WRIGHT STATE UNIVERSITY

ACT-R PGSS 2016

Learning to trust and trusting to learn

Ion Juvina, Othalia Larue, Michael G. Collins, & Peter Crowe

Department of Psychology
Wright State University
Dayton, OH

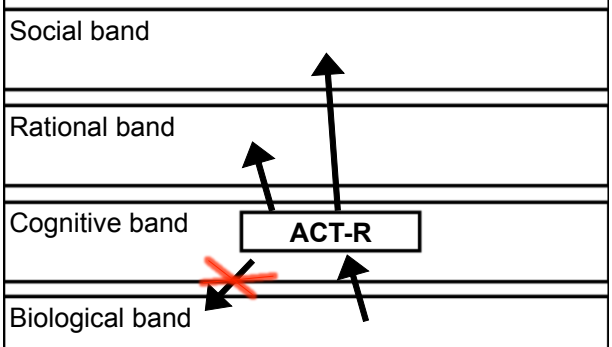
WRIGHT STATE UNIVERSITY

Outline

- Broader architectural issues
- A published model of learned trust
  - Validation studies
  - Critique: strengths and limitations
- A revised model
  - Validation
    - External validity
    - Predictions
- Conclusion and future work

WRIGHT STATE UNIVERSITY

Broader architectural issues



WRIGHT STATE UNIVERSITY

Broader architectural issues

- ACT-R is best positioned to inform the rational and social bands
  - Develop process models of decision making and strategic interaction
    - Lasting, repeated, cooperative or adversarial
    - Involving rational agents balancing multiple motives, constraints etc.
  - Facilitate incremental theory building and integration
  - Use cognitive models as intelligent agents/robots
    - e.g., companion robots

Companion robot

5

Broader architectural issues

- Representing agents vs. representing inanimate entities
  - Theory-of-mind
  - Impression formation
  - Trust
  - Human-machine interaction
    - Anthropomorphic features

6

Objective & approach

- Develop a unified theory of learned trust
  - grounded in general principles of human cognition
- *A priori* predictions
  - Predictions generated by a computational model before a human study is run (ex ante)
  - Model developed based on theory, literature, or prior studies
  - Generate predictions for
    - New task, new conditions / manipulations
  - Assess validity of model under new conditions
    - Study design and setup are identical for model simulations and human data collection
    - Simulation data are as rich and fine-grained as the human data
  - Revise model
  - **Model -> predictions -> experiments -> model …**

7

The published model

Juvina, I., Saleem, M., Martin, J.M., Gonzalez, C., & Lebiere, C. (2013). Reciprocal trust mediates deep transfer of learning between games of strategic interaction. *Organizational Behavior and Human Decision Processes*. *120*(2): 206-215.

Juvina, I., Lebiere, C., & Gonzalez, C. (2015). Modeling trust dynamics in strategic interaction. *Journal of applied research in memory and cognition. 4*(3): 197-211.

8

PD and CG game matrices

PD	A	B
A	-1, -1	10, -10
B	-10, 10	1, 1

CG	A	B
A	-10, -10	10, -1
B	-1, 10	1, 1

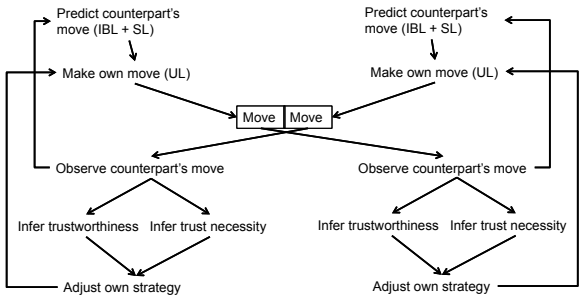
9

Cognitive model

- Awareness of interdependence (model vs. model)
  - Opponent modeling
- Generality
  - Learning
- Transfer of learning
  - Surface transfer
  - Deep transfer

10

Model's control flow



11

Trust mechanism

- **Determines which reward function to use**

Trust	Trust-invest	Reward function
+	+	Payoff1 + Payoff2 – Prev-Payoff2
+	–	Payoff1 + Payoff2 – Prev-Payoff2
–	+	Payoff2
–	–	Payoff1 – Payoff2

12

Model comparison

Model	Training sample		Testing sample	
	Correlation	RMSD	Correlation	RMSD
Trust	0.87	0.09	0.81	0.11
$P1_n$	0.66	0.19	0.65	0.19
$P1_n - P2_n$	0.01	0.46	− 0.001	0.46
$P1_n + P2_n$	0.65	0.29	0.61	0.30
$P1_n + P2_n - P2_{n-1}$	0.66	0.25	0.59	0.27
$P2_n$	0.62	0.34	0.62	0.34

13

Model validation

- Validation study 1
  - Same games (PD & CG)
  - A wider range of conditions
- A priori model predictions (published)
- Large set of human data (320 participants)

Collins, M.G., Juvina, I., & Gluck, K.A. (2016). Cognitive Model of Trust Dynamics Predicts Human Behavior within and between Two Games of Strategic Interaction with Computerized Confederate Agents. *Frontiers in Psychology, Section Cognitive Science 7*:49.

14

Terminology

- Trait trust = trust propensity, dispositional trust
- State trust = specific to a (repeated) interaction
- Trust necessity = need to invest in trust development

15

Setup: what was different

- Number of rounds
  - 50 instead of 200
- Payoff matrix
  - (-4,-1,1,4) instead of (-10,-1,1,10)
- Human vs. confederate agent
  - instead of human vs. human
- Participants
  - Amazon's Mechanical Turk workers
  - Instead of CMU undergraduates
- Order conditions
  - (PDPD, PDCG, CGPD, CGCG)
    - instead of (PDCG, CGPD)
- Strategy of confederate agent
  - Tit-for-tat
  - Pavlov-tit-for-tat
- Trustworthiness of confederate agent
  - Low trustworthiness
  - High trustworthiness
- Self-report measures
  - Trait trust (added)
  - State trust

16

Study design

Game Order		Confederate Agent	
First Game	Second Game	Strategy	Trustworthiness
PD	PD	T4T	High Trustworthiness
PD	CG	T4T	High Trustworthiness
CG	PD	T4T	High Trustworthiness
CG	CG	T4T	High Trustworthiness
PD	PD	T4T	Low Trustworthiness
PD	CG	T4T	Low Trustworthiness
CG	PD	T4T	Low Trustworthiness
CG	CG	T4T	Low Trustworthiness
PD	PD	PT4T	High Trustworthiness
PD	CG	PT4T	High Trustworthiness
CG	PD	PT4T	High Trustworthiness
CG	CG	PT4T	High Trustworthiness
PD	PD	PT4T	Low Trustworthiness
PD	CG	PT4T	Low Trustworthiness
CG	PD	PT4T	Low Trustworthiness
CG	CG	PT4T	Low Trustworthiness

17

Trait and state trust scales

- **Trait trust scale (24 items):**
  - I generally trust other people unless they give me a reason not to.
  - One is better off being cautious when dealing with strangers until they have provided evidence that they are trustworthy.
- **State trust scale (14 items):**
  - I believe that the other player wants to help me to make a good amount of payoff in this game.
  - The other player can be trusted.
  - I would not let the other player have any influence over my payoff.

18

Predicted vs. observed data

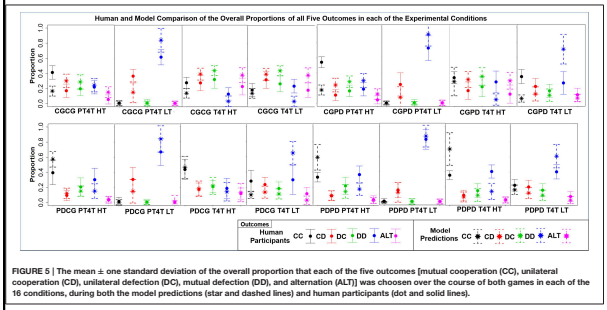
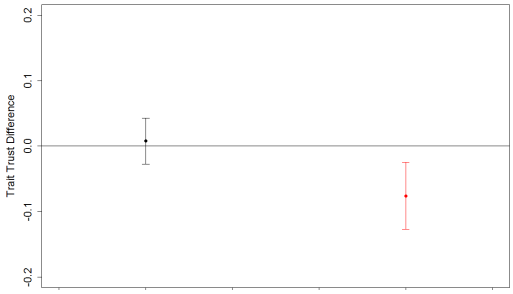


FIGURE 5 | The mean ± one standard deviation of the overall proportion that each of the five outcomes [mutual cooperation (CC), unilateral cooperation (CD), unilateral defection (DC), mutual defection (DD), and alternation (ALT)] was choosen over the course of both games in each of the 16 conditions, during both the model predictions (star and dashed lines) and human participants (dot and solid lines).

19

Trait Trust Difference



20

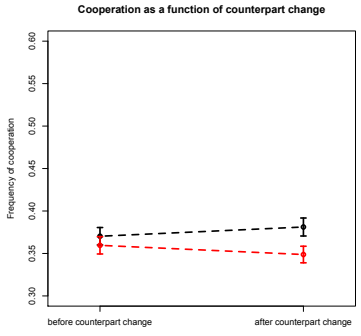
Study 1 Conclusions

- Majority of model predictions was observed
- Deviations from predictions
  - Humans cooperate
    - More against low-trustworthy agents than predicted
    - Less against high-trustworthy agents than predicted
  - Different setups, different populations
- Relationship between trait trust and state trust
  - Both trait trust and state trust are learned.
- Trust asymmetry

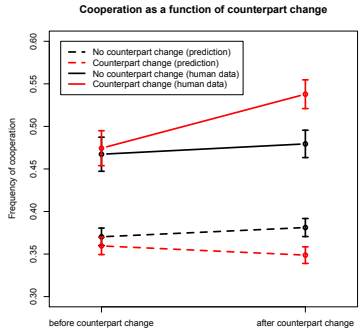
Model validation cont'd.

- Validation study 2
  - Same design as validation study 1 (16 conditions)
  - Added counterpart-change manipulation (32 conditions)
- *A priori* model predictions (published)
- Larger set of human data (640 participants)
  - Not published yet.

Results



Results



Critique of the published model

- Strengths (skip for now)
- Limitations

25

Limitations

- No link between trait and state trust
  - State trust starts at zero?
    - Initial trust is high (McKnight, Cummings, & Chervany, 1998)
    - Trait trust influences state trust (Berg, Dickhaut, & McCabe, 1995; Dirks & Ferrin, 2001)
  - State trust influences trait trust (Collins et al., 2016)
    - **Novel finding**
    - Relevant for interactions with multiple trustees in sequence

26

Limitations (cont'd)

- Learning equation was linear
  - $ST_t = ST_{t-1} + PET_t$
- Did not explain
  - Trust asymmetry (Slovic, 1993)
  - Early evidence more important than late evidence (Lount et al., 2008)
- Learning equations tend to be power functions (Newell & Rosenbloom, 1981)
- No link between cognitive ability and learned trust
  - Cognitive ability is one of the best predictors of learning and performance

27

The revised model

- Goal:
  - Overcome limitations of the published model
  - Expand the model's scope of applicability
  - External validity

Juvina, I., Collins, M.G., Larue, O., & de Melo, C. (2016). *Toward a unified theory of learned trust*. Paper presented at the International Conference on Cognitive Modeling, State College, PA.

28

Terminology revised model

- Initial state trust = trait trust
  - Modulo some incidental learning (emotions, gaze etc.)
- Trait trust = f(prior state trusts)
  - (Collins et al., 2016)
- Trait trust deviation = TT_end – TT_start
- Cognitive ability = accuracy of judgments of trustworthiness and trust necessity

29

Revised trust learning equation

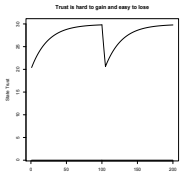
$$ST_t = ST_{t-1}^{a} + PET_t - b * TTD$$

- ST = state trust
- a = power exponent; a < 1
- PET = perceived evidence of trustworthiness
- TTD = trust propensity deviation
- b = perception bias

30

Trust asymmetry

- Trust is hard to gain but easy to lose
  - Empirical evidence?
    - Slovic, 1993; Collins et al., 2016.
  - Example: organizations (re)gaining the trust of their customers



31

TTD and perception bias

- Experience from a previous interaction influences perception of trustworthiness in current interaction
  - E.g., Black-hat/white-hat (or bad-cop/good-cop) effect
    - De Melo, Carnevale, and Gratch (2011); Hilty & Carnevale, 1993
    - a cooperative second bargainer is judged as more cooperative if the first bargainer was competitive rather than cooperative.
- Theories of adaptation (Helson, 1964).
  - people become accustomed to a reference point as a result of prior experience;
  - this point then serves as a comparison for the judgment of subsequent experiences.

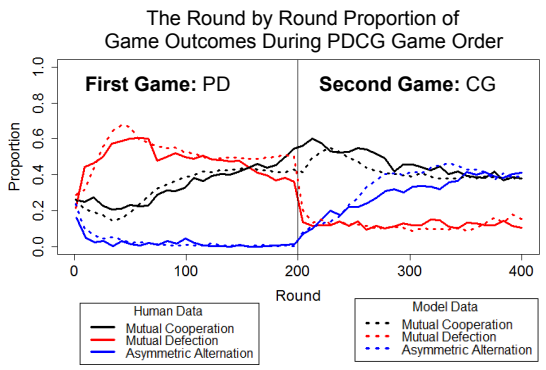
32

Validation of revised model

- Test that revised model:
  - Maintains published model's strengths
  - Explains other (seemingly unrelated) findings
  - Makes novel predictions

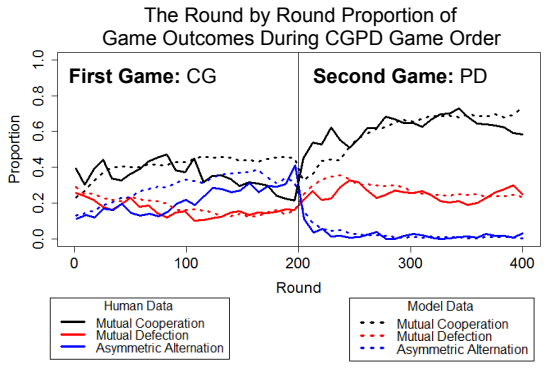
33

Fit of revised model to Juvina et al.(2013)



34

Fit of revised model to Juvina et al.(2013)



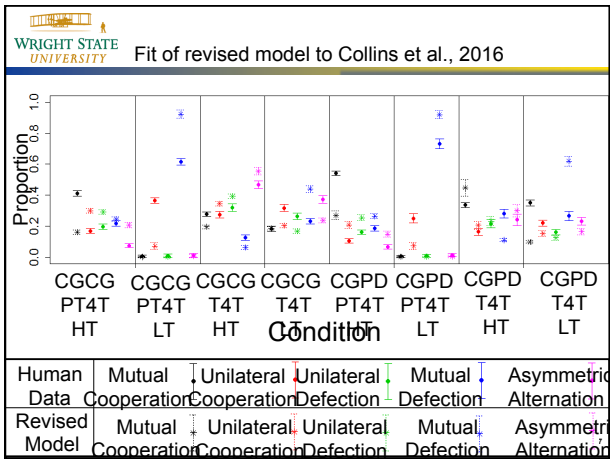
35

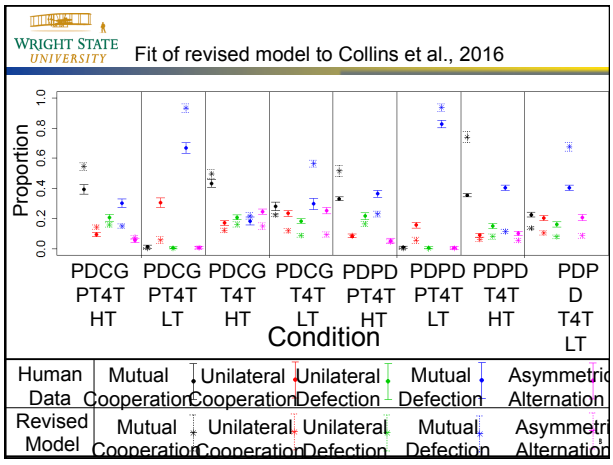
Best fitting parameter values and fit stats

Outcome	Published model		Revised model	
	Trust	Invest	Trust	Invest
CC	3	NA	6	NA
CD	-10	-1	-7	-1
DC	10	NA	9	NA
DD	-1	.18	-1*	.18*

- **Published model: *r* = .89, *RMSD* = .09**
- **Revised model: *r* = .90, *RMSD* = .07**

36



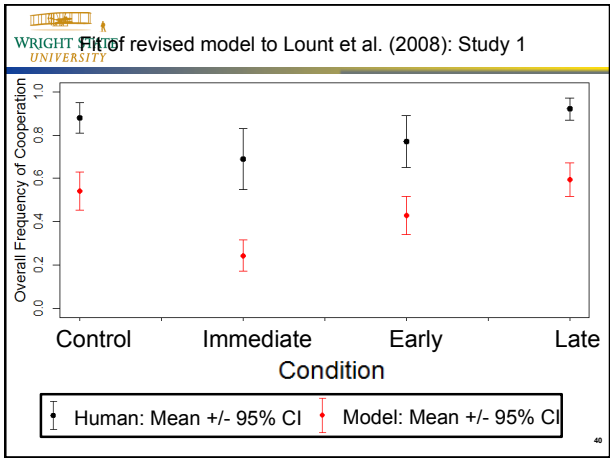


WRIGHT STATE UNIVERSITY

Fit stats Collins

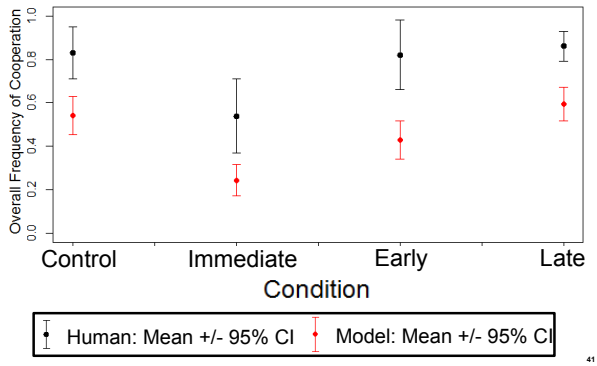
- **Published model: *r* = .64, *RMSD* = .33**
- **Revised model: *r* = .68, *RMSD* = .33**

39



40

Fit of revised model to Lount et al. (2008): Study 2



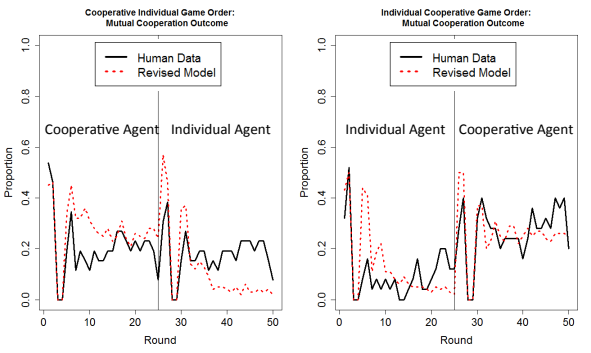
41

Fit stats Lount

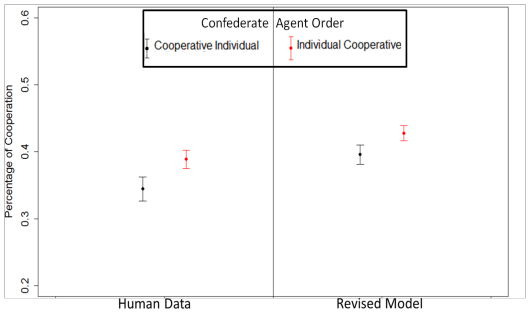
- **Revised model: *r* = .99, *RMSD* = .33**

42

Fit of De Melo et al., 2011



Fit of black-hat/white-hat effect



44

Fit stats De Melo

- **Revised model: *r* = .86, *RMSD* = .11**

45

Conclusions

- Model has potential to unify theories on learned trust
  - Cumulates learning from
    - History of prior interactions (trait trust)
    - Evidence of trustworthiness and trust necessity in current interaction
    - (Incidental learning from facial expressions)

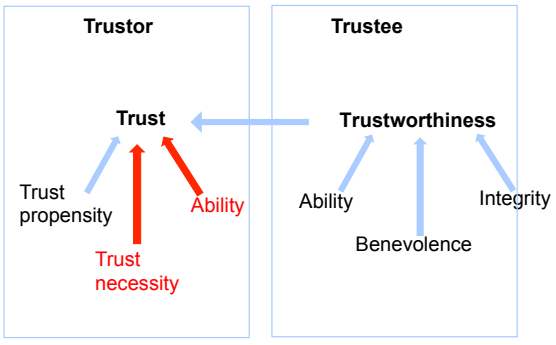
46

Predictions

- Trust decay
  - Evidence?
- Cognitive ability influences state and trait trust
  - Lyons, Stokes, & Schneider, 2011;
  - Sturgis, Read, & Allum, 2010;
  - Yamagishi, Kikuchi, & Kosugi, 1999

47

Contributions to trust theory



48

Work in progress and future work

- Empirical
  - Using models as autonomous agents
    - Study in progress
  - Study to test the trust decay prediction
- Modeling:
  - How state trust influences trait trust
    - TT = f(ST)
  - Dynamics of trait trust and perception bias
    - b = f(TTD)

49

Collaborators & funding

- **Kevin Gluck**
- **Celso de Melo**
- **Randall Green**

The work presented here was supported by The Air Force Office of Scientific Research grant number FA9550-14-1-0206 to Ion Juvina and the Oak Ridge Institute for Science and Education (ORISE) who supported this research by appointing Michael Collins to the Student Research Participant Program at the U.S. Air Force Research Laboratory (USAFRL), 711th Human Performance Wing, Human Effectiveness Directorate, Warfighter Readiness Research Division, Cognitive Models and Agents Branch administered by the ORISE through an interagency agreement between the U.S. Department of Energy and USAFRL.

50

Questions?

51