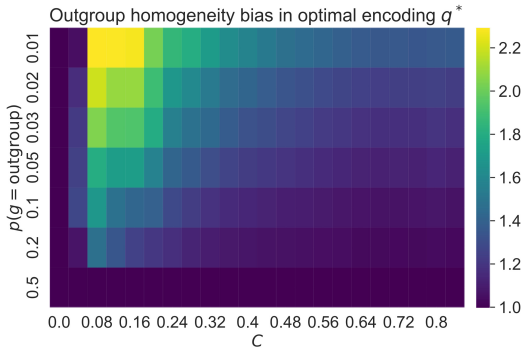

Figure 7: Outgroup homogeneity bias exhibited by the optimal encoder as a function of capaacity C and $p(g = \text{outgroup})$, averaged over 50 seeds.

choice prediction setting. In our first simulation experiment, we demonstrated that the degree to which a resource-rational observer relies on group labels over individuating information depends on both their cognitive capacity and the ratio between inter-group and intra-group variance. In our second simulation experiment, we showed that under minimal assumptions the same framework can be used to account for the outgroup homogeneity bias. Taken together, our results highlight the combined role of resource constraints *and* social environment structure in determining the extent to which a rational observer represents other agents categorically. While the framework we use casts the representational task in terms of an optimisation problem, it is important to note that we are not claiming that people solve Equation (6) themselves in each social environment or task context. Rather, the perspective taken in resource-rational analysis is that the problem has been approximately solved over time by evolutionary, developmental or learning processes, with observers simply executing the resulting policy (Lieder & Griffiths, 2020; Icard, 2023). In this way, the account we present can also be seen as an application of *ecological* rationality (Gigerenzer & Brighton, 2009).

We finish by outlining some limitations of the current work. Firstly, our 'experiments' were purely simulation-based, with no behavioural component. While direct comparisons to human behaviour are complicated by the difficulty of manipulating participants' cognitive capacity, future work could draw closer connections between the predictions of the IB model and the results of the cognitive load experiments referenced in the introduction. Another limitation is our use of a very simplistic social structure, with exhaustive and mutually exclusive groups. Real social environments are rarely this straightforward—group boundaries are often fuzzy, and people don't just have a single social identity that is static over time and across different contexts. Relatedly, real observers don't typically have direct access to others' 'true' social identity; rather people express various noisy cues that we use to make inferences about the social groups to which they might belong. Future work that incorporates these aspects as part of a more realistic task setting would be valuable. Finally, by focusing on the perspective of an idealised outside observer, the view we present here does not consider other potential factors involved in stereotype use that are fundamentally tied to a perceiver's *own* group identity. For instance, once people categorise themselves as members of a particular ingroup, they may represent outgroup individuals more negatively, or in a way that accentuates perceived differences between ingroup and outgroup (Tajfel et al., 1971; Tajfel, 1981; Gramzow et al., 2001; Dunham, 2018; Pietraszewski, 2020). Although this paper deals purely with the possible epistemic functions of social stereotype use, understanding the social functions is no less important, especially as relates to the aim of mitigating the negative impacts of stereotyping.

References

Alemi, A. A., Fischer, I., Dillon, J. V., & Murphy, K. (2017). Deep variational information bottleneck. In *International conference on learning representations.*

Allport, G. W. (1954). *The nature of prejudice.* Oxford, England: Addison-Wesley. (Pages: xviii, 537)

Arimoto, S. (1972). An algorithm for computing the capacity of arbitrary discrete memoryless channels. *IEEE Transactions on Information Theory*, *18*(1), 14–20.

Arumugam, D., Ho, M. K., Goodman, N. D., & Van Roy, B. (2024). Bayesian Reinforcement Learning With Limited Cognitive Load. *Open Mind*, *8*, 395–438. doi: 10.1162/opmi_a_00132

Berger, T. (2003). Rate-distortion theory. *Wiley Encyclopedia of Telecommunications*.

Bhui, R., Lai, L., & Gershman, S. J. (2021). Resource-rational decision making. *Current Opinion in Behavioral Sciences*, *41*, 15–21. doi: 10.1016/j.cobeha.2021.02.015

Binz, M., & Schulz, E. (2022). Modeling human exploration through resource-rational reinforcement learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), *Advances in neural information processing systems* (Vol. 35, pp. 31755–31768). Curran Associates, Inc.

Blahut, R. (1972). Computation of channel capacity and rate-distortion functions. *IEEE transactions on Information Theory*, *18*(4), 460–473.

Bodenhausen, G. V., & Lichtenstein, M. (1987). Social stereotypes and information-processing strategies: The impact of task complexity. *Journal of Personality and Social Psychology*, *52*(5), 871–880. doi: 10.1037/0022-3514.52.5.871

Bruner, J. S. (1957). On perceptual readiness. *Psychological Review*, *64*(2), 123–52. doi: 10.1037/h0043805

Cheyette, S. J., Wu, S., & Piantadosi, S. T. (2024). Limited information-processing capacity in vision explains number psychophysics. *Psychological Review*, *131*(4), 891–904. doi: 10.1037/rev0000478

Dunham, Y. (2018). Mere membership. *Trends in Cognitive Sciences*, *22*, 780-793.

Futrell, R. (2023). Information-theoretic principles in incremental language production. *Proceedings of the National Academy of Sciences*, *120*(39). doi: 10.1073/pnas.2220593120

Gigerenzer, G., & Brighton, H. (2009). Homo heuristicus: Why biased minds make better inferences. *Topics in Cognitive Science*, *1*(1), 107-143. doi: https://doi.org/10.1111/j.1756-8765.2008.01006.x

Gramzow, R. H., Gaertner, L., & Sedikides, C. (2001). Memory for in-group and out-group information in a minimal group context: the self as an informational base. *Journal of personality and social psychology*, *80 2*, 188-205.

Icard, T. F. (2018). Bayes, Bounds, and Rational Analysis. *Philosophy of Science*, *85*(1), 79–101. doi: 10.1086/694837

Icard, T. F. (2023). *Resource rationality.*

Icard, T. F., & Goodman, N. D. (2015). A Resource-Rational Approach to the Causal Frame Problem. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *37*(0).

Imel, N., & Zaslavsky, N. (2024). Optimal compression in human concept learning. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *46*(0).

Judd, C. M., Park, B., Yzerbyt, V., Gordijn, E. H., & Muller, D. (2005). Attributions of intergroup bias and outgroup homogeneity to ingroup and outgroup others. *European Journal of Social Psychology*, *35*(6), 677–704. doi: 10.1002/ejsp.281

Kinney, D., & Lombrozo, T. (2024). Building compressed causal models of the world. *Cognitive Psychology*, *155*. doi: 10.1016/j.cogpsych.2024.101682

Konovalova, E., & Le Mens, G. (2020). An information sampling explanation for the in-group heterogeneity effect. *Psychological Review*, *127*(1), 47–73. doi: 10.1037/rev0000160

Lai, L., & Gershman, S. J. (2024). Human decision making balances reward maximization and policy compression. *PLOS Computational Biology*, *20*(4). doi: 10.1371/journal.pcbi.1012057

Lieder, F., & Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, *43*. doi: 10.1017/S0140525X1900061X

Lippmann, W. (1922). *Public opinion*. Harcourt, Brace.

Macrae, C. N., Hewstone, M., & Griffiths, R. J. (1993). Processing load and memory for stereotype-based information. *European Journal of Social Psychology*, *23*(1), 77-87. doi: https://doi.org/10.1002/ejsp.2420230107

Macrae, C. N., Milne, A. B., & Bodenhausen, G. V. (1994). Stereotypes as energy-saving devices: A peek inside the cognitive toolbox. *Journal of Personality and Social Psychology*, *66*(1), 37–47. doi: 10.1037/0022-3514.66.1.37

Martínez, M. (2024). The information-processing perspective on categorization. *Cognitive Science*, *48*(2), e13411.

Mollica, F. (2024). A note on complexity in efficient communication analyses of semantic typology. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *46*(0).

Neisser, U. (1987). *Concepts and conceptual development: Ecological and intellectual factors in categorization*. Cambridge University Press.

Nolan, M. A., Haslam, S. A., Spears, R., & Oakes, P. J. (1999). An examination of resource-based and fit-based theories of stereotyping under cognitive load and fit. *European Journal of Social Psychology*, *29*, 641–663.

Oakes, P. J., & Turner, J. C. (1990). Is limited information processing capacity the cause of social stereotyping? *European Review of Social Psychology*, *1*(1), 111–135.

Ostrom, T. M., Carpenter, S. L., Sedikides, C., & Li, F. (1993). Differential processing of in-group and out-group information. *Journal of Personality and Social Psychology*, *64*(1), 21–34. doi: 10.1037/0022-3514.64.1.21

Ostrom, T. M., & Sedikides, C. (1992). Out-group homogeneity effects in natural and minimal groups. *Psychological Bulletin*, *112*(3), 536–552. doi: 10.1037/0033-2909.112.3.536

Palmer, S. E., Marre, O., Berry, M. J., & Bialek, W. (2015). Predictive information in a sensory population. *Proceedings of the National Academy of Sciences*, *112*(22), 6908-6913. doi: 10.1073/pnas.1506855112

Pietraszewski, D. (2020). Intergroup processes: Principles from an evolutionary perspective. In *Social psychology: Handbook of basic principles.*

Pratto, F., & Bargh, J. A. (1991). Stereotyping based on apparently individuating information: Trait and global components of sex stereotypes under attention overload. *Journal of Experimental Social Psychology*, *27*(1), 26-47. doi: https://doi.org/10.1016/0022-1031(91)90009-U

Quattrone, G. A., & Jones, E. E. (1980). The perception of variability within in-groups and out-groups: Implications for the law of small numbers. *Journal of Personality and Social Psychology*, *38*(1), 141–152. doi: 10.1037/0022-3514.38.1.141

Rubin, J., Ulanovsky, N., Nelken, I., & Tishby, N. (2016, 08). The representation of prediction error in auditory cortex. *PLOS Computational Biology*, *12*(8), 1-28. doi: 10.1371/journal.pcbi.1005058

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*(3), 379-423. doi: https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Sims, C. R. (2016). Rate–distortion theory and human perception. *Cognition*, *152*, 181-198. doi: https://doi.org/10.1016/j.cognition.2016.03.020

Stangor, C., & Duan, C. (1991). Effects of multiple task demands upon memory for information about social groups. *Journal of Experimental Social Psychology*, *27*(4), 357-378. doi: https://doi.org/10.1016/0022-1031(91)90031-Z

Tajfel, H. (1981). *Human groups and social categories: Studies in social psychology*. Cambridge University Press.

Tajfel, H., Billig, M., Bundy, R. P., & Flament, C. (1971). Social categorization and intergroup behaviour. *European Journal of Social Psychology*, *1*, 149-178.

Taylor-Davies, M., & Lucas, C. G. (2023). Balancing utility and cognitive cost in social representation. *arXiv preprint arXiv:2310.04852*.

Tishby, N., Pereira, F. C., & Bialek, W. (1999). The information bottleneck method. *arXiv preprint physics/0004057*.

Tishby, N., & Zaslavsky, N. (2015, 03). Deep learning and the information bottleneck principle. *2015 IEEE Information Theory Workshop, ITW 2015*. doi: 10.1109/ITW.2015.7133169

Wei, X.-X., & Stocker, A. A. (2015). A Bayesian observer model constrained by efficient coding can explain 'anti-Bayesian' percepts. *Nature Neuroscience*, *18*(10), 1509–1517. doi: 10.1038/nn.4105

Zaslavsky, N., Hu, J., & Levy, R. P. (2021). A Rate–Distortion view of human pragmatic reasoning? In A. Ettinger, E. Pavlick, & B. Prickett (Eds.), *Proceedings of the Society for Computation in Linguistics 2021* (pp. 347–348).

Zaslavsky, N., & Tishby, N. (2019). Deterministic annealing and the evolution of optimal information bottleneck representations.