To ascertain the extent to which self-categorization theory provides a better or worse explanation of our data, we will directly compare these competing explanations using Bayesian ordinal logistic models with priors that reflected the different assumptions of each explanation in relation to how ingroup and outgroup descriptive norms are expected to influence preferences.

Each of the following models represent Bayesian versions of ordinal logistic regression, which predicts the proportions of responses on an ordinal scale while assuming that certain variables (in our case, the descriptive norms) change the odds of making higher or lower responses on the scale. Specifically, the variables are parameterized in terms of the natural log odds of favouring a higher response (more strongly preferring to not report the robber). We can represent this as shown in Equation 1:

(1)

Here, *I* represents the ingroup norm condition (and the corresponding direction of the outgroup norm), B represents both norms shown and *I × B* represents the outgroup norm, while *bin****­****, bboth and bout a*re parameters representing the effects of changing these conditions. We outline the priors for these parameters which we have assigned to capture self-categorization theory and the alternative hypothesis below. We describe our reasoning for each below.

According to self-categorization theory, individuals should only follow a supposed ingroup norm if they actually self-identify with that group and should only want to avoid following an outgroup norm when they consider that outgroup as separate from their self-identity. We thus include these interactions in the self-categorization model, as outlined in Equation 2, where ingroup agree is coded as 1 if the participant reports identifying with the ingroup and 0 otherwise. Equivalent binary coding is used for outgroup disagree.

(2)

The ingroup agree and outgroup disagree variables effectively act as switches, determining whether the self-categorization model assumes the participant will be affected by the ingroup and outgroup descriptive norms respectively. Scores on these two variables will have no bearing on a participant’s inclusion in the data analysis. The alternative hypothesis assumes that identification with the group that a descriptive norm comes from does not influence the effect of that descriptive norm and thus, these variables are ignored by the alternative model.

#### Prior assumptions: bin

Given the similarity of our ingroup norm condition to the experiments reported in Pryor, Perfors and Howe [22], we will use the observed effect of the ingroup norm in those experiments to inform our prior for the ingroup norm effect in the current analysis. The log odds ratio estimated across the relevant experiments from Pryor, Perfors and Howe [22] was 1.02 with a standard deviation of 0.19 (see S2 Text). A notable difference between those experiments and the norms that will be presented in the current experiment is that the previous experiments presented a stronger ingroup norm (75% of ingroup did X) than the current experiment (60% of ingroup did X). To adjust for this difference and account for increased uncertainty in this parameter estimate, we will set the prior distribution for the effect of the ingroup norm to be a folded normal distribution with a mean of , and a standard deviation of 0.5 for both the self-categorization and alternative models. Given both models assume that ingroup norms will shift responses towards the options favoured by the ingroup norm (i.e. the effect of the ingroup norm will be positive), we fold this normal distribution such that the prior is restricted to be greater than 0.

#### Prior assumptions: bboth

The parameter represents a possible main effect on responses of merely presenting both an ingroup and outgroup norm compared to only an ingroup norm, independent of the direction of those norms. Including this effect in the models is important as it allows for the possibility that the outgroup norm is more effective in one direction than in another. It also allows for the possibility that being presented with two opposing norms shifts people’s bias. For example, the increased ambiguity caused by having opposing norms presented may elicit an omission bias [28], wherein taking no action (i.e. not reporting the robber) is favoured more often, independent of the actual direction of the norms. Given that the presentation of the outgroup norm will be independent of the direction of either norm, we have no clear, theoretical reason to predict a strong systematic effect in either direction due to merely presenting an outgroup norm, independent of that norm’s direction. Thus, the self-categorization and the alternative model both will adopt a weakly informative prior for *bboth*, represented by a normal distribution with a mean of 0 and standard deviation of 0.5.

#### Prior assumptions: bout

The parameter *bout* is the key manner in which the self-categorization explanation of the descriptive norm effect differs from that of the alternative hypothesis that people follow the overall descriptive norm. This parameter represents the extent to which presenting an outgroup norm that is opposite to the ingroup norm shifts preferences towards or away from the option favoured by the ingroup norm.

For self-categorization theory, presenting an outgroup norm that opposes the ingroup norm is expected to increase conformity with the ingroup norm. We represent this with a normally-distributed prior that is restricted to be greater than 0 (specifically a half-normal distribution with a mean of 0 and standard deviation of 0.5).

Contrasting with self-categorization theory, the alternative hypothesis assumes that people care about the overall norm, regardless of whether it comes from an ingroup or outgroup. Under it, ingroup and outgroup norms are assumed to affect preferences equivalently, only differing to the extent that the strength of these norms differ. Given that our outgroup norm will always be opposite to the ingroup norm and stronger (85% of the outgroup are said to be allocated to imagine performing the same action compared to 60% for the ingroup), we represent this expectation by setting *bout* to be a transformation of *bin*, such that .

#### Model comparison

We will assess the relative evidence for the self-categorization model and the alternative model provided by the data using a Bayes Factor (BF) calculated with the “Bridge Sampling” package in R. This Bayes Factor represents the probability of the observed data occurring under the alternative model relative to the probability of the observed data occurring under the self-categorization model. Specifically,