Individual differences in dynamic belief updating during trust learning

V. Guigon[1,2], S. Topel[3] and Caroline J. Charpentier[1,2,4]

1. Department of Psychology, University of Maryland, College Park, MD, USA
2. Program in Neuroscience And Cognitive Science, University of Maryland, College Park, MD, USA
3. Faculty of Social and Behavioral Sciences, Leiden, Netherlands
4. Brain and Behavior Institute, University of Maryland, College Park, MD, USA

SCAN ME

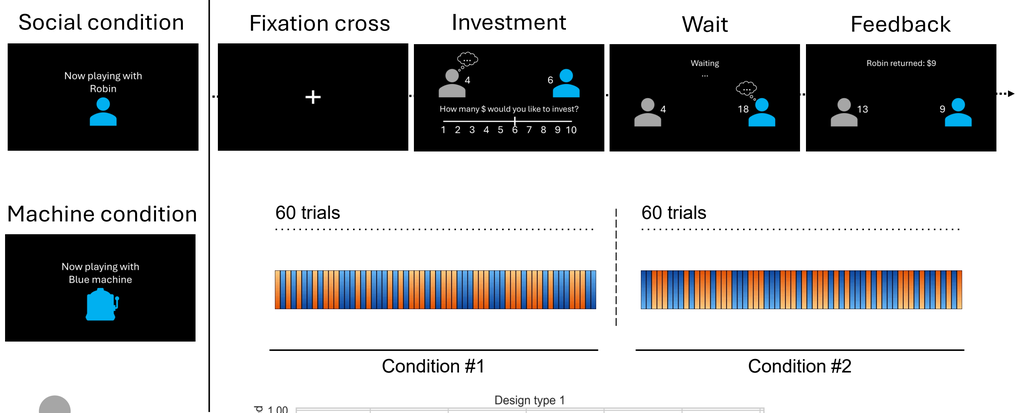
CONTEXT

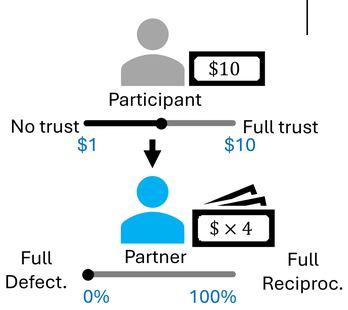
- **Trust** is fundamental to social interaction. It is defined as the willingness to be **vulnerable** to **another being** on the basis of **positive expectations** of their intentions and behaviors (1).
- In repeated interactions, **learning to trust others** involves cognitive processes that integrate uncertainty, context, and potential betrayal.
- **Different strategies** can guide trust behavior. Some rely on heuristics (e.g., fixed rules or triggers). Others use reinforcement learning, adjusting expected value through associative updates, or Bayesian belief updating, which integrates uncertainty into probabilistic inferences about partners' intentions (2, 3).
- **Individuals differ** in how they deploy cognitive **processes**, which may impact their **strategies**. Yet, this heterogeneity has not been systematically characterized. Here, we aim to do so:
  - How does dynamic trust learning integrate sensitivity to betrayal, uncertainty and context?
  - How do these processes vary across individuals and reveal distinct cognitive profiles?

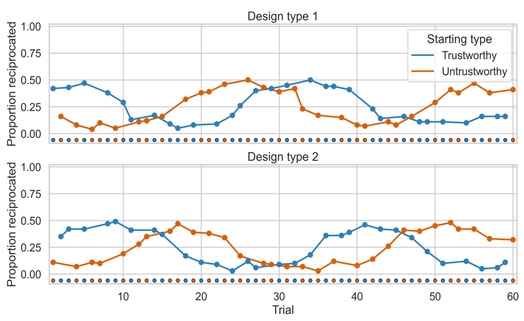
HYPOTHESES

- **Gain/loss learning asymmetry**
  Higher learning rate after losses than gains.
- **Context modulation**
  Betrayal aversion leads to stronger learning rates after losses in the social condition.
- **Individual differences**
  - A subset of participants deploy heuristics (non-learners).
  - Among learners, variability in learning asymmetry and betrayal aversion is captured by model fit/parameters

METHODS

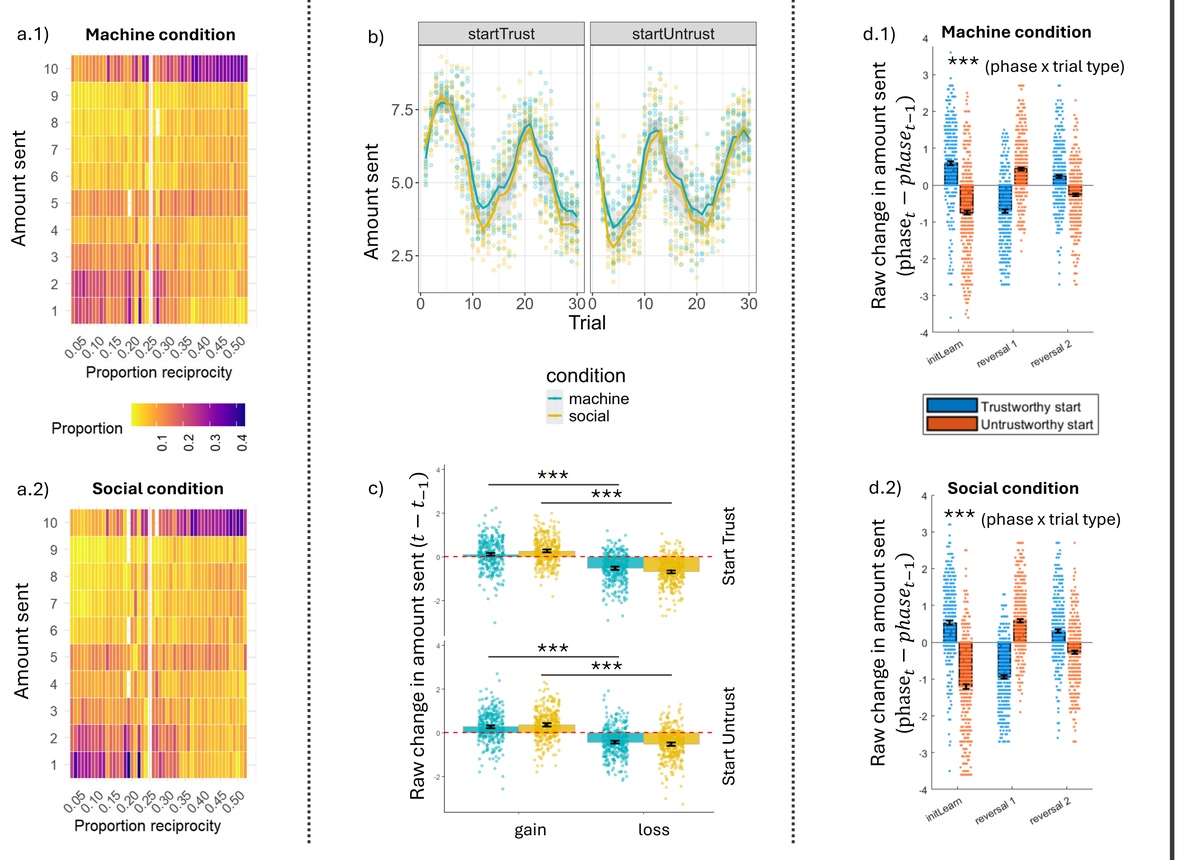






- *Prolific*
- *N = 334 (F=139, M=115)*
- *Age = 35*±11 [19-65]
- Indefinitely repeated multinomial Trust game
- 2 pre-prog. partners
- 2 pre-prog. machines

BEHAVIOR



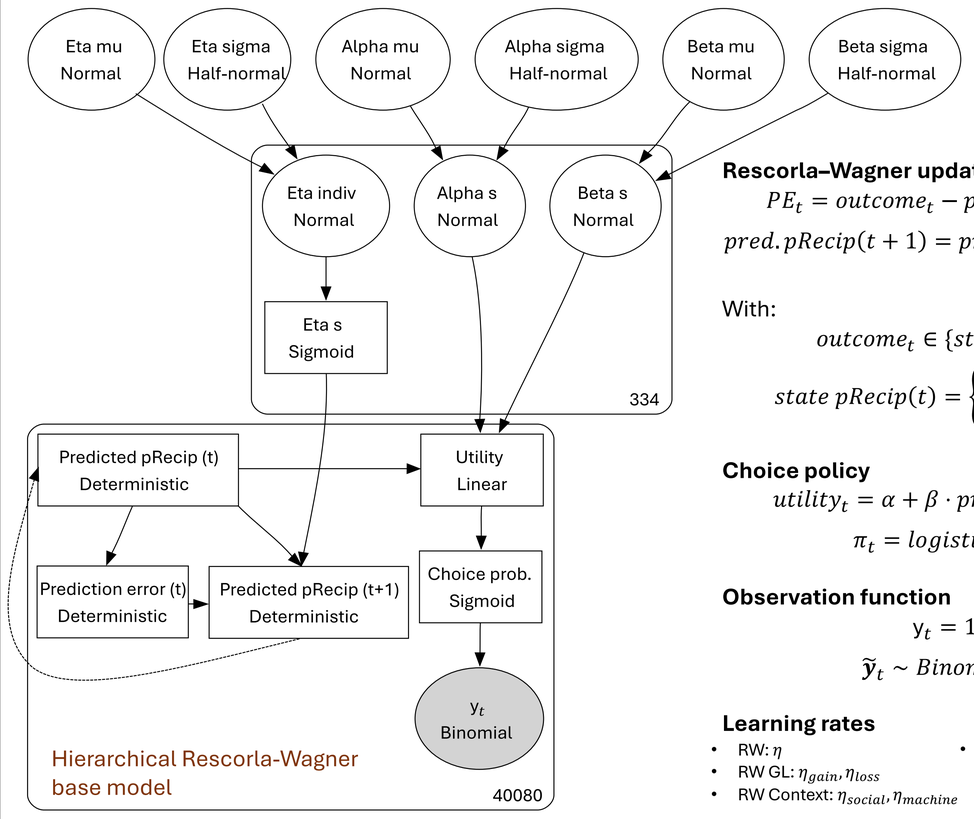
a) In machine (a.1) and social (a.2) conditions, the proportion of amounts sent clustered around ~1 token when reciprocity was below 25% and ~10 tokens when reciprocity exceeded 25%, consistent with utility maximization. A stripe around ~5 tokens, regardless of reciprocity, indicates non-maximizing behavior.

b) Participants adjusted their behavior to track partners' reciprocity over time, reflecting learning.

c) Amount sent changed more after losses than after gains, particularly in the social condition, consistent with a gain/loss asymmetry and betrayal aversion.

d) Following reversals of partner trustworthiness in machine (d.1) and social (d.2) conditions, participants adapted their amounts sent, with loss sensitivity strongest initially and declining across reversals.

GAIN-LOSS AND CONTEXT DEPENDENCY



Rescorla–Wagner update rule

$$PE_t = outcome_t - predicted\ pRecip(t)$$

$$pred.pRecip(t+1) = pred.pRecip(t) + \eta \cdot PE_t$$

With:

$$outcome_t \in \{state\ pRecip(t)\}$$

$$state\ pRecip(t) = \begin{cases} 1 & if\ gains\,(t) > 0 \\ 0 & otherwise \end{cases}$$

Choice policy

$$utility_t = \alpha + \beta \cdot predicted\ pRecip(t)$$

$$\pi_t = logistic(utility_t^*)$$

Observation function

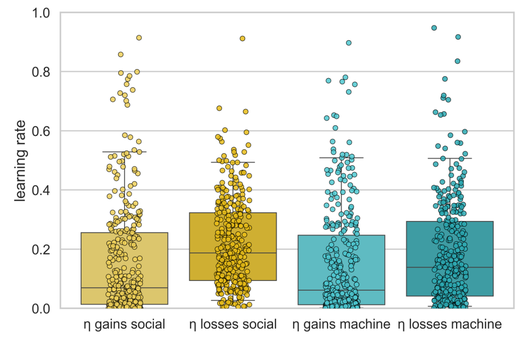
$$y_t = 1 + \tilde{y}_t$$

$$\tilde{y}_t \sim Binomial(9, \pi_t)$$

Learning rates

- RW: η
- RW GL: η_{gain}, η_{loss}
- RW Context: $\eta_{social}, \eta_{machine}$
- RW GL Context: $\eta_{social\ gain}, \eta_{social\ loss}, \eta_{machine\ gain}, \eta_{machine\ loss}$

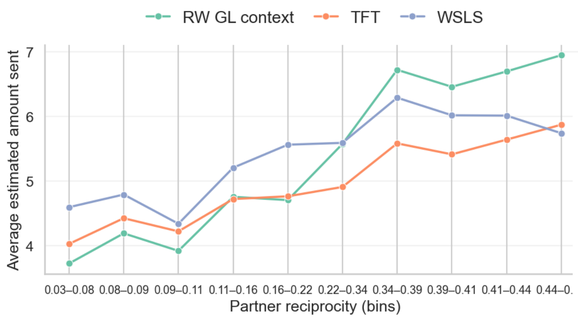
model	elpd_waic	p_waic	elpd_waic_diff	weight	elpd_loo	p_loo	elpd_loo_diff	weight
RW_GL_Context	-91821	6864	0	0.63	-91102	6145	0	0.63
RW_Context	-96854	3840	5033	0.16	-96842	3828	5740	0.16
RW_GL	-97169	3675	5349	0.18	-97179	3685	6078	0.18
RW	-100343	2852	8522	0.03	-100352	2861	9250	0.04



INDIVIDUAL DIFFERENCES

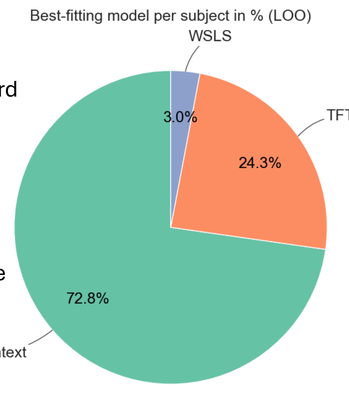
Heuristics:

- Tit-for-tat : Each subject has a personal default level and adjusts up or down depending on the partner's last reciprocity.
- WSLS : After a trustworthy outcome (*gain*), the subject goes to their personal high-send level; after a betrayal (*loss*), they drop to their personal low-send level.

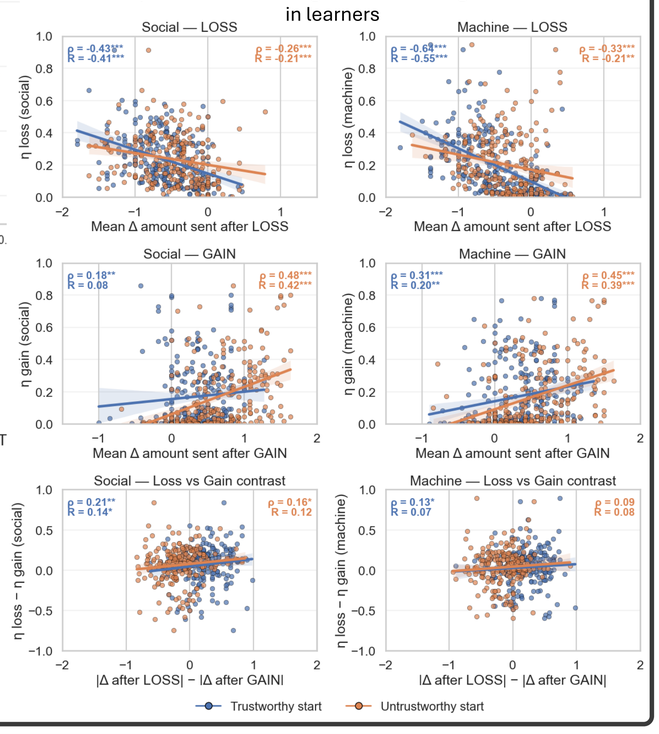


Average profiles differ by strategy:

- RW GL context adapts to partners and is pulled toward utility-maximizing regions.
- TFT and WSLS are concentrated around 5–6, consistent with the heatmap stripe.



Relationships between learning rates and changes in amount sent in learners



CONCLUSIONS & NEXT STEPS

- Trust behavior shows gain/loss asymmetry, context modulation, and individual variability in strategy use.
- Best-fitting model was **RW GL context**, capturing 72.8% of participants. The rest was best described by heuristics.
- Among learners, the model captures variability in betrayal aversion & context sensitivity: Loss (gain) learning rates correlated with decreases (increases) in amount sent after losses (gains) in both social and machine conditions.
- Next, we will extend the model space to include Bayesian updating and sensitivity to uncertainty (4).
- We will examine links between model parameters and individual-difference measures.
- All findings will be validated in a replication dataset.

REFERENCES

- (1) Alós-Ferrer, C., & Farolfi, F. (2019). *Frontiers in neuroscience*, *13*, 887.
- (2) Bellucci, G., & Dreher, J. C. (2021). *The Neurobiology of Trust, Cambridge University Press, Cambridge*, 185-218.
- (3) Engle-Warnick, J., & Slonim, R. L. (2006). *Economic theory*, *28*(3), 603-632.
- (4) Lamba, A., Frank, M. J., & FeldmanHall, O. (2020). *Psychological science*, *31*(5), 592-603.