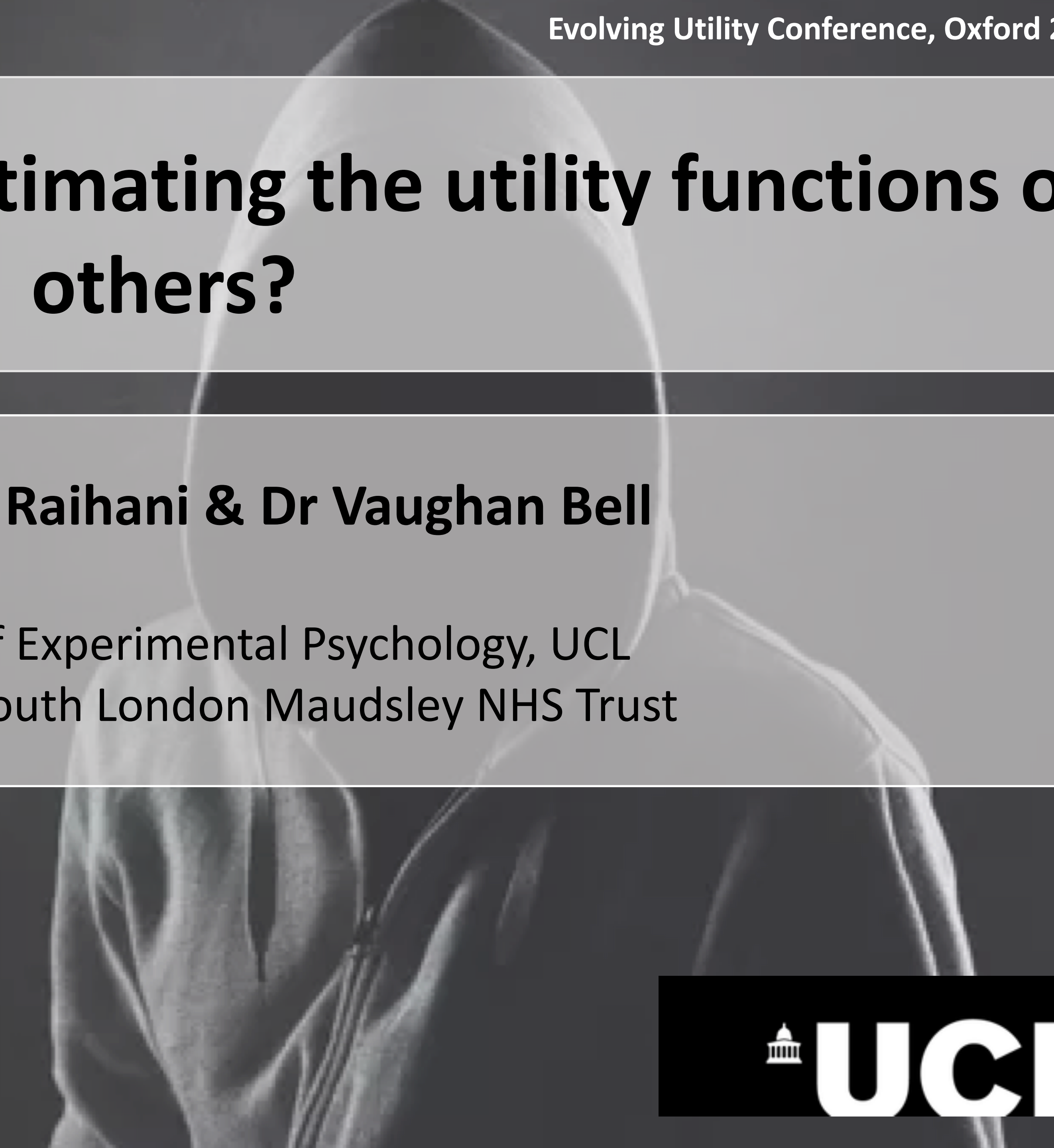
Evolving Utility Conference, Oxford 2018

Paranoia: biases in estimating the utility functions of others?

Prof. Nichola Raihani & Dr Vaughan Bell

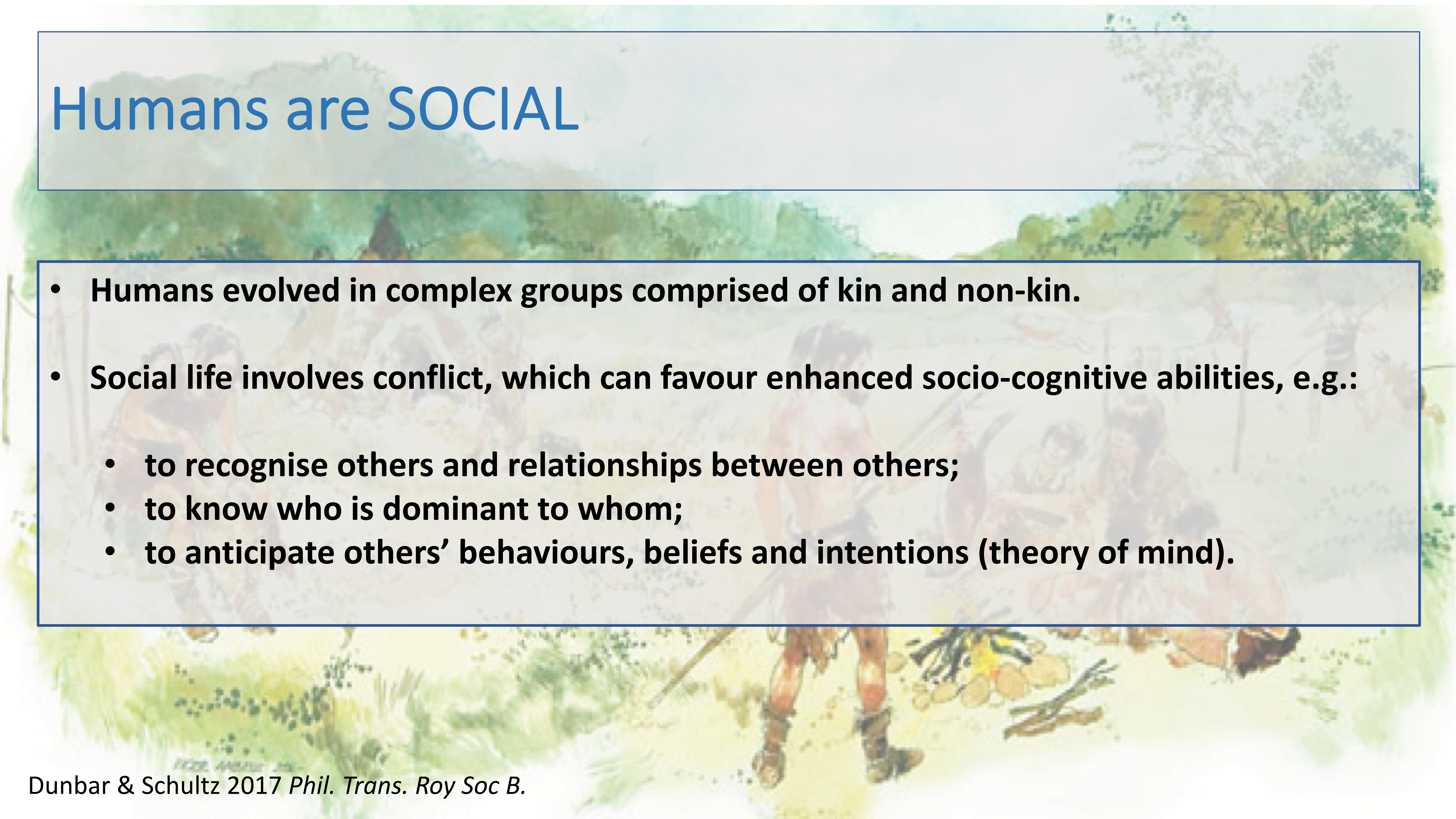
Department of Experimental Psychology, UCL
UCL Psychiatry South London Maudsley NHS Trust



THE ROYAL SOCIETY

UCL

Humans are SOCIAL



- **Humans evolved in complex groups comprised of kin and non-kin.**

- **Social life involves conflict, which can favour enhanced socio-cognitive abilities, e.g.:**

  - **to recognise others and relationships between others;**
  - **to know who is dominant to whom;**
  - **to anticipate others’ behaviours, beliefs and intentions (theory of mind).**

Dunbar & Schultz 2017 *Phil. Trans. Roy Soc B.*

Forms of social conflict in human societies

- **Lethal raiding**
- **Dominance competitions.**
- **Reputation damage, including witchcraft accusations, stigmatisation & gossip, leading to ostracism, persecution, maybe death.**

40p

IN DEPTH COVERAGE OF BRITAIN'S POLITICAL CRISIS

i

The essential daily briefing

FRIDAY
1 JULY 2016
Number 1,747
i@inews.co.uk
Twitter: @theipaper
facebook.com/theipaper

EURO 2016
Dewch 'mlaen Cymru!*

(*Come on Wales)

Et tu, Gove?

» Boris Johnson betrayed by key ally at the last moment – so quits his campaign to become Prime Minister

» Savage criticism after he walks away from Brexit fall-out: 'He's ripped Tories apart and knocked billions off people's savings'

» Theresa May is now the favourite to be Britain's next PM

» Rival candidate Gove says only he can deliver full Brexit

Interest rates could be cut this summer P11

The Somme: 100 years on, we remember P25

Woman wins appeal to use daughter's eggs P21

INSIDE PUZZLES P52 | MARK STEEL | THIS WEEK'S TV | AB FAB | METRONOMY

Daily Mail

20 unspoilt SEASIDE GEMS

GLORIES OF SECRET BRITAIN

SUPERB FREE PULLOUT INSIDE THIS PAPER

Exclusive: Health tourism loophole exposed

FOREIGNERS CHARGE NHS FOR CARE IN THEIR OWN COUNTRY

Daily Mail INVESTIGATIONS UNIT

Harper Beckham still uses dummy at the age of four

Daily Mail

FREE INSIDE TOMORROW

GREAT BRITISH BAKE OFF

SUPERB 16-PAGE RECIPE GLOSSY

As numbers break all records ...

MIGRANTS: HOW MANY MORE CAN WE TAKE?

Jamie's war on junk food ads in TV talent shows

Macfarlan et al., 2014, *PNAS;* Boyer et al. 2015 *Persp. Psych. Sci.;* Gershman 2016; Mace et al. 2018 *Nat. Hum. Behav.*

Raihani & Bell *in prep*

Social conflict: common themes

- Perpetrators use low-cost opportunities to harm targets.

- Targets can be singled out for belonging to different group, being ‘different’ (e.g. disabled), being high / low status, and blamed for unlucky ‘accidental’ misfortunes.

- Individuals in human groups face a persistent, yet relatively unpredictable threat of being selectively persecuted or singled out for harm by conspecifics. For some individuals / in some environments, this threat may be greater than for others.

- Selection for psychological threat-detection mechanisms that anticipate, avoid or deflect coalitional threat.

Avoiding social threat

- Ability to predict the intentions and motives of others (their utility functions) can help individuals avoid costly social conflict.

- But we observe behavior, not preferences.

- Behaviours often happen in ambiguous scenarios, leaving scope for variation and error in intention attribution.

- Paranoia might result in systematic alterations in the way people estimate others' utility functions. It also might be an adaptive response to environmental social threat.

- Biased threat detection might be a feature not a bug (error-management).

Paranoia - definition

- Tendency to attribute hostile intentions to others when true intentions are unknown or ambiguous.
- The most common presenting symptom of psychosis.
- But also common in the general population, ranging from mildly out-of-proportion socio-evaluative concerns to frank paranoid delusions.

Freeman & Garety 2000 *Brit. J. Clin. Psych;* Freeman et al. 2010 *Psych. Med.*

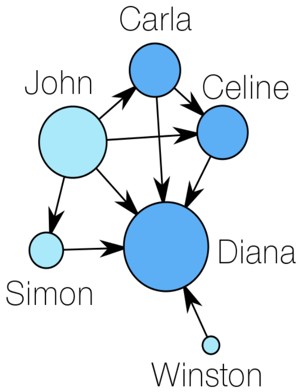
Evidence that paranoia is linked to coalitional psychology

- Paranoia is predicted by environmental variables that signify social threat, including:

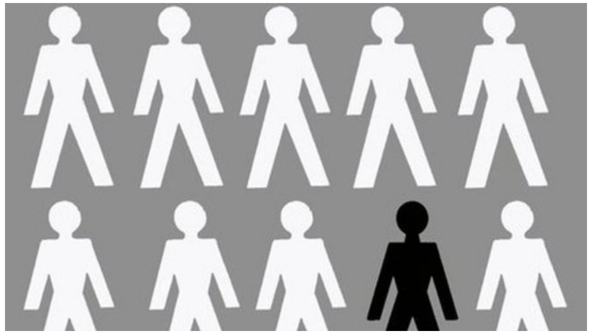
Being victimised

Being bullied

Being low status



Having a small social network

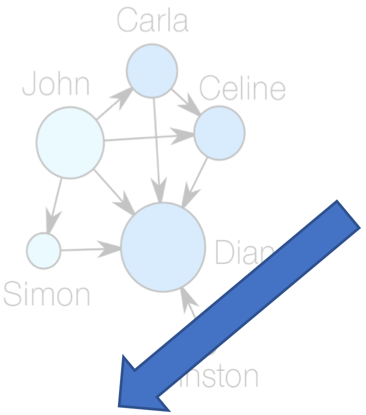


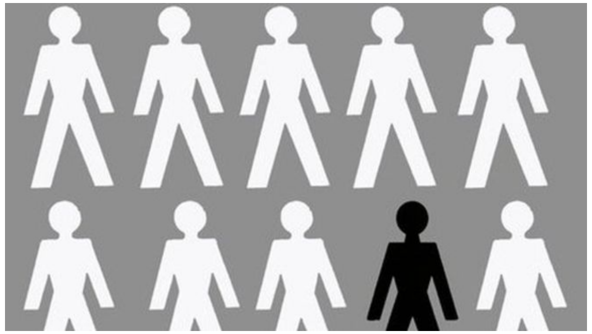
Being an ethnic minority

Shaikh et al., 2016 *Psych. Res.;* Kirkbride et al., 2008 *J. Brit. Psych.;* Gayer-Anderson & Morgan 2013 *Epidem. Psych. Sci.*

Evidence that paranoia is linked to coalitional psychology

- Paranoia is predicted by environmental variables that signify social threat, including:





BUT: living at high ethnic density buffers against this risk, as coalitional psychology model would predict

Being an ethnic minority

Bosqui et al. 2014 *Soc. Psych. Epidemiol.*

Experiments

- Exp 1. Does pre-existing paranoia result in a bias towards inferring harmful intent in social interactions?
- Exp 2. Does experimentally-induced social threat also bias estimates of others’ utility functions?
- Exp 3. How do biased estimates of others’ utility functions affect responses to social behaviour?

General approach

- Participants recruited via Amazon Mechanical Turk (online crowdsourcing platform, with access to more diverse sample than possible when using undergraduate pool).
- Large N (> 2,000 participants per study); pre-registered predictions; open data and code.
- Assess people for ***trait* paranoia**, using Green et al. Paranoid Thoughts Scale.
- Test-retest method, where participants recalled ~10 days after psychometric test to take part in experiment.
- **Live** social interactions, using game-theory paradigms.
- Statistical approach: multi-model selection with model averaging.

Quantifying trait paranoia

Ideas of social reference

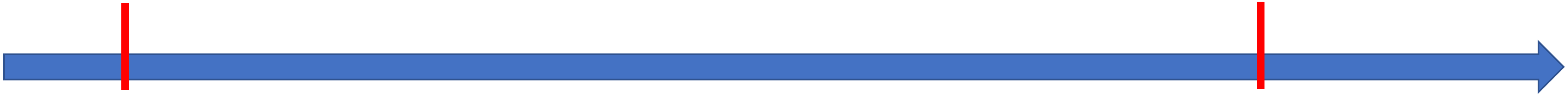
e.g. People talking about me behind my back upset me

Ideas of persecution

e.g. I was sure someone wanted to hurt me

General population mean: 48.8

Clinical mean: 101.9



Lowest score: 32

Highest score: 160

The Green et al. Paranoid Thought Scales (2008)

Exp 1: Does paranoia bias estimates of utility functions?

www.nature.com/scientificreports

SCIENTIFIC REPORTS

OPEN

Paranoia and the social representation of others: a large-scale game theory approach

Received: 8 February 2017
Accepted: 15 May 2017
Published online: 03 July 2017

Nichola J. Raihani[1] & Vaughan Bell[2,3]

Current definitions of paranoia include two key components: unfounded ideas of harm and the idea that the harm is intended by others. However, attributions of harmful intent have been poorly studied and mainly using artificial scenarios rather than participation in genuine social interactions where genuine resources are at stake. Using a large non-clinical population (N = 3229) recruited online, we asked people to complete a measure of paranoid ideation before playing a modified Dictator Game, where the 'dictator' can allocate money to the partner (the 'receiver'). Participants were allocated to the role of receiver or of an uninvolved observer; and evaluated to what extent they believed dictator decisions were motivated by (i) self-interest or (ii) harmful intent. All participants attributed more harmful intent to unfair as opposed to fair dictators. Paranoia had a positive effect on harmful intent attribution, for both fair and unfair dictators. Paranoia did not interact with attributions of self-interest. Importantly, highly paranoid participants attributed equally strong harmful intent to the dictator in the observer role as in the receiver role. This challenges the assumption that paranoia is mainly due to an exaggerated sense of personalised threat and suggests instead that paranoia involves a negative social representations of others.

Paranoia involves two key components: a person having unfounded ideas that harm will occur to them, and the idea that the harm is intended by others[1]. Current accounts of the formation and maintenance of paranoia include a cognitive style involving worry, negative thoughts about the self, interpersonal sensitivity, anomalous experiences, insomnia and reasoning biases as key components[2]. The threat anticipation model suggests that these converge and lead to an over-interpretation of potential harm to the self – a state that self-maintains in people with high levels of paranoia primarily due to anxiety-driven avoidance of disconfirmatory evidence[3,4].

Differences in threat perception have been well-established and studies have consistently found threat-related material engages attention to a far greater degree in clinically paranoid people (reviewed in ref. 5). It is unlikely that this is explained by differences in general 'theory of mind' ability, despite the obvious misattributions of mental state. Although past studies include both positive and negative findings[6], a meta-analysis of theory of mind in schizophrenia reported no overall association with paranoia[7]. However, numerous studies have investigated attributional biases in the interpretation of events, and have shown a general tendency for attributing responsibility to other people rather than situations for negative events[8].

Notably, attributions of whether harm is intended, rather than where responsibility lies for negative events, have been far less studied despite this being a key aspect of paranoia. Studies using the Ambiguous Intentions Hostility Questionnaire[9] have reported that people with high levels of non-clinical and clinical paranoia are more likely to rate vignettes describing ambiguous scenarios as involving hostile intent[9–12]. More recently, Freeman *et al.*[13] used a paradigm from Ames and Fiske[14] to test differing attributions from vignettes describing intentional and unintentional harm with identical consequences. Counter to the researchers' predictions, they found that people with high levels of paranoia were less likely to make attributions of intent and blame in the vignette describing intentional harm than people with low levels of paranoia but more likely to ascribe blame and intent in the vignette describing non-intentional harm. However, they suggested that these third-person vignettes may not have sufficiently engaged self-relevant concerns, and that if participants were directly involved in an interaction,

[1]Department of Experimental Psychology, University College London, London, United Kingdom. [2]Division of Psychiatry, University College London, London, United Kingdom. [3]South London and Maudsley NHS Foundation Trust, London, United Kingdom. Correspondence and requests for materials should be addressed to V.B. (email: Vaughan.bell@ucl.ac.uk)

SCIENTIFIC REPORTS | 7: 4544 | DOI:10.1038/s41598-017-04805-3 1

Raihani & Bell 2017 Scientific Reports

- In social dilemmas, unfair behaviour can reflect self-interest or a desire to harm the partner.

- Example utility function

- **U (*x* , *y*) = *x* + *ay***

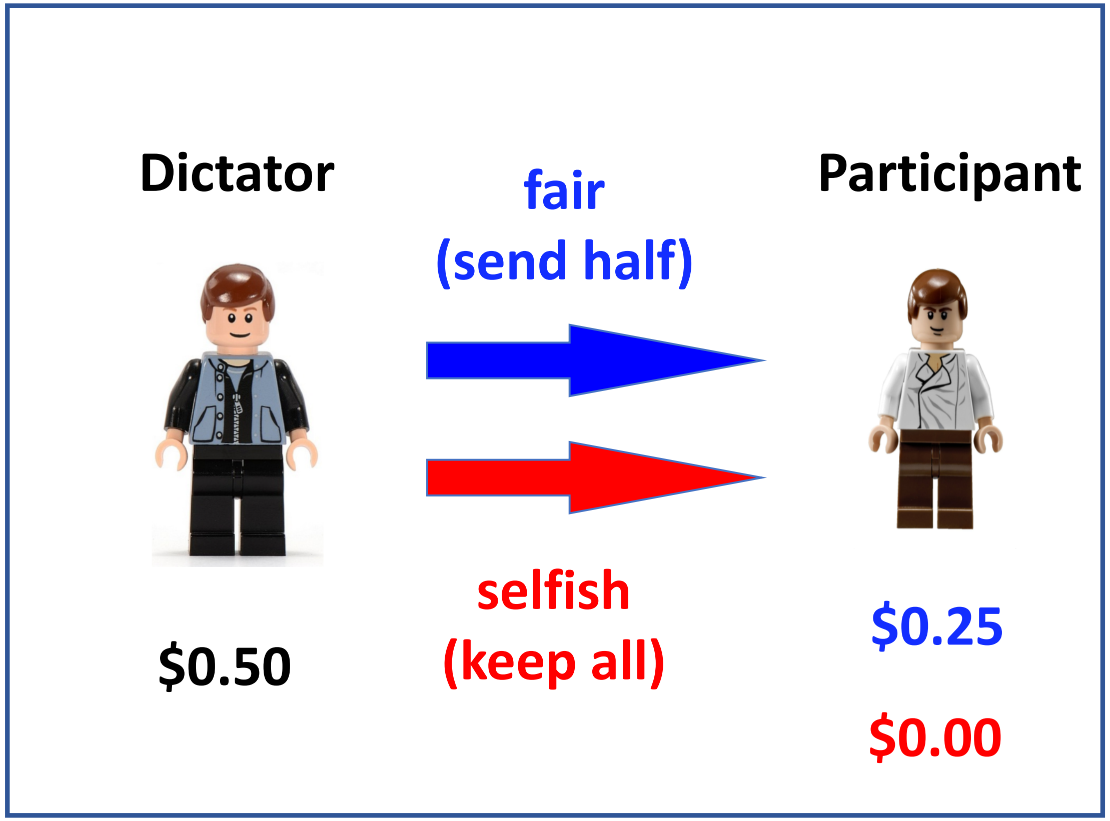
- Where ***x*** and ***y*** are payoffs to players **x** & **y**, and a represents the weight ***x*** places on ***y's*** payoff.

- a = -1 implies **x** prefers to maximise payoffs relative to y
- a = 1 implies **x** prefers to maximise joint payoffs
- a = 0 implies **x** prefers to maximise individual payoffs

Experiment 1. Method

- N = 3,229 people (53 % female; age: 18-80 years).
- *Step 1*: Green et al. Paranoid Thoughts Scale -> trait paranoia.
- *Step 2*: Measure how trait paranoia affects attribution of harmful intentions versus attributions of self-interest to partner in Dictator Game.

The Dictator Game



Dictator

$0.50

fair
(send half)

selfish
(keep all)

Participant

$0.25

$0.00

- The Dictator Game (Kahneman et al. 1986).

- Motives underpinning dictator decisions are **ambiguous** with respect to harmful intent and could reflect Dictator's desire to:

  - **earn more money (self-interest)**

  or

  - **prevent partner from getting any money (harmful intent)**

Attributions of (i) self interest and (ii) intent to harm

- Participant assigned to role of receiver / uninvolved observer

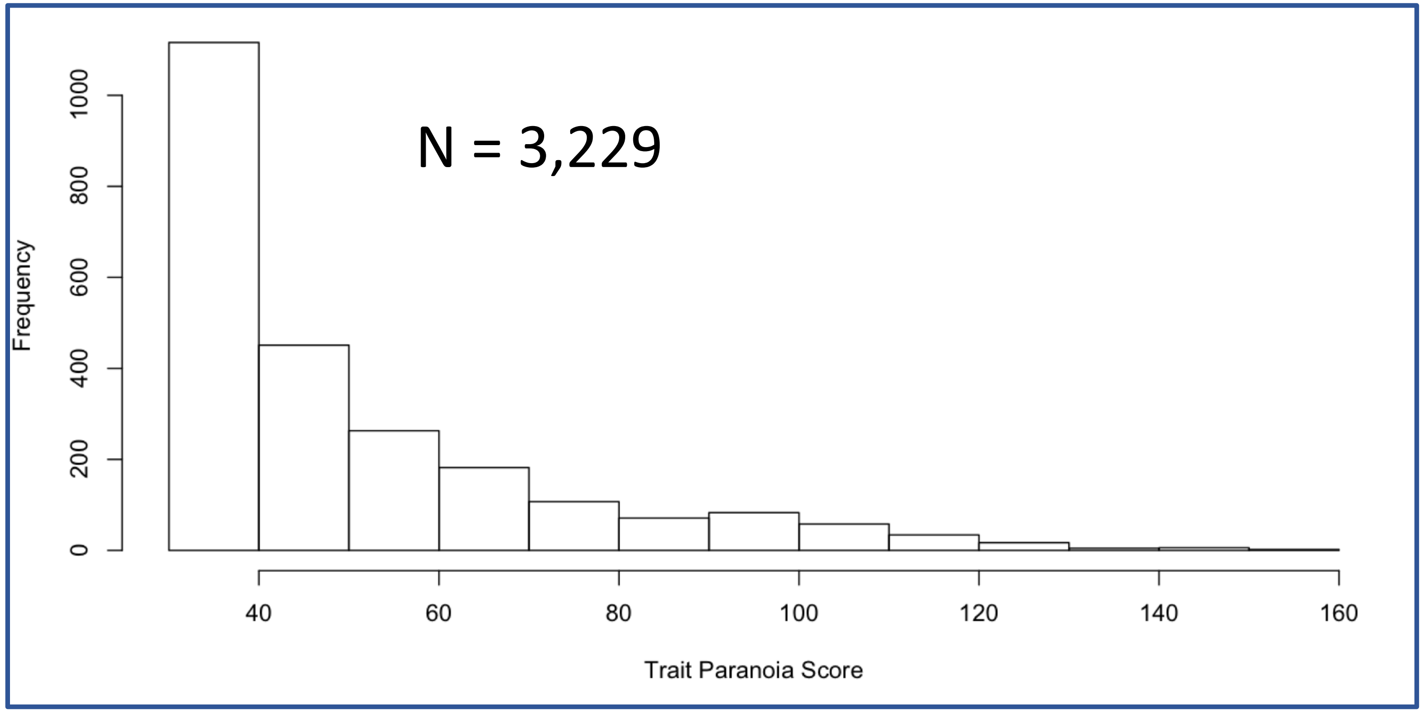
- Dictator makes decision (fair / selfish)

- Participant asked to make two separate ratings (slider scales of 0-100) about the extent to which they believe the dictator’s decision is motivated by:

  - **“Desire to earn more money” (self-interest)**
  - **“Desire to reduce your (the other person’s) bonus” (harmful intent)**

Experiment 1. Results

- Mean paranoia score = 50.7 ± 0.47; range 32-160 (Green et al. mean: 48.8 ± 1.00).



N = 3,229

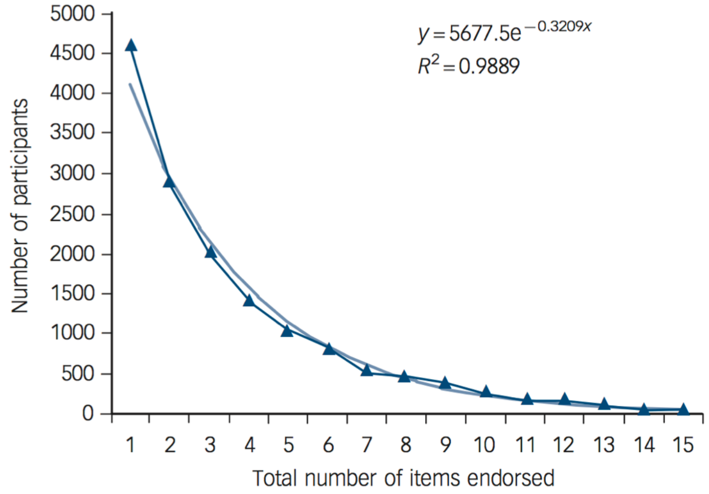
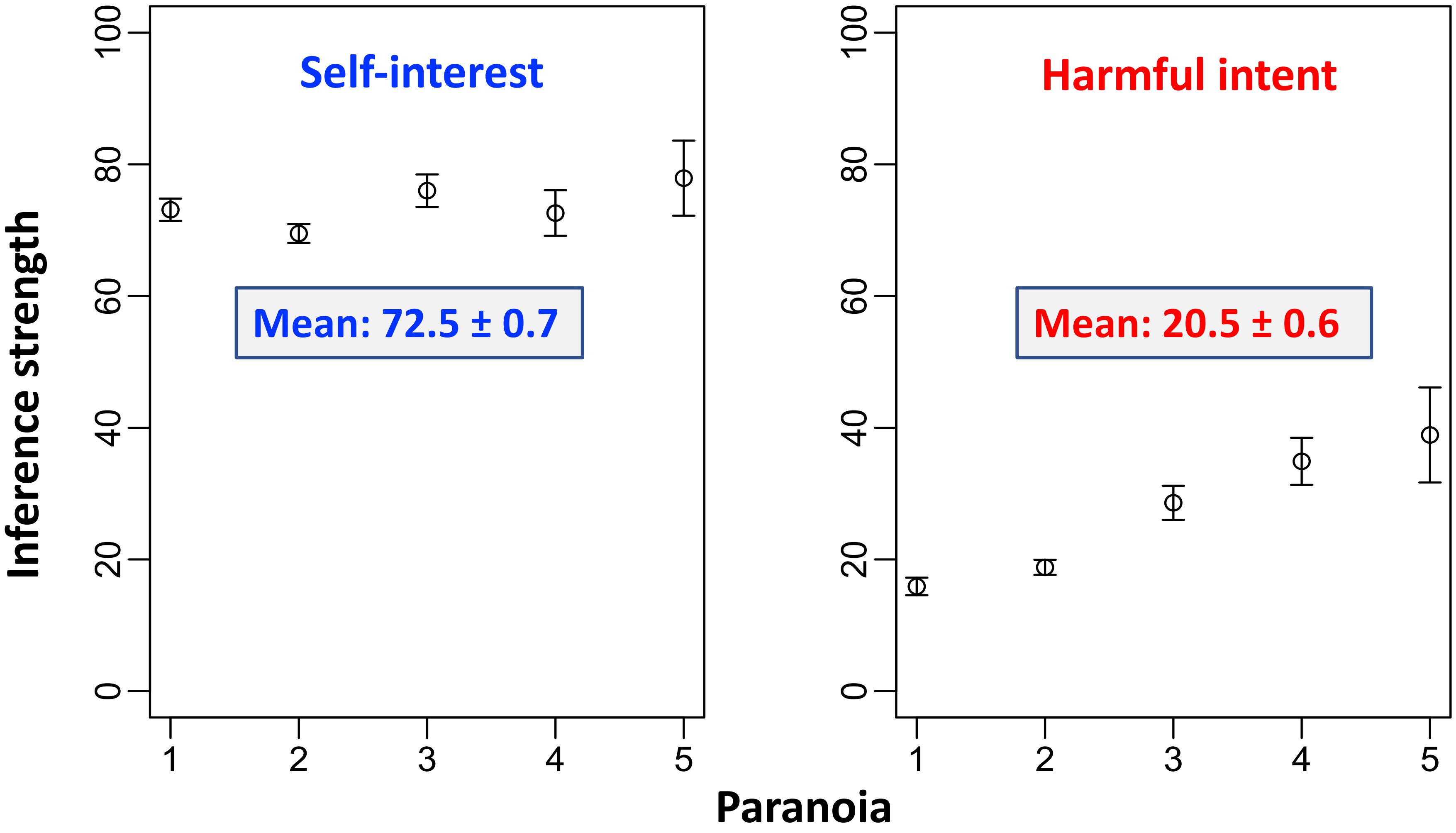
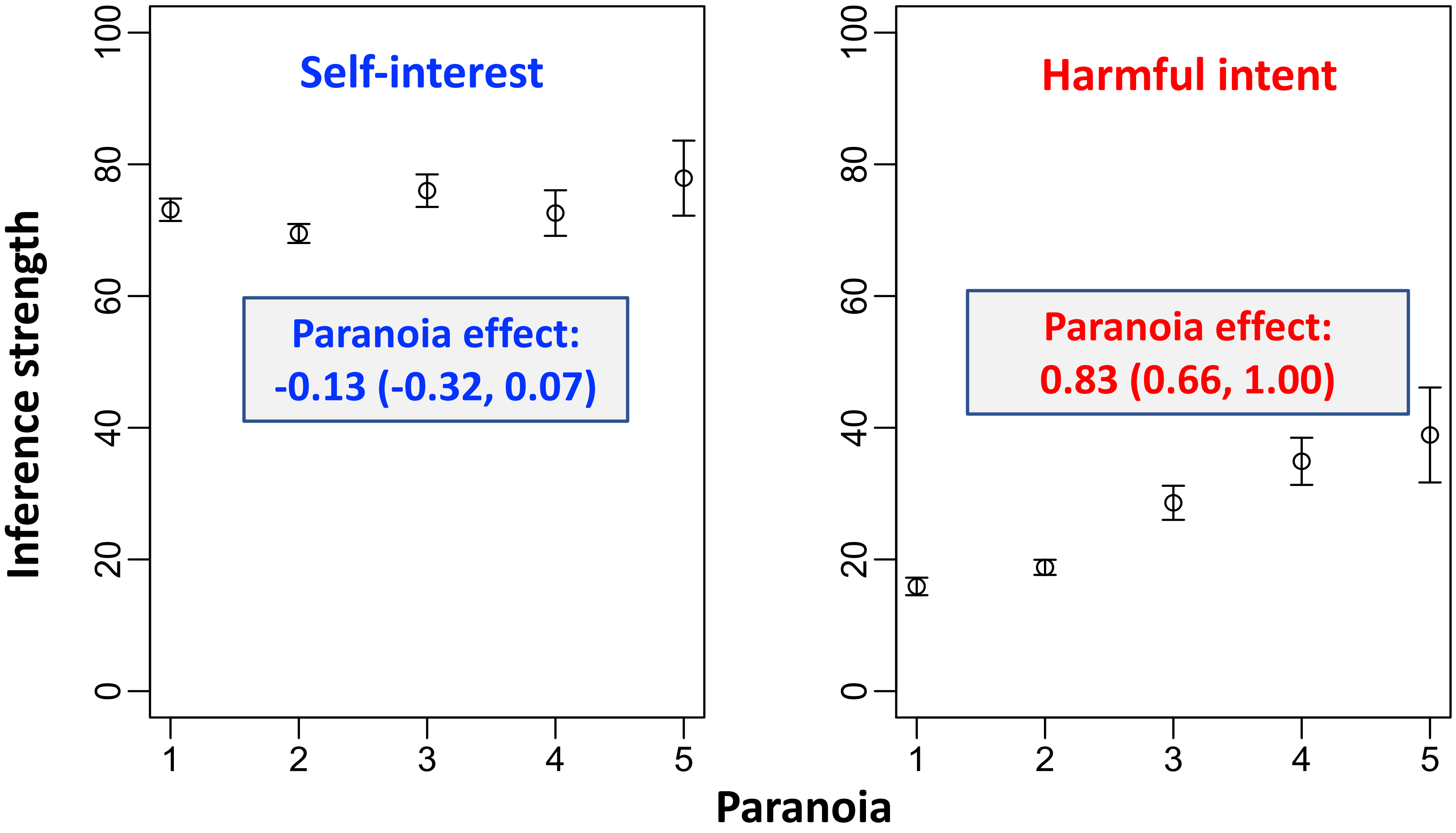


Fig. 2 The distribution of total paranoia scores in the general population.

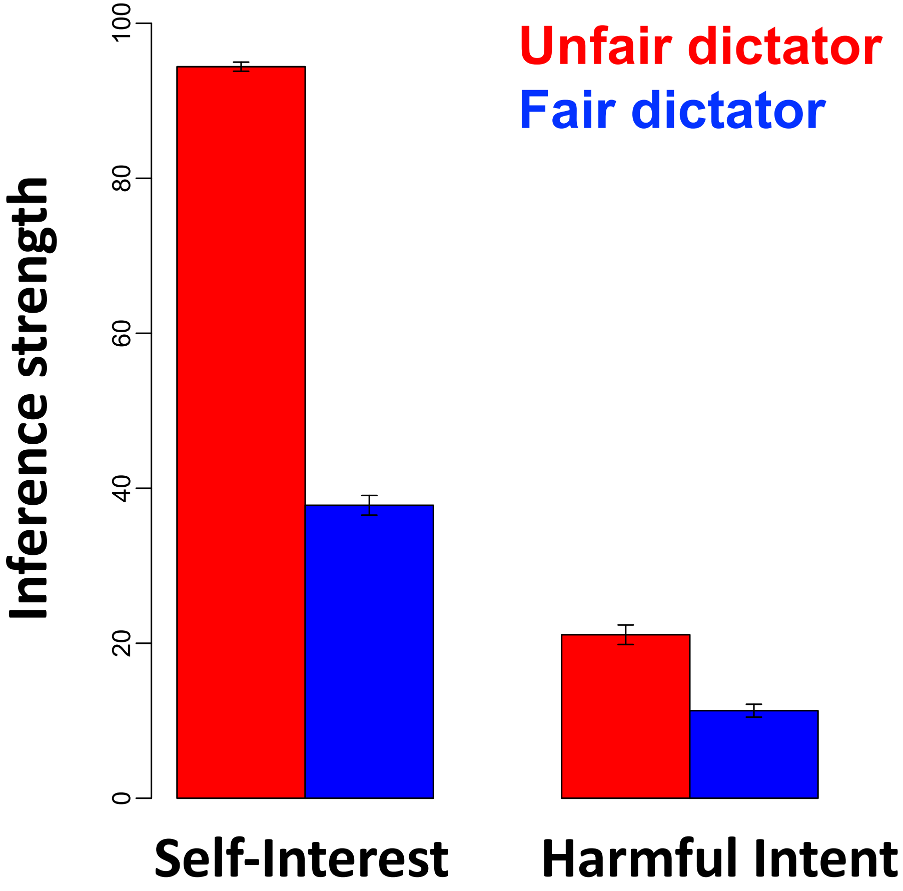
Experiment 1. Results



Experiment 1. Results



Experiment 1. Results



- ↑ self-interest & ↑ harmful intent when responding to **unfair partners**.
- No fairness x paranoia interaction on harmful intent attribution.
- (i.e. paranoia does not result in exaggerated responses to unfairness)

Experiment 1. Summary

- **Pre-existing paranoia results in higher attributions of harmful intent for the same outcomes -> suggests a bias in estimates of others' utility functions.**

- **Paranoia reflects over-perception of hostile intent specifically, rather than more general negative social evaluations of others** (paranoia doesn't affect judgements of self-interest).

- **"Live" paranoid ideation is generally labile** (everyone attributes more harmful intent to unfair dictators; no fairness x paranoia interaction).

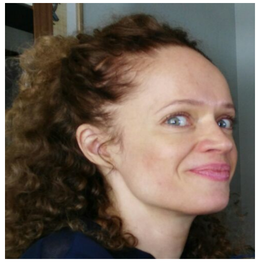
Exp 2: Does experimental social threat bias utility function estimations?

- How does experimental social threat affect hostile intent attributions in ambiguous settings?
- How does pre-existing paranoia interact with social threat to affect hostile intent attributions?

Experiment 2. Method

Anna Greenburgh
PhD candidate

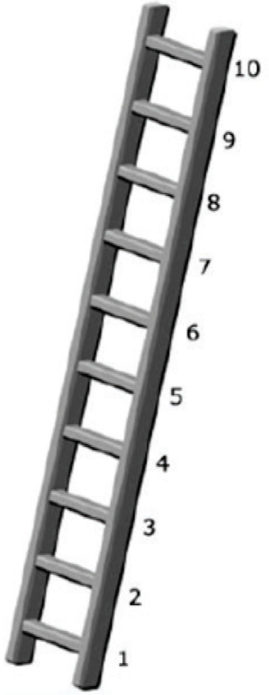
Zeina Ramadan
MSc Student



Vanessa Saalfeld
MSc Student

- N = 2,030 people (58 % female; age: 18-98 years; a new sample to Exp 1).
- *Step 1*: Green et al. Paranoid Thoughts Scale -> trait paranoia.
- Mean paranoia score: 54.8 ± 0.57 (range: 32-160).
- Social threat manipulations: (a) interact with someone higher status; (b) interact with political adversary.

Experiment 2a. Social status manipulation



Think of the ladder below as representing where people stand in your country.

At the **top** of the ladder are the people who are the best off - those who have the most money, the most education and the most respected jobs. At the **bottom** are the people who are the worst off - those who have the least money, the least education and the least respected jobs, or no job. The higher up you are on this ladder, the closer you are to the people at the very top; the lower you are on this ladder, the closer you are to the people at the very bottom.

Where would you place yourself on this ladder?

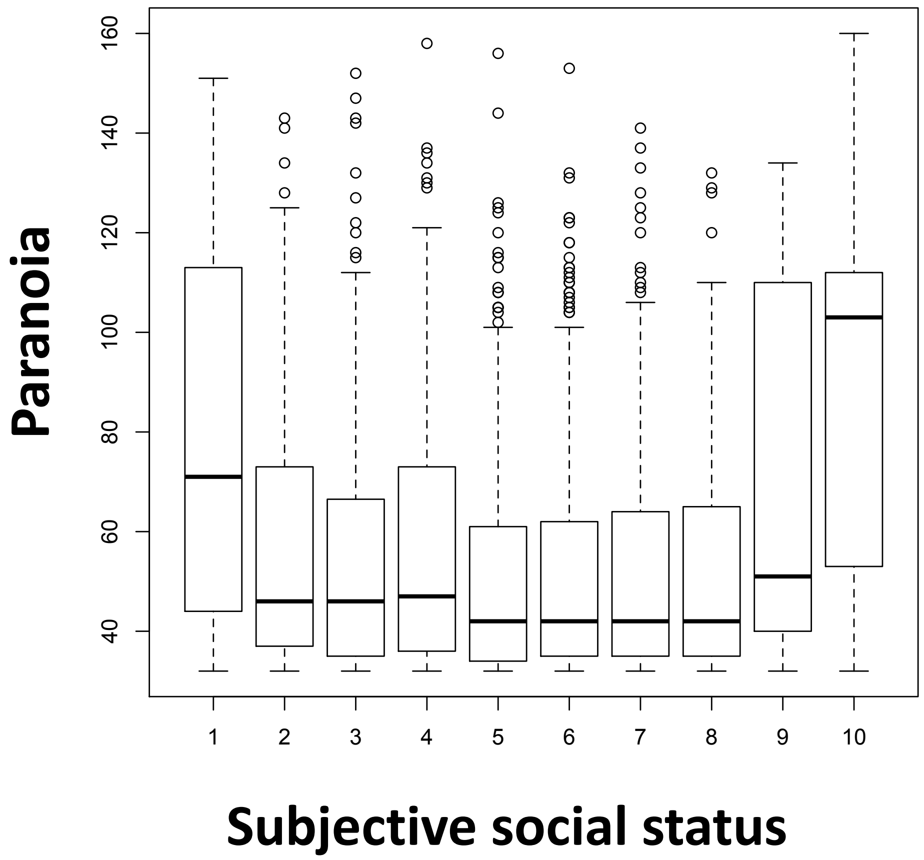
Adler et al. 2000 *Health Psych.*

Experiment 2a. Social status manipulation

10
9
8
7
6
5
4
3
2
1

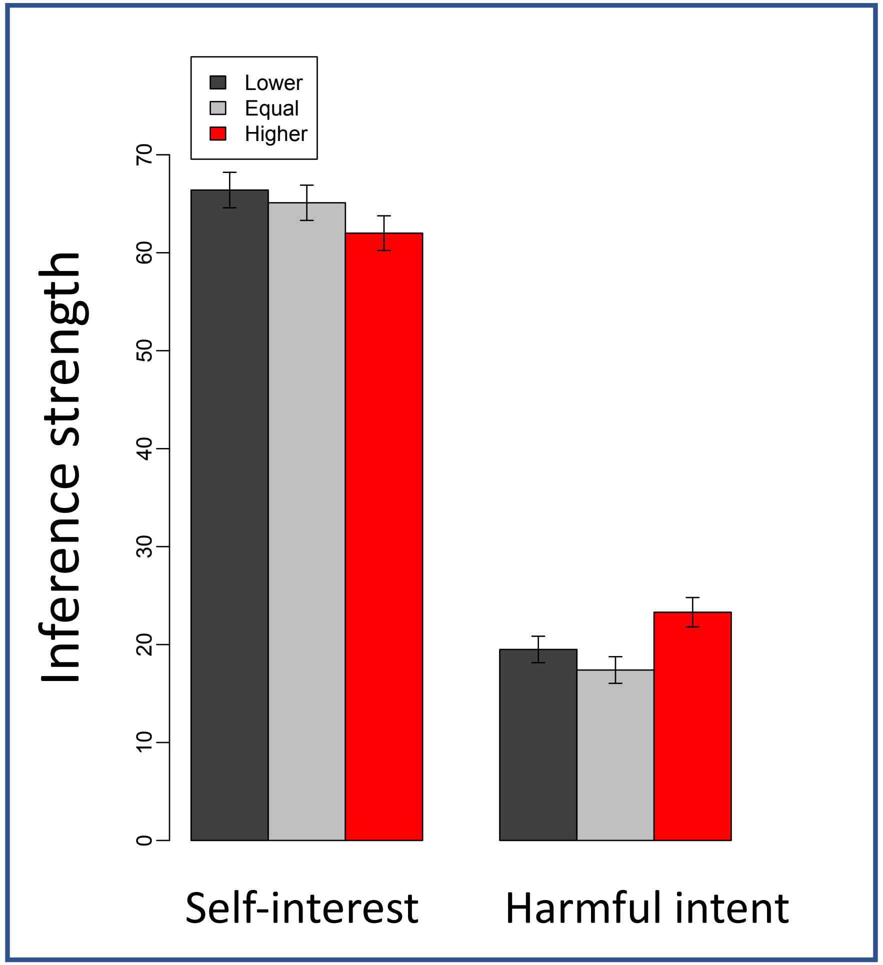
- Participants provide subjective social status.
- Mean SSS score: 5.01 ±0.0 (range: 1-10)
- Participants matched to dictator who is lower / same / higher subjective status.
- Participants infer whether dictators motivated by (i) self-interest and (ii) harmful intent (as before).

Experiment 2a. Results



- Lower status individuals more paranoid (effect: -0.26; CI: -0.45, -0.06).
- But also evidence for a “paranoia of the elite” (effect: 1.31, CI: 0.67, 1.95).
- Needs more investigation….

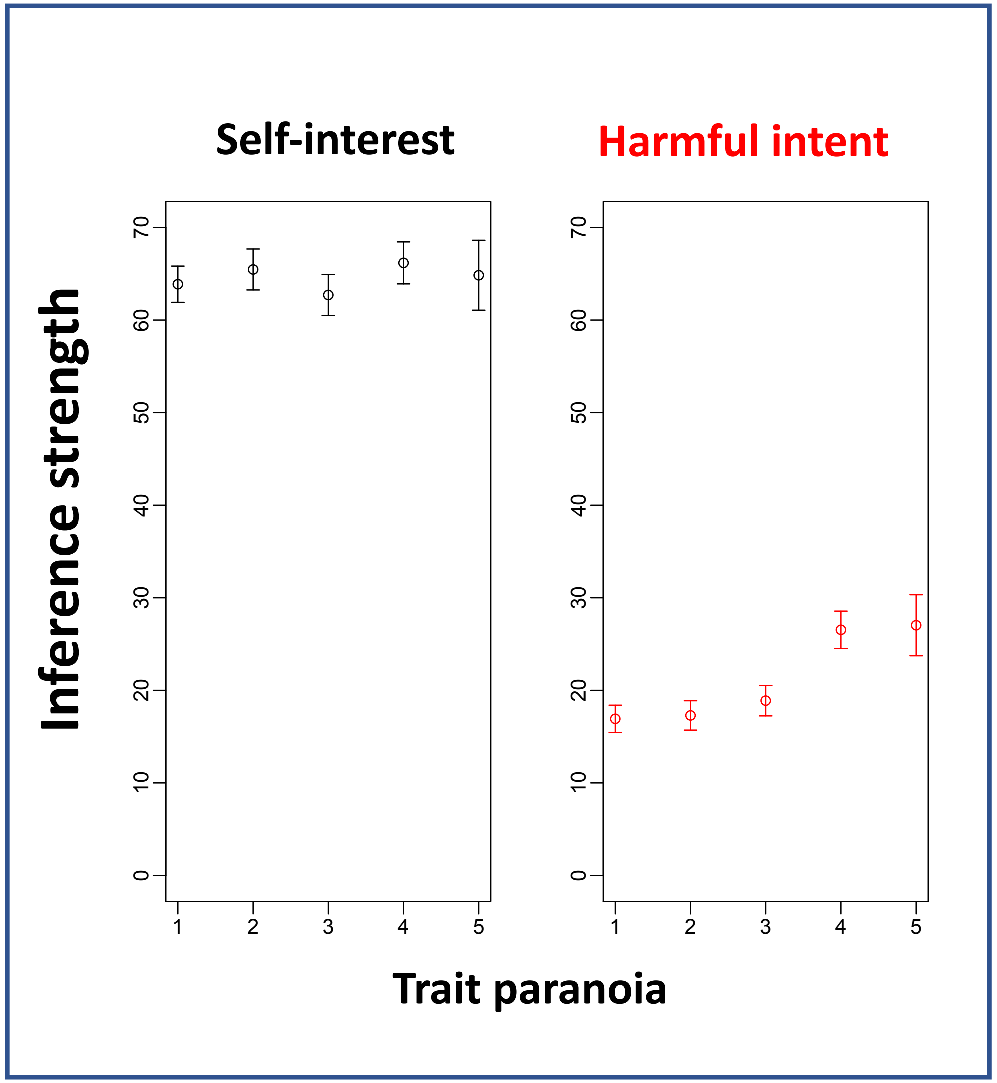
Experiment 2a. Results



- Self-interest attributions > harmful intent (as before).

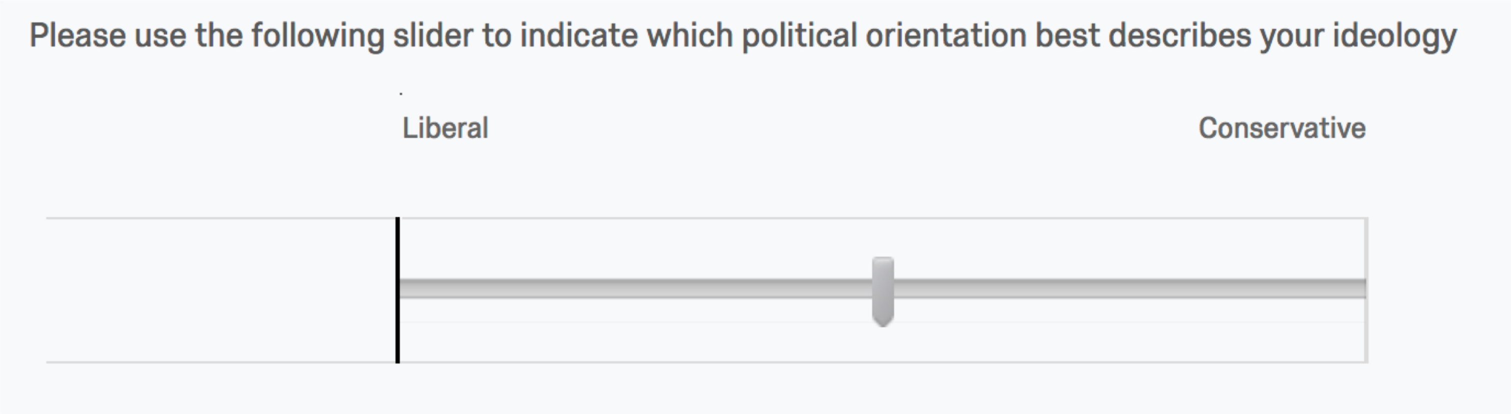
- Stronger harmful intent attributions when **dictators were higher status** than participants.

Experiment 2a. Results



- Trait paranoia predicted harmful intent attribution (effect: 0.43, CI: 0.19, 0.67) but not attributions of self-interest.
- This replicates Exp 1.
- No interaction between paranoia x relative social status on harmful intent attribution.

Experiment 2b. Group membership manipulation

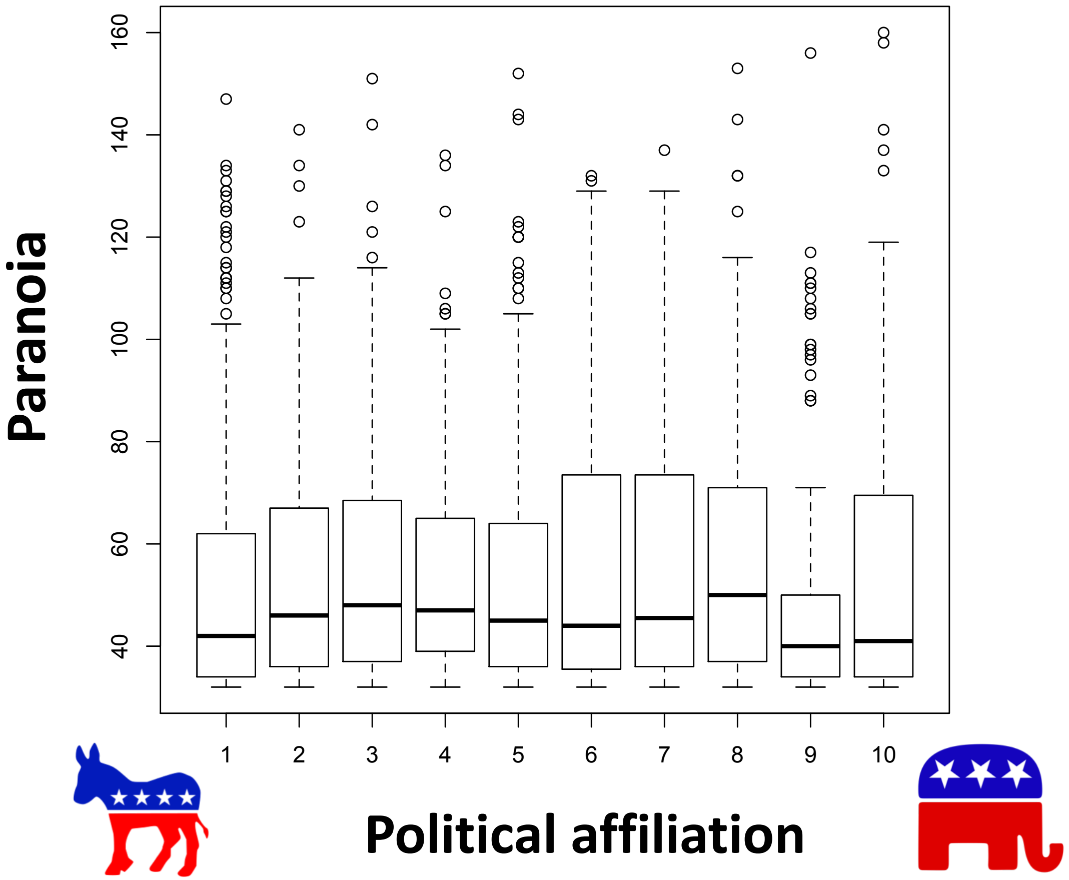


Please use the following slider to indicate which political orientation best describes your ideology

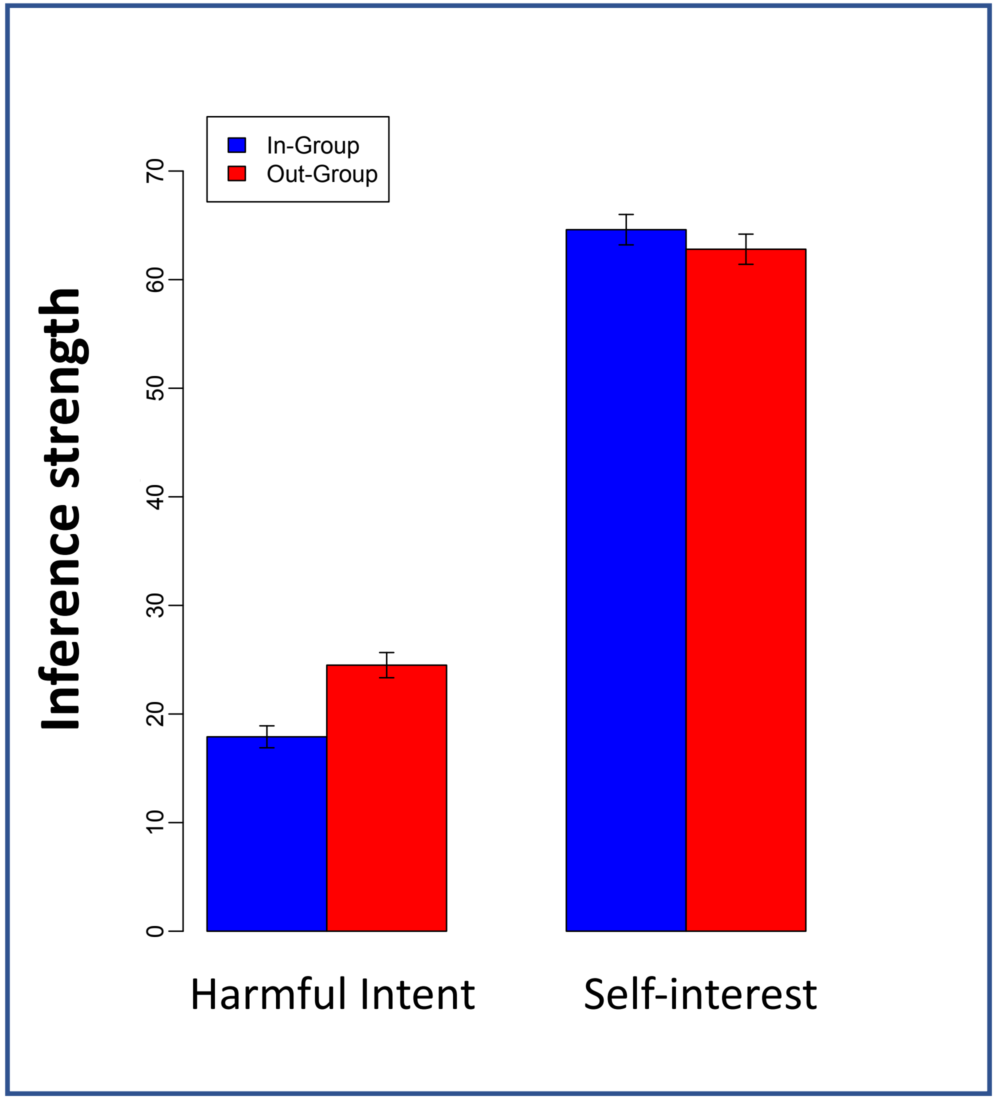
- Participants (US-based) rate their political ideology (0=liberal; 100=conservative).
- Slight liberal bias (mean: 41.8 ± 0.67; range = 0-100).
- Participants matched with dictator of same / different political ideology.
- Rate dictator's (i) self-interest and (ii) harmful intentions (as before).

Experiment 2b. Results

- Political conservatives more paranoid than political liberals (effect: 0.18, CI: 0.03, 0.34).



Experiment 2b. Results



- **Out-group** dictators rated as having higher harmful intent than **in-group** dictators.

- Paranoia positively predicted harmful intent attribution (effect: 0.55, CI: 0.33, 0.78) but not attribution of self-interest (i.e. another replication).

- No paranoia x group membership interaction on harmful intent attribution.

Experiment 2. Summary

- **Increasing paranoia associated with increasing attributions of harmful intent but no effect on attributions of self interest** (2 x replications of Exp 1 result).

- **Paranoid thinking can be dialed up and down in response to social threat** (social threat increases harmful intent attributions in most subjects).

- **Paranoia seems to reflect a lower baseline for detecting social threat, rather than impaired reactivity to it** (no paranoia x social threat interaction on harmful intent attribution in any experiment).

Experiment 3: How do biased estimates of utility functions affect social behaviour?

- Previous work suggests that paranoia is associated with increased aggressive / hostile / violent tendency, but the factors mediating this effect are unclear.

- Is increased aggression in paranoia mediated by paranoid people's perceptions that others intend them harm?

Psychological Medicine, Page 1 of 11. © Cambridge University Press 2017
doi:10.1017/S0033291717003075

ORIGINAL ARTICLE

Conflict and cooperation in paranoia: a large-scale behavioural experiment

N. J. Raihani[1] and V. Bell[2,3]*

[1] *Department of Experimental Psychology, University College London, London, UK*
[2] *Division of Psychiatry, University College London, London, UK*
[3] *South London and Maudsley NHS Foundation Trust, London, UK*

Background. Paranoia involves thoughts and beliefs about the harmful intent of others but the social consequences have been much less studied. We investigated whether paranoia predicts maladaptive social behaviour in terms of cooperative and punitive behaviour using experimental game theory paradigms, and examined whether reduced cooperation is best explained in terms of distrust as previous studies have claimed.

Methods. We recruited a large population sample ($N = 2132$) online. All participants completed the Green *et al.* Paranoid Thoughts Scale and (i) a Dictator Game and (ii) an Ultimatum Game, the former with an option for costly punishment. Following distrust-based accounts, we predicted highly paranoid people would make higher offers when the outcome depended on receiving a positive response from their partner (Ultimatum Game) but no difference when the partner's response was irrelevant (Dictator Game). We also predicted paranoia would increase punitive responses. Predictions were pre-registered in advance of data collection. Data and materials are open access.

Results. Highly paranoid participants actually made lower offers than non-paranoid participants both in the Dictator Game and in the Ultimatum Game. Paranoia positively predicted punitive responses.

Conclusions. These findings suggest that distrust is not the best explanation for reduced cooperation in paranoia and alternative explanations, such as increased self-interest, may apply. However, the tendency to attribute harmful intent to partners was important in motivating punitive responses. These results highlight differing motivations underlying adverse social behaviour in paranoia and suggest that accounts based solely on the presenting features of paranoia may need to be rethought.

Received 22 March 2017; Revised 19 September 2017; Accepted 19 September 2017

Key words: Delusion, game theory, psychosis, schizophrenia, social cognition.

Introduction

Paranoia lies on a continuum, ranging from low-level paranoid thoughts to frank paranoid delusions that centre on concerns about others' harmful intent (Freeman & Garety, 2000). Paranoia has typically been investigated using psychometric measures or tests that measure cognitive biases, reflecting the common focus on individualistic processes involved in evaluating personal threat (Freeman & Garety, 2014). However, relatively little research has explored social interaction despite adverse social behaviour being a common characteristic of paranoia (Freeman, 2016).

Paranoia strongly relates to processes involved in social decision-making and cooperation (Boyer *et al.* 2015; Patrzyk & Takáč, 2017) and there is now a nascent literature investigating paranoia using game theoretic paradigms. These paradigms involve social interactions where cooperative or non-cooperative decisions determine each participant's economic payoff (Camerer, 2003). These tasks are particularly relevant to paranoid concerns because they allow for cooperation and mutual benefit but also allow for harm, in the form of economic costs, and cheating, by gaining resources at the expense of the partner. These approaches therefore have the potential to reveal how paranoia affects social behaviour, and – because they are designed to selectively isolate specific social motivations – can help test and develop psychological theories.

Paranoia and cooperative behaviour

Previous game theory studies report that paranoia predicts reduced cooperation in the Prisoner's Dilemma Game (Ellett *et al.* 2013) and reduced investments in the Trust Game (Fett *et al.* 2012, 2016;

* Address for correspondence: V. Bell, Division of Psychiatry, UCL, 6th Floor, Maple House, 149 Tottenham Court Road, London W1T 7NF, UK.
(Email: Vaughan.Bell@ucl.ac.uk)

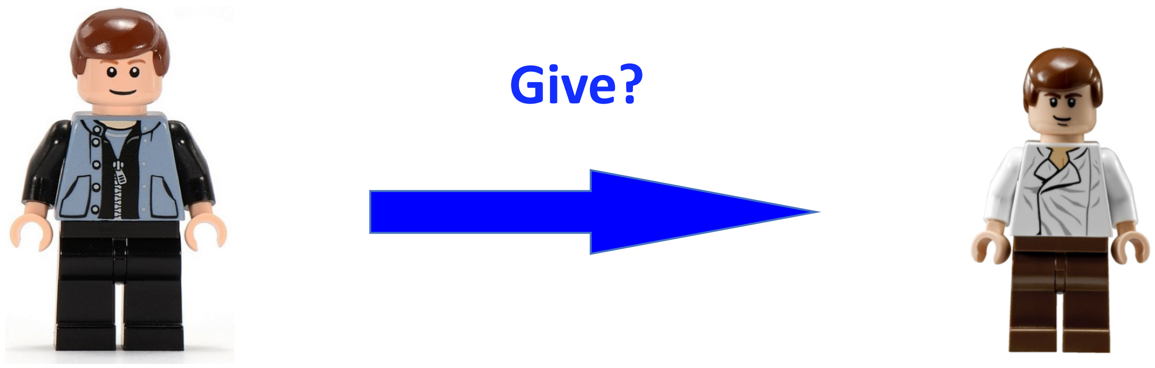
Raihani & Bell 2017 *Psychological Medicine*

Experiment 3: Method

The Dictator Game

Dictator

Receiver



Give?

$X

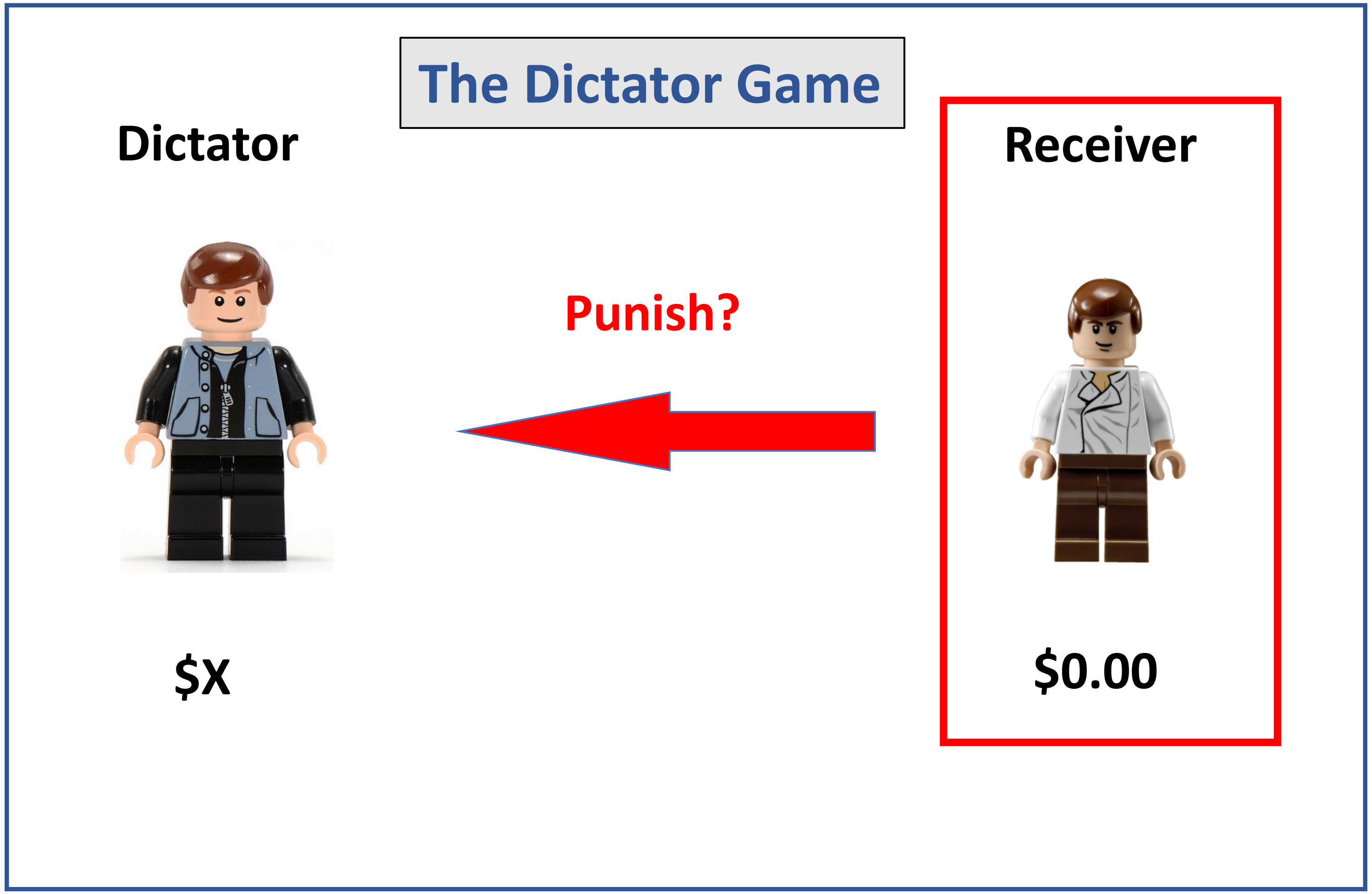
$0.00

Dictator motives are **ambiguous**

Selfishness can reflect (i) greed or (ii) harmful intent

- Paranoia predicts **increased attribution of harmful intent** (but not self-interest) in the DG (Raihani & Bell 2017).

Experiment 3. Method



The Dictator Game

Dictator

Receiver

Punish?

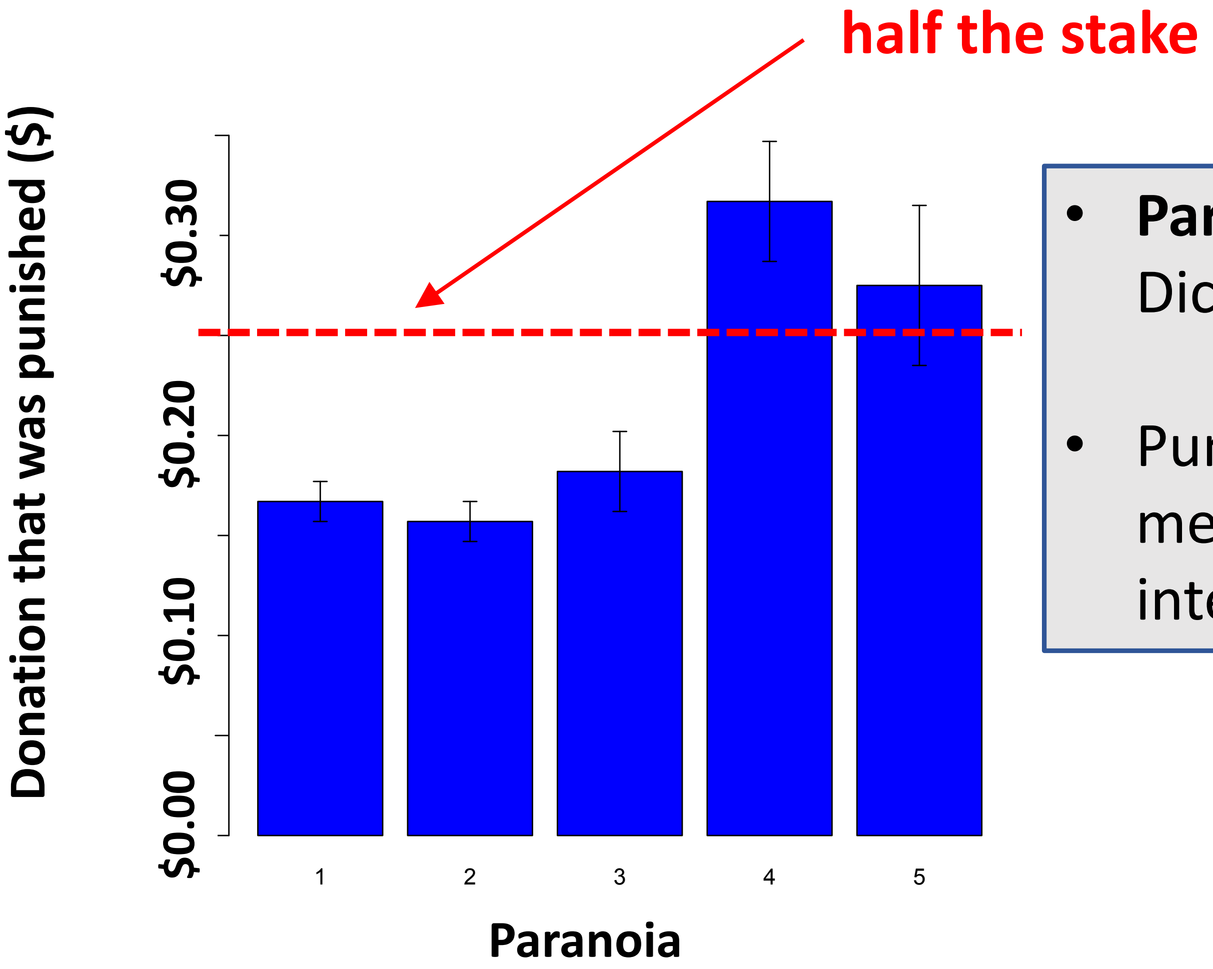
$X

$0.00

- Paranoia predicts **increased attribution of harmful intent** (but not self-interest) in the DG (Raihani & Bell 2017).
- Punishment depends on beliefs about the target's **intentions** as well as outcomes.

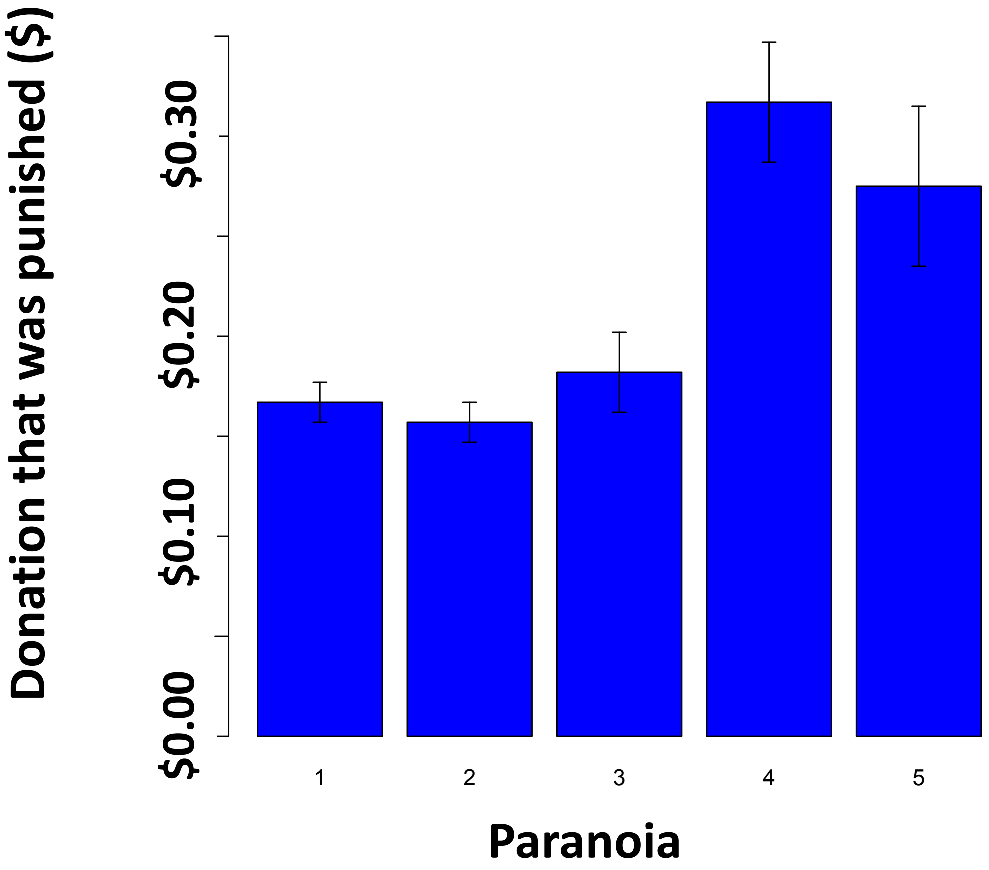
- **Does paranoia increase punitive tendency?**
- **Is this mediated by harmful intent attribution?**

Experiment 3. Results



- **Paranoia positively predicted punishment** in the Dictator Game (effect: 0.86, CI: 0.57, 1.16)
- Punishment decisions partially but not fully mediated by tendency to attribute harmful intentions to dictators.

Experiment 3. Results



- **We just replicated this result in a new sample (n > 1100).**

- We also found that paranoia is positively correlated with 'negative social potency' which measures how much people enjoy being 'cruel, callous and using others for personal gain.'

- Tendency to punish is mediated by negative social potency→ people who enjoy harming others also punish more.

Overall summary

- Humans are social and have an evolutionary history of dealing with threats from conspecifics.
- Selection can favour mechanisms that attempt to anticipate / deflect these threats by inferring unseen intentions/preferences from observed behavior.
- Uncertainty in perception means intention attribution is error-prone. Might be biased towards over-estimating harmful intent – this might be the basis of paranoia.
- Paranoia responds to experimentally-induced social threat and affects social behavior.
- Can this evolutionary perspective help us to understand where risk factors for clinical paranoia and how to treat it?

Bibliography for papers, data & code.

- **Exp 1:** Raihani & Bell (2017) Paranoia and the social representation of others: a large-scale game theory approach. *Scientific Reports* 7, 4544.

- **Exp 2a & b:** Saalfeld et al. (2018) Experimentally-induced social threat increases paranoid thinking. PsyArXiv: https://psyarxiv.com/jxkv3/

- **Exp 2c:** Greenburgh et al. (2018) Paranoia and conspiracy: group cohesion increases harmful intent attribution in the Trust Game. PsyArXiv: https://psyarxiv.com/mgzjr/

- **Exp 3:** Raihani & Bell (2017) conflict and cooperation in paranoia: a large-scale behavioural experiment. *Psychological Medicine,* 76, 1-11.

Questions / comments / suggestions?

UCL

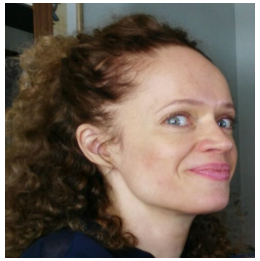
THE ROYAL SOCIETY

Email: nicholaraihani@gmail.com
Twitter: @nicholaraihani

Dr Vaughan Bell
UCL Psychiatry
South London
Maudsley NHS
Trust

Anna Greenburgh
PhD candidate

Zeina Ramadan
MSc Student



Vanessa Saalfeld
MSc Student