from trust in groups to trust in individuals

@philipparnamets

Philip Pärnamets, Tobias Granwald & Andreas Olsson

philip.parnamets@nyu.edu

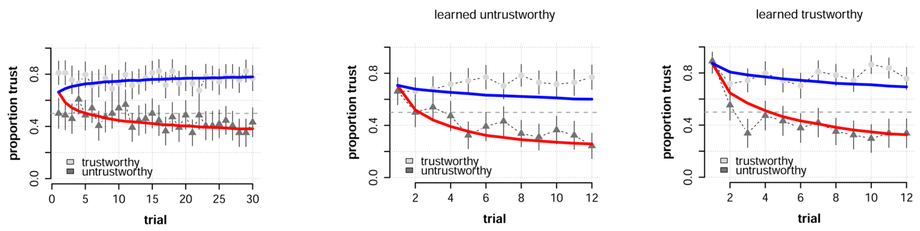
background & aim

- trust is central for social interaction.
- recent work investigated how we learn to trust other individuals by tracking their average trustworthiness using basic learning mechanisms
- however, our social world is structured by group information and little is known about how such information affects, if at all, decisions to trust

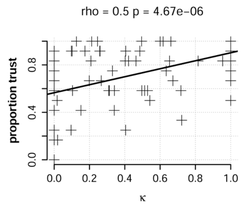
we investigated how trust in individuals is learned and unlearned in the context of minimal social groups

we aimed to characterize the dynamics both of trial-by-trial updating of trust (learning) and of within-trial variation in trust decisions

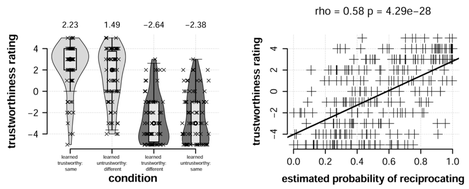
results – learning to trust



- participants **rapidly adjust to actual behavior** of partner
- only **small influence** of prior group learning on aggregate behavior
- **no difference** in based on group membership on final trial
- at start of character phase participants **discount prior learning**, (captured by κ)
- **higher learning rate** in character phase to **negative prediction errors** compared to positive ones



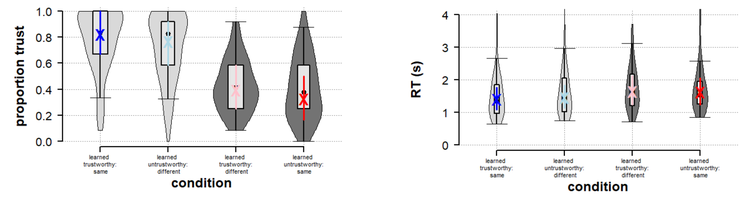
- **κ correlates** with aggregate trust to **actually trustworthy** individual fom untrustworthy group



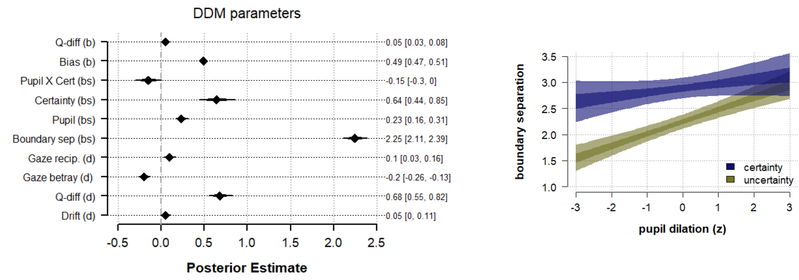
- **trustworthiness ratings** of each partner **correlate with trustworthiness estimates** from RL model

results – deciding to trust

posterior predictive checks



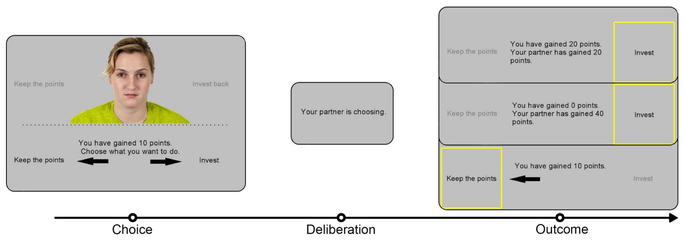
model results



- participants' choices & RTs can be explained by **drift-diffusion model** using quantities **derived from RL model**
- trustworthiness estimates affect both **drift rates and starting points** of sampling process
- **looking behavior** during deliberation has additive effect on **drift rates**
- **certainty** in partners' trustworthiness **and pupil-linked arousal** regulates **boundaries** of decision process

procedure

- **trust games** against computerized opponents
- groups identified by **t-shirt color**
- one group reciprocated with **P=.75** and other group with **P=.25**
- task **incentivized**
- two phases of experiment (see below)



learning phase

(learn about groups)
- multiple trust games
- trustworthiness rating of groups
- selection of group for another round

… x60 trials

character phase

(learn about individuals)
- repeated trust games
- trustworthiness rating of individual partners

… x12 trials each

RL model

hybrid PH-RW model

$$\text{decision} \begin{cases} EV(trust) = Pr_t * 2 + \theta \\ p(trust) = \frac{e^{EV(trust)/\beta}}{e^{EV(trust)/\beta} + e^{1/\beta}} \end{cases}$$

$$\text{learning} \begin{cases} d_t = R_t - Pr_t \\ Pr_{t+1} = Pr_t + \alpha A_t d_t \begin{cases} \alpha_+, & d_t > 0 \\ \alpha_-, & d_t < 0 \end{cases} \\ A_{t+1} = \lambda * |d_t| + (1 - \lambda) A_t \end{cases}$$

learning phase

$Pr_{blue} = Pr_{yellow} = 0.5$

$A_{blue} = A_{yellow} = 1$

character phase

κ : discount parameter

$Pr_{blue\_1} = Pr_{blue\_2} = Pr_{blue} + (1\text{-}Pr_{blue})^*\kappa$

$Pr_{yellow\_1} = Pr_{yellow\_2} = Pr_{yellow} + (1\text{-}Pr_{yellow})^*\kappa$

$A_{blue\_1} = A_{blue\_2} = A_{yellow\_1} = A_{yellow\_2} = 1$

conclusions

- group information (stereotypes) **can be quickly overcome given individuating information in available**, with minor long-term effects on aggregate behavior
- participants compute **expectations of reciprocation** (as predicted by RL model) and use these to guide both **behavior and judgments**
- combination of **discounting of prior learning** together with **faster learning rates for negative prediction errors** allow participants to **avoid myopia and effectively manage learning**
- participants' trust decisions depend on **trial-by-trial variations in visual attention and pupil-linked arousal**

NEW YORK UNIVERSITY

Karolinska Institutet