The role of serotonin and of perceived social differences in infering the motivation of others

Michael Moutoussis[1 *] , Joe Barnby[2] , Anais Durand[1], Megan Croal[3] , Robb B. Rutledge[1,4] and Liam Mason[1,5]

1. Department of Imaging Neuroscience, University College London (inc. Wellcome Centre for Human Neuroimaging and Max-Planck - UCL Centre for Computational Psychiatry and Ageing).
2. Department of Psychology, Royal Holloway University of London
3. COMPASS Pathways Plc, London
4. Department of Psychology, Yale University
5. Department of Clinical, Educational and Health Psychology, UCL

* Corresponding author:
Michael Moutoussis,
Email: m.moutoussis@ucl.ac.uk

Author Contributions:
Michael Moutoussis: investigation, data curation (supporting), review & editing (equal), conceptualization, formal analysis, methodology, software, writing – original draft preparation (all lead), **Joe Barnby:** Conceptualization, formal analysis, methodology, software, writing – original draft preparation (all supporting), review & editing (equal). **Anais Durand:** Investigation, data curation (equal), writing – original draft preparation (supporting), review & editing (equal)**, Megan Croal** Investigation, data curation (equal), writing – original draft preparation (supporting), review & editing (equal)**, Robb B. Rutledge** Methodology, project administration, supervision, writing – original draft preparation (all supporting) review & editing, (equal), funding acquisition (lead). **Liam Mason** Conceptualization, methodology, writing – original draft preparation, investigation, data curation, funding acquisition (all supporting), review & editing (equal), project administration, supervision (both lead).

Competing Interest Statement: None of the authors has any financial or intellectual competing interests in connection with the manuscript to declare.

Classification: Major: Biological and Social Sciences (dual). Minor: Neuroscience.

Keywords: Computational neuroscience, SSRI, attribution, prejudice.

Abstract

Attributing motives to others is a crucial aspect of mentalizing, which is disturbed by prejudice and is also affected by common psychiatric disorders. Thus it is important to understand in depth the neuro-computational functions underpinning mentalizing and social reward. We hypothesized that people quickly infer whether other's motives are likely beneficial or detrimental, then refine their judgment. Such 'Classify-refine', active inference models of mentalizing motives might improve on traditional models, and hence allow testing the hypothesis that serotonergic antidepressant drugs improve function partly by inducing more benign views of others. In a week-long, placebo vs. Citalopram study using an iterated dictator task, 'Classify-refine' models accounted for behaviour better than traditional models. Citalopram did not lead to more magnanimous attributions of motives, but we found evidence that it may help refine attributions about others' motives through learning. With respect to social differences, model comparison clearly indicated that ethnicity-dependent, in-task biases played no role in attributing motives for the large majority of participants. This is a very encouraging result which further research should seek to replicate, and, if replicated, celebrate. Lower subjective socio-economic status was associated with lower attributions of harm intent to others. We discuss how classify-refine social cognition may be highly adaptive. Future research should examine the role of Serotonergic antidepressants in clinical studies over longer time spans.

Significance Statement

We developed computational models to study how, in social situations, first impressions count a lot, but people also gradually refine their views to do justice to others. In our study, such 'classify-refine' models clearly outperformed the ones based on simpler learning. Modeling analyses indicated that refining one's views was facilitated by serotonergic antidepressants. We expected that those who perceived themselves as socially disadvantaged, and those interacting with people of color, would tend to attribute less benign motives to others. However in our low income, highly educated, young sample, we found evidence against both these biases. Further studies are needed to test how far these encouraging neuropharmacological and psychosocial findings apply to other populations.

Main Text

Introduction

Our relationships with other people are a key part of our well-being and social harmony, and our mental health suffers when important relationships deteriorate. Indeed, such is the influence of inter-personal function on mood that one of the therapies with best evidence for depressive disorders is Interpersonal Therapy, which focuses on navigating relationships. Similarly, perceiving others as a source of harm is very important in conditions ranging from PTSD (assaults) to social anxiety (humiliation) to paranoia (conspiracy). In the presence of Citalopram therapy, relationship factors are associated with improvements in depressive symptoms (1). However, whether the SSRI itself causes such improvements, and what the social-cognitive mechanisms involved might be, is as unclear as it is clinically important.

We asked whether week-long treatment with an SSRI antidepressant affects perception of others, depending on 1) their actions and 2) their apparent ethnicity. We used the attribution-focused repeated dictator task, with the aid of a rigorous computational model of attributions about others. We examined the pre-registered hypotheses that 1) Citalopram would result in attribution of more beneficial motives overall and 2) more negative attributions would be made for ethnically out-group others (2). At the same time, we sought to deliver a task and a validated model for probing attributions in mood disorders, and potential treatment mechanisms affecting those. In our improved task, participants saw four different Dictators would could make either fair (5:5) or unfair (10:0) splits of 10 pence, between themselves and the participant (2 Dictators 80% fair, 2 Dictators 20 % fair). Participants had to make

attributions of motivation, but also to report their expectations about fair behaviour by the Dictator (Fig. 1).

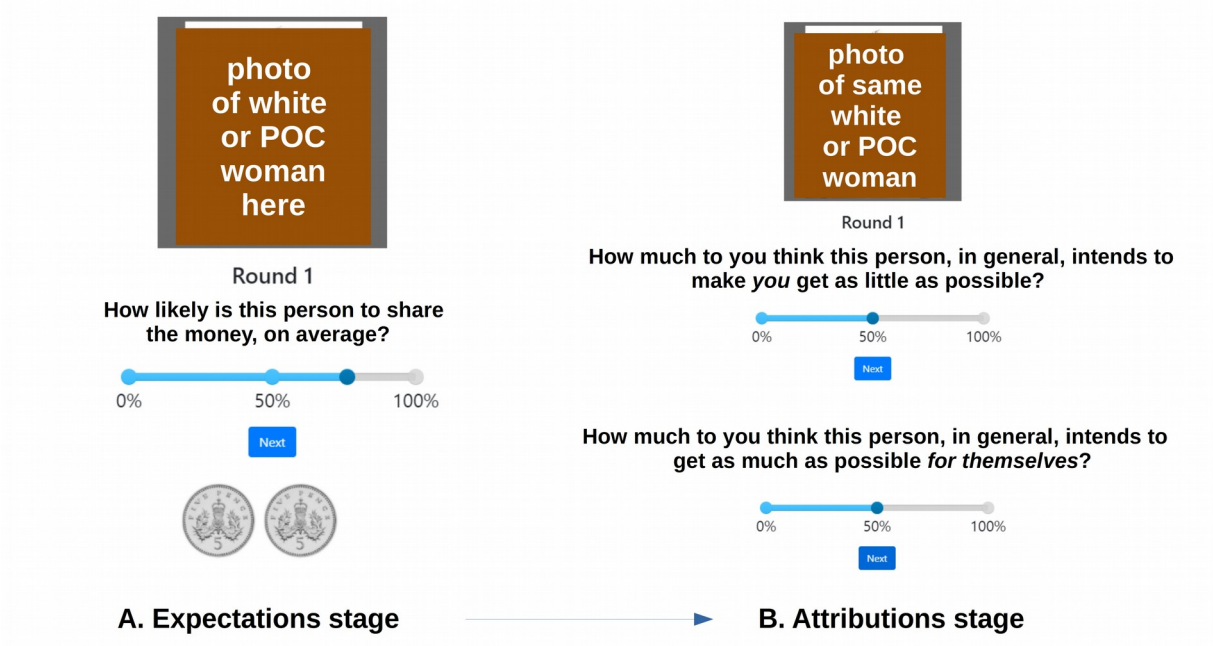

Fig. 1. One trial of the revised iterated dictator task, a.k.a. 'Sharing game'. **A.** Expectation stage. Participants were asked to estimate the probability that this Dictator would split two coins fairly. **B.** Having observed the split (not shown here), participants have to infer the degree of Self-interest and Harm-intent motivating the Dictator. The order of these questions was randomized across trials. The expectation stage preceded attributions within a trial, but computationally it is best thought as being the result of updates and beliefs formed on the basis of the observations made so far, especially in the trial before. That is, given one's learning, attributions are independent of the fairness reporting of the first stage. Unlike many economic games, we displayed ecologically valid Dictator images and modeled whether they elicited motivational attributions more beneficial or detrimental to the participant.

Belief-based models have performed well in attribution tasks, over and above associative models (3, 4), but the initial characterization of others appears faster than most models suggest (5, 6). We took inspiration from recent work on black-and-white thinking (7) to hypothesize that in interpersonal situations, people attempt to quickly classify others' attributes as 'beneficial or detrimental' to themselves. Then, they may refine their beliefs through learning, as per our previous associative and Bayesian models. Classifying into a small number of categories is cognitively easy and does not necessarily stop people from asking, "she's selfish, but how selfish?". We implemented models much like our previous ones (3, 4) within the active-inference framework, which naturally accommodates this process of fast classification, followed by belief refinement (8).

In the task participants were presented with partners having two salient dimensions of motivation: Harm-intent (HI) and Self-interest (SI). In classify-refine models, agents could classify the partner as either beneficial (low HI or SI) or detrimental (Fig. 2A) along each dimension - but could then refine the nature of these classes. Generative models also included a Partner's response function, which probabilistically furnished fair or unfair splits at each trial (Fig. 2B). In classic models, as in our previous work, agents updated belief distributions over fine-grained grids of SI and HI, which were subject to refinement.

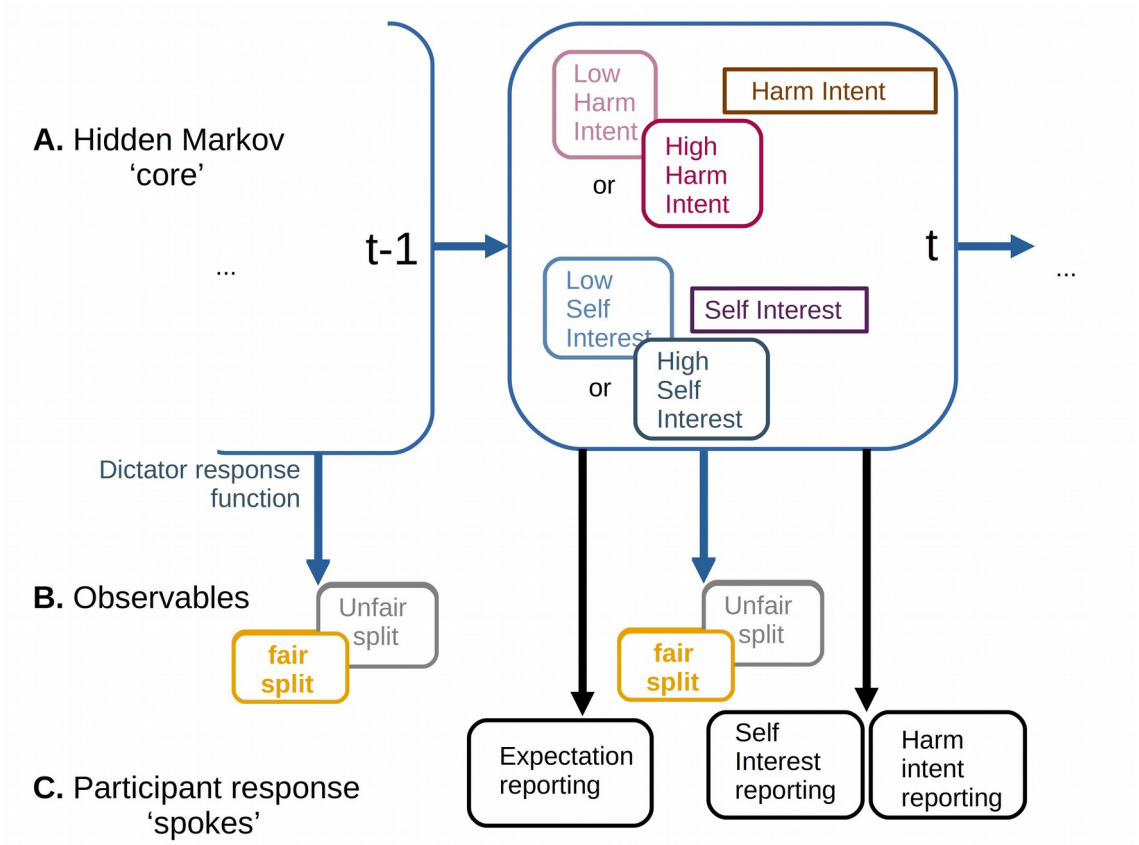

Fig. 2. Process diagram of the classify-refine model. Pure learning, or pure inference, models have many levels, or continua, for attributions instead of two, much like the response scales that our experiments make available to the participants. However, they also have a fixed map between the level of a partner's disposition and their policy (likely responding).

Results

Seventy-four participants completed the Sharing Game (revised iterated dictator task) on at least one occasion (44/74 self-identified Female, 29/74 Male, 0/74 other, 1 missing). 42 participants were randomized to Citalopram. The commonest ethnicity was Chinese. The sample was highly educated, young (median age = 25), and of low income (Supplemental Fig. S2). 66 participants completed the Sharing Game again one week later. At both baseline and follow-up participants faced four different partners, for 12 trials each. The partners were represented by naturalistic photos of people, who were all female but of varying apparent age and ethnicity. We fitted participants' expectations about how the Dictator would split the sum, and their attributions of the likely self-interest and harm-intending motives of the Dictator, using models parameterized as per Table 1, and further described below.

Table 1. Meaning and role played by model parameters

Parameter	Meaning	Role in winning model
Replication parameters (analogous to previous winning models)		
initHarmInt (a.k.a. *pHI0*)	central tendency of **init**ial, or prior, beliefs over attribution of **Harm Int**ent	Included
initSelfInt (a.k.a. *pSI0*)	central tendency of **init**ial, or prior, beliefs over attribution of **Self Int**erest	Included
IntentAttrEv (a.k.a. *dEv*)	certainty of prior belief in positive or negative **Intent Attribution**. 'd' refers to the active inference convention for prior beliefs over states, 'Ev' for amount of **ev**idence.	Included
evidRatio (a.k.a. *EvRat*)	**Ratio** of initial **evid**ence over Harm Intent vs. Selfish Intent	Excluded: ratio=1 in winning model.
decisPrec (a.k.a. alphaPrec)	baseline **decis**ion **prec**ision, or inverse-temperature (=mean of prior on expected free energy precision)	Included
wHarmInt (a.k.a. *wH*)	**w**eight of **Harm Int**ent in Dictator's policy fairness	Included
wSelfInt (a.k.a. *wS)*	**w**eight of **Self**ish **Int**ent Intent in Dictator's policy fairness	Included
fairnessB (a.k.a. *w0*)	**B**ias (intercept) in Dictator's policy **fairness** function	Included
λ_{other}	learning rate measure from one dictator to the next	Included
Exploratory parameters (newly introduced)		
typingConf, (a.k.a. *aEv*)	**Conf**idence certainty of prior belief over **typing** the level of niceness of ‘positive’ and ‘negative’ character classes. 'a' refers to the active inference symbol for the **likelihood** map, 'Ev' for amount of **ev**idence.	Included
learnRetn (a.k.a. ω)	extent of **learn**ing **ret**ention from trial to trial	Included
POCbias	Bias in expecting more positive or negative attributions for non-white Dictators.	Excluded: bias=0 in winning model

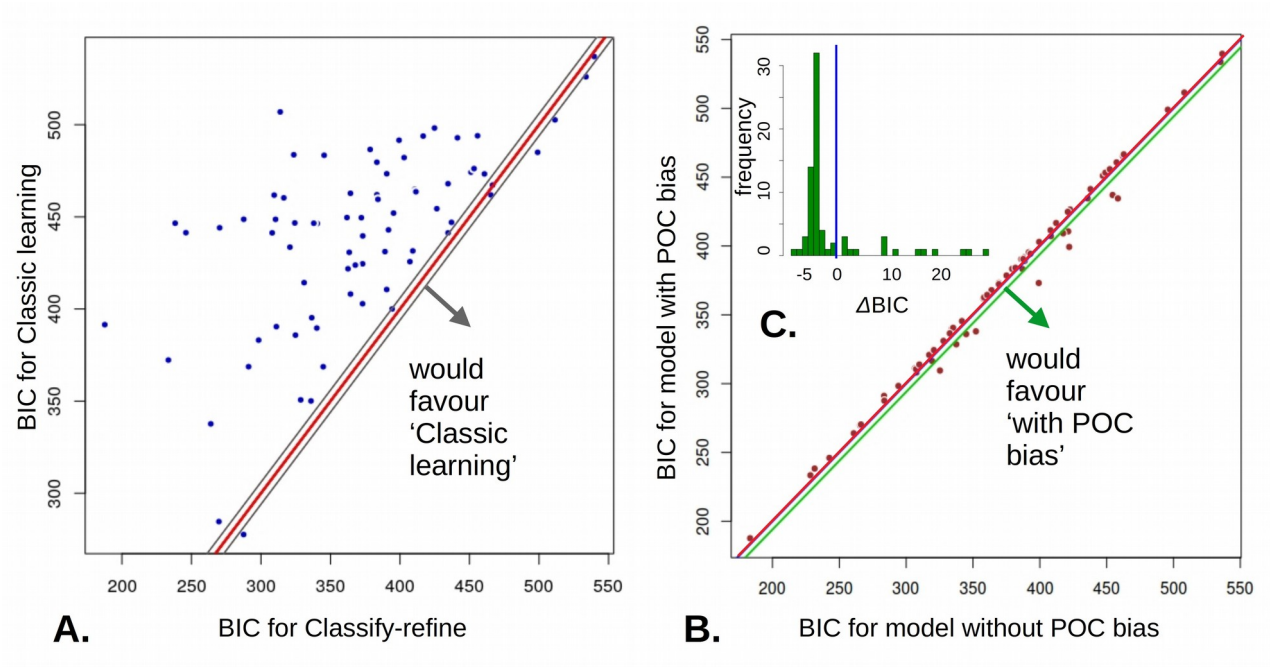

Fig. 3. Key model comparison results. Red lines = BIC equality. **A**. Classify-refine models clearly outperform classic learning. For 64/74 participants, the BIC difference is 6 or more (i.e. to the left of gray, diff=+6 gray line; conventionally at least modest evidence. Right gray line is diff=-6). Blue line is identity. **B, C.**: Coalitional bias based on ethnicity. **B.** The model without person-of-color bias gave a more parsimonious fit for 63 out of 74 participants (dots above the equal BIC line). For 10/73 participants, the model including a POC bias parameter gave an advantage of at least 6 BIC points ('modest evidence') **C.** (inset) Histogram of the difference, showing a main peak for the simpler no-bias model and a rightwards tail for the more complex model.

Refined learning accompanies fast classification

In the baseline sample, a model whereby people first classified their interaction partner as either having high or low harmful intent and self-interest, and then gradually refined their response after subsequent information was available, was clearly superior to a model of learning over a fine scale of attribute values (the 'classical learning' model; Fig. 3A. median BIC 372.2050 vs. 446.69, Wilcoxon rank sum p < 1E-5). People refined the attributions they made over trials more slowly if they had a higher parameter, *aEv*. This quantified the initial amount of evidence underpinning the so-called a-matrix, i.e. the beliefs that each type of partner would follow each specific policy (see Table 1 and Methods). Further model comparisons performed on the baseline data indicated the necessity of including learning from one dictator to the next (λ_{other} in Table 1) and imperfect memory for retaining learning from trial to trial (ω). We then fixed one parameter at a time, in order to discover more parsimonious models, or replicate our previous successful models as per pre-registration. The necessity for all parameters replicated, except the need for separate uncertainties over prior Harm and Selfishness intent (analogous to uSI0 = uPI0 in our previous model, and replicating novel work (9)). The only remarkable correlation between parameters was between *initHarmInt* and *initSelfInt* (baseline r=0.438, p_{uncorr} = 0.000170, follow-up r=0.428, p_{uncorr} = 0.000372, Fig. S6).

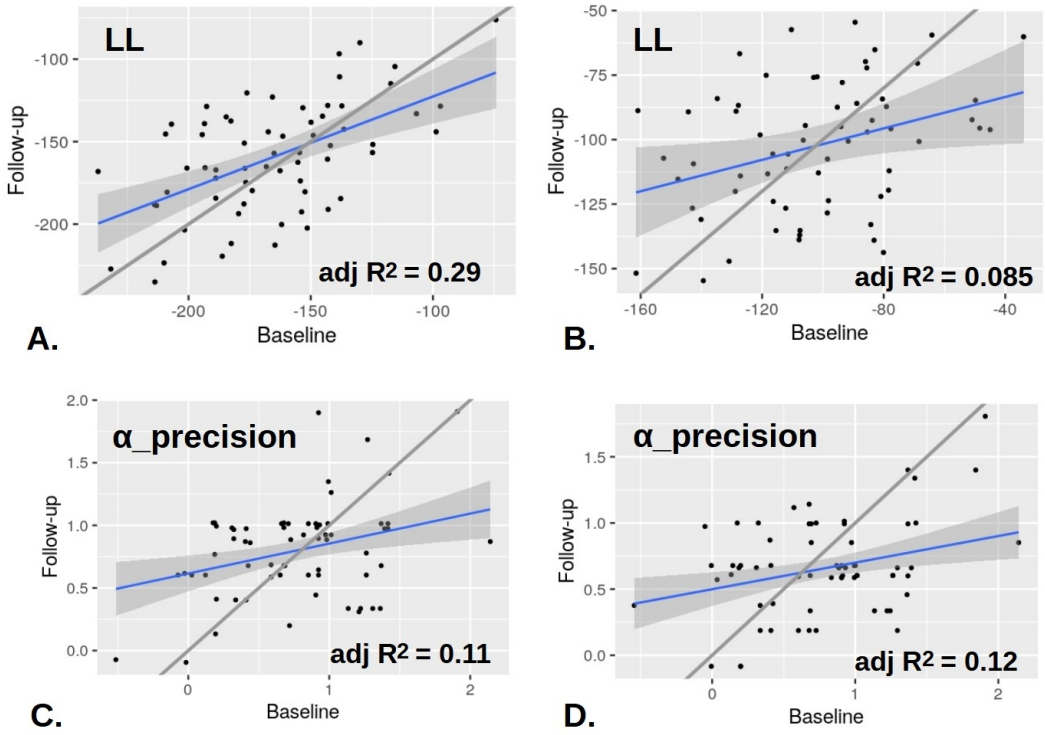

Figure 4. Task and model stability. **A., B.** Log-likelihood was the most stable measure, but in the prediction-attribution task more variance was shared between the baseline and follow-up, typical of the general pattern. However, for the learning over partners and decision noise, here in **C., D.** the attribution only, established task performed as well.

Coalitional factors: lower subjective status may attenuate attributions of harm intent

To examine the pre-registered hypothesis that coalitional stereotypes would modulate attributions of harmful intent and self interest, we introduced a parameter that shifted the effective attributions as they were communicated from the 'core belief' module to the 'Dictator response function' (Fig. 2A to B), based on a coalitional feature that is very important in terms of stereotyping, namely the apparent ethnicity of the Dictator. This modeling approach has specific advantages, compared to a post-hoc analysis (see Methods).

Happily, we found evidence, in terms of model fit, against our hypothesis that a positive or negative bias towards people of color (POC) would improve model fit. The model including apparent ethnicity had median BIC of 372.2, vs. for 366.2 for not including ethnicity, Wilcox. p = 8.5E-5). However, 10 of 73 participants appeared better fit by a model including POC bias (Fig. 2B, C).

In pre-registration, we hypothesized that across-participant coalitional threat, operationalized as perceiving oneself as of lower socioeconomic rank, would increase harm-intent attributions – but we found evidence for the opposite. Average harm attributions increased with Subjective SES (SSES) score, beta=0.109, Std.Err.=0.048, p=0.030. This was unchanged controlling for testing wave and drug (`HIAv ~ SSES + wave*drug + (1|participant)`: p=0.034, beta=0.109, Std.Err.=0.050).

The success of the winning model replicated between baseline and follow-up waves

A winning model over the baseline sample motivated the hypothesis that the same model would be the best at follow-up. This hypothesis held, in that the same model had the best total BIC in the follow-up testing. Hence, we examined the hypothesis that each parameter of the model would show stability, i.e. that it would be correlated between baseline and follow-up. To do this, we regressed the follow-up values of the parameters on the baseline ones, controlling for group allocation (placebo vs. SSRI). We found evidence for stability of the following parameters: *initHarmInt* (p=0.0096, adj.R^2 = 0.096 excl. one outlier; see Methods), decision noise *decisPrec* (p=0.0024, adj.R^2 =0.11), prior attribution certainty *typingConf* (p=0.00011, adj. R^2 = 0.21 excl. one outlier), bias *fairnessB* (p=1.5E-5, adj. R^2 = 0.253 excl. one outlier), memory *learnRet* (p=0.00122, R^2 =0.132), and learning over dictators λ_{other} (p=0.012, R^2 =0.071). Other parameters were poorly correlated (pS0: p=0.194, R^2 =0.064; dEv: p=0.731, adj.R^2 = -0.032 excl. one outlier; wH: p=0.806, R^2 =0.010), wS (p=0.436, R^2 =-0.022). As often found, model fit was the most stable measure (Log-likelihood: p=2.17E-6, adj. R^2 = 0.287). Again as expected, model fit improved on re-testing (See Fig. S4, B. vs. A.)

Including predictions along with attributions, improves task and model stability

We the fitted the equivalent of the winning model only to the harmful intent and self-interest attributions (i.e., approximating the previous version of the task). Here, fewer measures were correlated at conventional levels of significance, and baseline values generally explained less variance of the follow-up ones. As expected, model fit was most stable (log likelihood: p=0.0074, adj. R^2 = 0.085) and the overall bias *fairnessB*, decision noise *deciPrec* and learning over dictators λ_{other} also showed significant correlations (w0: p=0.012, adj. R^2 = 0.106; alphaPrec: p=0.00202, adj. R^2 = 0.122; λ_{other}: p=0.0208, adj. R^2 = 0.055; Fig. 4). However, the other measures did not attain conventional significance by this simple measure (p-value for initHarmInt: 0.235, initSelfInt: 0.90, aEv: 0.814, typingConf:0.108, wHarmInt: 0..228, wSelfInt: 0.948, learnRet: 0.155). The above is as per pre-registered hypothesis C.

Citalopram may reduce typing confidence, thus enhancing refinement of views

We first assessed the stability of the measure of average attribution levels for an individual per testing wave, as we pre-registered the hypothesis that Citalopram would reduce this measure levels across waves. These measures showed good stability, as shown by correlating follow-up attributions with baseline ones, while controlling for the effect of Citalopram. Average harm-intent attributions, HIAv, correlated at p=0.0015, beta=0.442, R^2 = 0.135. Average Self-interest, SIAv, showed p= 0.0023, beta=0.409, R^2 = 0.152. Predictions (predAv) showed p=1.17E-5, beta=0.437, R^2 = 0.249.

We used linear mixed effects analysis, `measure ~ wave*drug + ( 1 | participant)` to assess the effect of Citalopram. Contrary to our hypothesis, we found no evidence that Citalopram affected harm-intent attributions, p=0.11 for the `wave*drug` term for `measure`=HIAv. However, there was modest evidence that self-interest attributions were *enhanced* by Citalopram (`wave*drug` for SIAv: beta = 0.269, Std.Err.=0.134, p =0.0484). Average expectation was unaffected, p for predAv = 0.48.

We found a possible novel effect of Citalopram on the confidence of typing characters, *typingConf* upon examining model-based measures. Otherwise, these measures were consistent with the analysis of average attributions, i.e., there was a modest effect of SSRI increasing *initSelfInt* (pSI0: beta=1.32, Std.Err.=0.62, p=0.037), consistent with the SIAv result above. The possible novel effect was a decrease *typingConf* by Citalopram (aEv: `wave*drug` p=0.023, beta=-1.18, St.Err.=0.50). This would make the attribution-refining process faster. Notably, this was on a background of *initSelfInt* reducing with wave (pS0: p=0.014, beta=-2.612, Std.Err.=1.033), unlike *initHarmInt* (pH0: p=0.153). The wave*drug effect for all other measures was unremarkable (p values: LL: 0.729, *initHarmInt*: 0.217, *dEv*: 0.743, *alphaPrec*: 0.7638, *fairnessB*: 0.568, *wHarmInt:* 0.979, *wS*: 0.11*, ω:* 0.303*,* λ_{other}: 0.898*)*

Psychometric findings and the effect of Citalopram

We administered the Hypomanic Personality Scale (HPS) at baseline, and the Patient Health Questionnaire (PHQ-9), Rosenberg Self Esteem (RSE), Mood and Anxiety Symptom Questionnaire (MASQ), Perceived Stress Scale (PSS), Generalized Anxiety Disorder - 7 (GAD), and the Achievement Motivation Inventory (AMI) at both baseline and follow-up. We examined the dependence of follow-up scores on Citalopram treatment, controlling for baseline scores, gender, SSES and HPS. Citalopram had no statistically significant effect on any of the scores (Supplemental Fig. S2).

In order to reduce dimensionality and increase sensitivity, we performed a factor analysis on the baseline state-like symptoms (i.e., not the HPS), resulting in two factors on the basis of parallel analysis and scree plot (Fig. S5). We tested the validity and stability of this by deriving factor scores on baseline and follow-up based on the baseline loadings only. The first factor scores, 'anxious depression' correlated r=0.791, p < 1E-10. The second factor scores, 'stressful amotivation', correlated at r=0.803, p < 1E-10, both showing good stability and hence validity. Anxious depression was not affected by SSRI (`anxDep ~ wave*drug + (1 | participant)` gave p=0.759) but stressful amotivation marginally improved, p=0.052, beta = -0.25. Exploratory linear regressions found no evidence that variance in psychometric ratings, or changes in these ratings, was related to model parameters.

Discussion

In this work, we set out to test whether antidepressant treatment may help in affective disorders by promoting attributions of more pro-social motives to others. We sought to refine our understanding of the neuro-computational basis of attributions of motives through active-inference modelling, and examined whether factors often postulated to recruit subtle coalitional dynamics, namely apparent ethnicity and subjective socioeconomic status, affected attribution of magnanimous motives. One week's treatment with Citalopram did not appreciably change self-reported affective states, and, contrary to our hypothesis, did not result in more magnanimous attributions (2). Importantly, we found strong evidence in favour of our hypothesised active-inference models, which postulated that participants first classified others' attributes as positive or negative, and then sought to refine attributions. Contrary to our hypotheses regarding coallitional factors, ethnicity did not affect positive or negative attributions, supporting a view of our participant's decision-making as unbiased in this respect. In addition, low subjective socioeconomic status reduced rather than increased attributions of harmful intent.

In line with our pre-registered hypotheses regarding modelling, this study replicated the set of features, or parameters, needed to model the data according to our previous work - with one exception (4). The exception was that prior belief uncertainty over Harm-intent and Self-interest could be condensed into a single uncertainty parameters. This is important, as the replication entails a task in a different population, in a laboratory based pharmacology experiment, with added task features of naturalistic photos of the 'partner', and questions about expected frequency of fair outcomes (Fig. 1). We found good evidence that the version of the task including these new questions was more stable, or reliable, than the original - as we predicted ((2), main hypothesis C). Our close replication of a robust correlation between intial beliefs in harm-intent and self-interest (3) suggests that future studies consider their commonality (harm-benefit intent) and contrast (other-self intent).

The astute reader may have noticed that we presented no tests of two pre-registered secondary hypotheses, because these have been rendered moot by the success of Classify-refine models. Indeed, it was work on the first hypothesis (H.) that partly led to these models. Specifically, when we fitted models to each 'Partner' separately, we noticed that the fitted values of initHarmInt and intSelfInt followed the actual proportions of fair returns experienced in the block ahead. This could be explained by the very fast learning shown by Bone, Pike and co-workers and others (5, 6), who noticed that initially high learning learning rates decayed exponentially, more quickly than a classic Bayesian learning pattern.

Classify-refine learning is much like posterior-belief-based update allocation (10), and has been considered in some detail by Story and co-authors in the context of 'black and white thinking', using active inference (7). Refining this, we suggest that classify-refine updating may be a normal, adaptive social strategy, rather than elicited by the task itself, or mostly found in suboptimal black-and-white thinking. It is advantageous to rapidly distinguish friend from foe, and then do more justice to the true nature of others. Here, maladaptive black-and-white thinking would be about the failure of the 'refine' process, e.g. by circular causation tit-for-tat behaviour. The success of classify-refine models suggests a mechanistic explanation for the decaying learning rates of the model or Pike, and for the inability of classic models to account for fast initial learning.

Citalopram treatment did not result in attribution of more pro-social states, either in terms of increasing the proportion of fair returns that participants expected, their average HI and SI attributions, nor the parameters pHI0 and pSI0. Indeed, there was weak evidence for a relative increase in average SI and pSI0 (both at the 0.05 level uncorrected). Citalopram did not affect participants' psychometric measures either (which were healthy), so the possibility remains that in may only ameliorate depressive attributions. We found intriguing evidence – much in need of replication – that Citalopram may reduce certainty over character types, allowing for faster learning of others' true nature, and that it may reduce stress-induced amotivation. The simplest interpretation of our findings, however, is that treatments for depression that ameliorate inter-personal cognition may act via pathways only peripherally, or indirectly, affected by antidepressants (11).

Our findings about the role of coallitional processes were unexpected, and in a sense cause for celebration. Models that included a parameter that would change motivational attributions depending on the ethnicity of the partner, a parameter which was able to capture any direction of such effect at the individual level, were rejected as they did not improve model fit – although a few individuals may be prone to this bias. This contrasts with literature showing affective biases against people of color (12) and is consistent with decision-making findings where appropriately educated participants behaved fairly towards POC others, even despite low-level affective bias (13). Lower socio-economic self-ranking was associated, again contrary to our hypotheses, with lower attributions of harm intent. We note that our participants were 74 predominantly young, highly educated women from BAME backgrounds, the most common ethnicity being Chinese. Groups with a different experience of social disadvantage, or indeed advantage, may attribute motives differently. Although, therefore, our sample is not representative of the UK population, our inability to detect signs of implicit bias and socioeconomic-based suspiciousness are cause for celebration.

In terms of limitations, our study did not involve participants with clinically significant affective symptoms, nor did Citalopram have an effect on their mood and anxiety, rendering moot the question of whether the attributional problems of depression are ameliorated by SSRIs. It would be important, therefore, for CBT treatment trials to include social-cognitive tasks such as the one used here. Similarly, the happy absence of detectable biases related to low subjective socio-economic status or ethnicity need replication in samples more representative of the general population.

In conclusion, Classify-refine models appear to hold much promise for social neuroscience, have normalizing implications for 'black-and-white' thinking, are well served by the active inference framework, and have learning-relevant parameters that may depend on Serotonin function. SSRIs appear to have little effect on the motives and expectations healthy people have from each other. Importantly, computational neuroscience studies should be equipped to interrogate burning issues such as social inequality or racial biases, but, as importantly, also to elucidate mechanisms behind 'the glass being half full'.

Materials and Methods

Study population

Healthy participants were recruited from a University College London subject pool and gave informed consent for a week-long Citalopram 20mg vs. placebo study, of which the present experiment was part.

This was approved by the UCL Ethics Committee, ID 19601/001. All participants were UK residents, and were tested in person at the laboratory at the beginning and end of the week. They had no history of psychiatric or neurological disorder and pledged not to become intoxicated by drugs or alcohol during the study. Of the 42 Citalopram participants, 25 experienced subjective drug effects (all minor) and correctly guessed that they took SSRI.

Task and modeling

We sought to increase the stability and sensitivity of tasks, so as to increase the accuracy of detecting individual differences, including drug effects. We thus increased task length by one 12-trial block, but most importantly by collecting more data per trial (i.e., Fig. 1A), and testing a new set of computational models. We considered models that had a central 'learning core', whose generative model was a simple hidden Markov process (HMM), and was implemented as a one-level, 12-trial Active Inference MDP. Each also had three 'reporting' processes, generating fairness expectation, Harm-intent attributions and Self-interest attributions. Fig S1 serves as an illustration and reference to the workings of the classify-refine model. More detailed explanations and equations follow in the Supplement. Here we describe the key features of the models.

All models were derived from a published, successful Bayesian model (4). Agents inferred the character of others over a grid of Self-Interset $\times$ Harm-intent values, starting from indepently parametrized priors over each dimension (Table 1). In their generative model, each character (SI, HI) determined a probability of fair-split through a logistic function (Supplement Eq. S1). Observed splits allowed them to invert this model and update their beliefs. Beliefs held at a central 'core' were translated to self-reports via three very simple active-inference processes, 'fairness prediction', 'HI report' and 'SI report'.

In Classic models, HI and SI grids had 6 bins each with fixed boundaries, covering 0-100%, but for the classify-refine models, the central core considered only two states along each attribution dimension, giving four combinations. These, however, did not have fixed location, but could be refined via learning. Initially, each is taken to include low (0.05) and high (0.95) levels of each attribute, so as to represent the psychological meaning of 'low' and 'high', and to span the meaningful range of each attribute. Crucially, uncertainty over the location of the bins was parametrized by *typing confidence* parameter *aEv* (Table S1). Beliefs about location of the 'high' and 'low' states were refined by allocating the evidence for 'fair' or 'unfair' split at each trial in proportion to the (posterior) belief that said state underlied the trial (10).

Coalitional dynamics did not affect learning, but simply modulated the beliefs outputted from the learning to the reporting processes. That is, given the same evidence, both HI and SI beliefs were shifted in the same direction for a white person vs. a person of colour. The magnitude and direction of this shift was fitted individually to each participant, allowing modelling of the coalitional bias in any direction - including having making more optimistic attributions for their out-group.

Model fitting

Maximum-a-posteriori fitting with weakly informative priors defined over native parameter space was used for all models (14). The sum-log-likelihood at the MAP estimate was then used to derive Bayesian Information Criterion (BIC; corrected for small samples) and Akaike IC (AIC) for each participant. In cases where a parameter was fitted across the whole participant sample, we modified the complexity penalty associated with that parameter to be consistent with BIC/2 being an approximation to model evidence.

Gradient-descent methods (including matlab *fmincon* and SPM *spm_dcm_mdp*; (15, 16)) encountered problems with local minima, hence a custom adaptive grid search optimisation with multiple initial conditions was used.

Regression analyses

Weak fitting priors in native parameter space mostly resulted in normal-like parameter distributions in transformed space, where regression analyses were performed. However, it also allowed for outliers in this transformed space. We therefore first used robust regression *rlm* in R (17), and reported hypothesis testing is based on the t-value of the coefficient in question, e.g. `rlm( pHI0_follow-up ~ pHI0_baseline + drug_group + sex + subective_socioecon_status )`. We then performed the equivalent OLS regression using *lm* and identified points with Cook's distance > 1 and/or standardised residual distance > 3 from the theoretically predicted on Q-Q plots. We excluded these suspects from further analyses.

After the very few outliers were excluded, we performed the longitudinal analyses based on linear mixed effects analyses as above.

Acknowledgments

We wish to thank Giles Story, Alex Pike, Kate Button, Janina Hoffman, Katie Hobbs, Ryan Smith and the WCHN Methods group for their support. The Wellcome Centre for Human Neuroimaging is supported by a Platform Grant from the Wellcome Trust. The Max-Planck centre for computational neuroscience and ageing is supported by UCL and the Max-Plan Society. R.B.R. is supported by a NARSAD Young Investigator Award from the Brain & Behavior Research Foundation, P&S Fund. L.M. is supported by a Medical Research Council Clinician Scientist Fellowship (MR/S006613/1).

References

1. N. T. Joseph, *et al.*, Support and Undermining in Interpersonal Relationships Are Associated with Symptom Improvement in a Trial of Antidepressant Medication. *Psychiatry: Interpersonal and Biological Processes* **74**, 240–254 (2011).

2. M. Moutoussis, *et al.*, Do SSRIs promote more magnanimous attributions about others? (2022) https:/doi.org/10.17605/OSF.IO/HT95F (February 15, 2023).

3. J. M. Barnby, V. Bell, M. A. Mehta, M. Moutoussis, Reduction in social learning and increased policy uncertainty about harmful intent is associated with pre-existing paranoid beliefs: Evidence from modelling a modified serial dictator game. *PLOS Computational Biology* **16**, e1008372 (2020).

4. J. M. Barnby, M. Mehta, M. Moutoussis, The computational relationship between reinforcement learning and social inference in paranoia (2022) https:/doi.org/10.31234/osf.io/x4d3f (May 23, 2022).

5. J. Bone, *et al.*, Computational mechanisms underlying social evaluation learning and associations with depressive symptoms during adolescence (2021) https:/doi.org/10.31234/osf.io/9m7vr (January 27, 2023).

6. A. K. Hopkins, R. Dolan, K. S. Button, M. Moutoussis, A Reduced Self-Positive Belief Underpins Greater Sensitivity to Negative Evaluation in Socially Anxious Individuals. *Computational Psychiatry* **5**, 21–37 (2021).

7. G. W. Story, *et al.*, A Social Inference Model of Idealization and Devaluation (2021) https:/doi.org/10.31234/osf.io/yvu2b (September 1, 2021).

8. R. Smith, K. J. Friston, C. J. Whyte, A step-by-step tutorial on active inference and its application to empirical data. *Journal of Mathematical Psychology* **107**, 102632 (2022).

9. J. M. Barnby, V. Bell, Q. Deeley, M. Mehta, M. Moutoussis, D2/D3 dopamine supports the precision of mental state inferences and self-relevance of joint social outcomes. 2023.05.02.539031 (2023).

10. H. M. Dorfman, R. Bhui, B. L. Hughes, S. J. Gershman, Causal Inference About Good and Bad Outcomes. *Psychol Sci* **30**, 516–525 (2019).

11. C. L. Nord, *et al.*, Neural effects of antidepressant medication and psychological treatments: a quantitative synthesis across three meta-analyses. *The British Journal of Psychiatry* **219**, 546–550 (2021).

12. J. L. Davies, O. H. Turnbull, Affective bias in complex decision making: Modulating sensitivity to aversive feedback. *Motiv Emot* **35**, 235–248 (2011).

13. J. Correll, *et al.*, Across the thin blue line: Police officers and racial bias in the decision to shoot. *Journal of Personality and Social Psychology* **92**, 1006–1023 (2007).

14. M. Moutoussis, A. K. Hopkins, R. J. Dolan, Hypotheses About the Relationship of Cognition With Psychopathology Should be Tested by Embedding Them Into Empirical Priors. *Front. Psychol.* **9** (2018).

15. matlab, MATLAB (2019).

16. SPM development team, SPM (2022).

17. R Core Team, R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/. (2020).