Autonomous agents and human cultures in the trust–revenge game

Authors: Amos Azaria, Ariella Richardson, Avi Rosenfeld

Presented by: The Whales
Pedro Albuquerque, Siya Kunde, Rubi Quiñones
30 November 2017

Overview

- Introduction
- Related Work
- Experiment Procedure
- Results
- Conclusions
- Our Conclusions

TRUST

FEAR

REVENGE

Game Techniques

- To determine whether the behavior of non game theory agents (NGTE) is similar to human behavior from difference cultures

- The Dictator Game
- The Investment Game
- The Trust-Revenge Game

Dictator Game

1. Player A starts with all the money.
2. (TRUST) Player A may contribute any amount to Player B.
   a. Player A cannot attempt to gain anything from transferring chips over the Player B in the trust stage.
3. (RECIPROCATE) Player B may return some (or all) the money received from Player A.
   a. Any amount transferred in this setting may be attributed to generosity.

The Investment Game

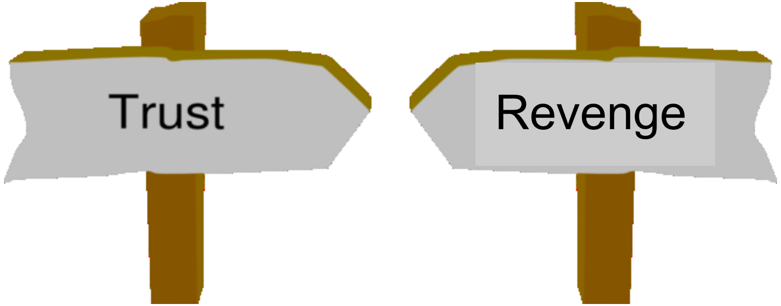
1. Player A and Player B are given 10 chips each at the beginning of the game.
2. (TRUST) Player A can give some or all of their chips to. Player B.
    a. The number of chips that Player A decides to give is multiplied by 3 (trust rate).
3. (RECIPROCATE) Player B can give back some or all of what he was given.

Trust rate are common knowledge and revealed to both players at the beginning of the game.

The Trust-Revenge Game

1. Player A and Player B are given a certain number of chips each at the beginning of the game.
2. (TRUST) Player A can give some or all of their chips to. Player B.
   a. The number of chips that Player A decides to give is multiplied by 3 (trust rate).
3. (RECIPROCATE) Player B can give back some or all of what he was given.
4. (REVENGE) Player A pays any number of chips to the operator.
   a. Player B must pay to the operator the number of chips Player chose for revenge multiplied with the revenge rate.

Trust rate and revenge rate are common knowledge and revealed to both players at the beginning of the game.



Related Work

1. Willinger [38] compared French and German players using the investment game
   a. **Results**: German players invested more than French players
   b. **Results**: Reciprocating was no different between the groups
2. Berge [25] conducted experiments with students to test the subgame-perfect equilibrium
   a. **Results**: Students did not follow the equilibrium
3. Gneezy [27] experimented with the trust-revenge game
   a. **Results**: Player A takes revenge on Player B when Player B keeps all of the money

- Agent design by game theory expert vs. non game theory experts
- Subgame perfect equilibrium : SPE of a game G is a Nash Equilibrium of G that corresponds to a Nash Equilibrium in every subgame of G.

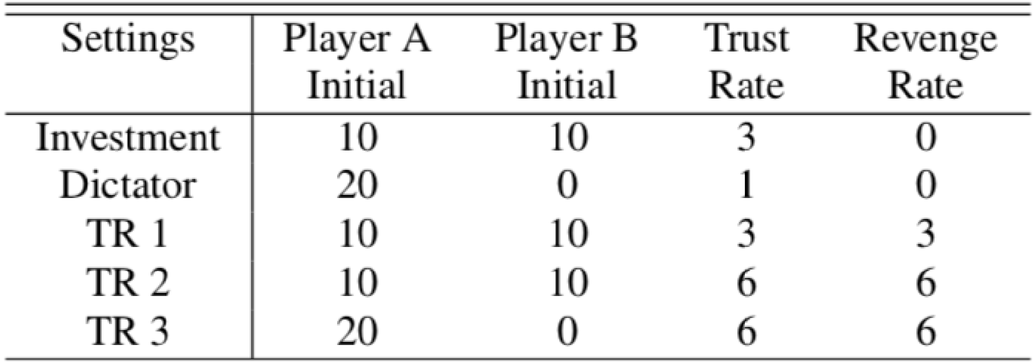
Objective:

To determine whether the behavior of NGTE agents is similar to human behavior from different cultures.

Experimental Setup

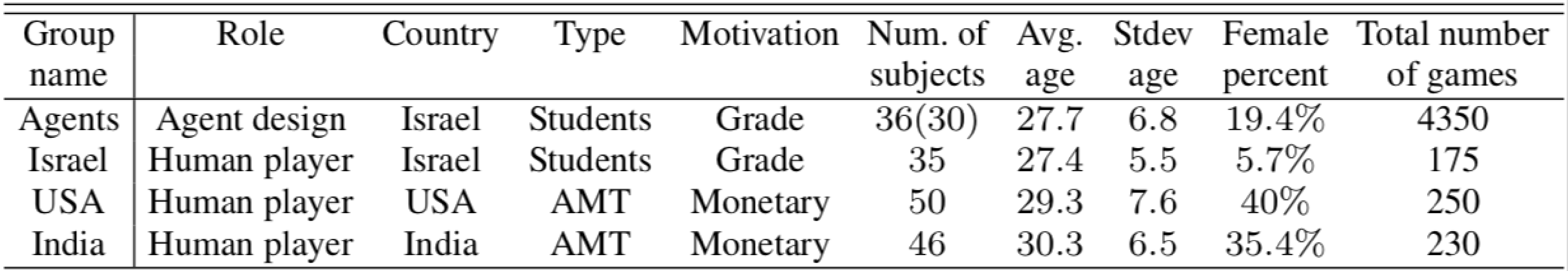
- Game Settings
- Subjects
- Number of Games and Motivation

Game Settings



Settings	Player A Initial	Player B Initial	Trust Rate	Revenge Rate
Investment	10	10	3	0
Dictator	20	0	1	0
TR 1	10	10	3	3
TR 2	10	10	6	6
TR 3	20	0	6	6

Subjects

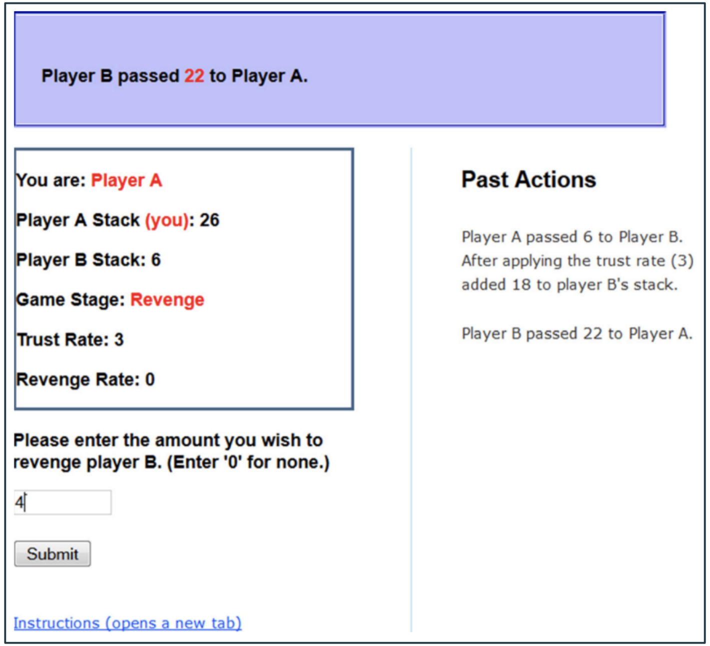


Group name	Role	Country	Type	Motivation	Num. of subjects	Avg. age	Stdev age	Female percent	Total number of games
Agents	Agent design	Israel	Students	Grade	36(30)	27.7	6.8	19.4%	4350
Israel	Human player	Israel	Students	Grade	35	27.4	5.5	5.7%	175
USA	Human player	USA	AMT	Monetary	50	29.3	7.6	40%	250
India	Human player	India	AMT	Monetary	46	30.3	6.5	35.4%	230

Number of Games and Motivation

- Autonomous agents played 290 games.
- Human agents played 10 games.
- Motivated by grades and monetary incentives.

UI



Player B passed 22 to Player A.

You are: Player A

Player A Stack (you): 26

Player B Stack: 6

Game Stage: Revenge

Trust Rate: 3

Revenge Rate: 0

Please enter the amount you wish to revenge player B. (Enter '0' for none.)

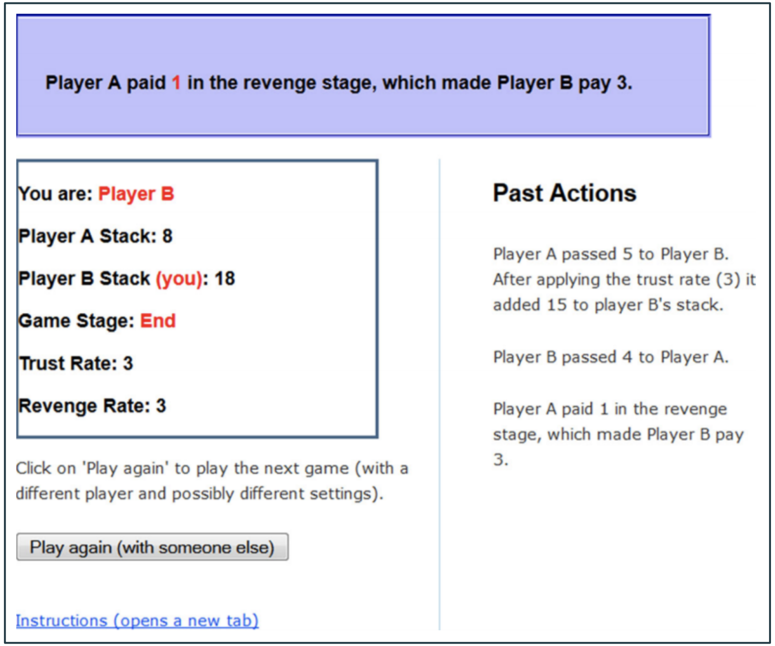
4

Instructions (opens a new tab)

Past Actions

Player A passed 6 to Player B. After applying the trust rate (3) added 18 to player B's stack.

Player B passed 22 to Player A.



Player A paid 1 in the revenge stage, which made Player B pay 3.

You are: Player B

Player A Stack: 8

Player B Stack (you): 18

Game Stage: End

Trust Rate: 3

Revenge Rate: 3

Click on 'Play again' to play the next game (with a different player and possibly different settings).

Instructions (opens a new tab)

Past Actions

Player A passed 5 to Player B. After applying the trust rate (3) it added 15 to player B's stack.

Player B passed 4 to Player A.

Player A paid 1 in the revenge stage, which made Player B pay 3.

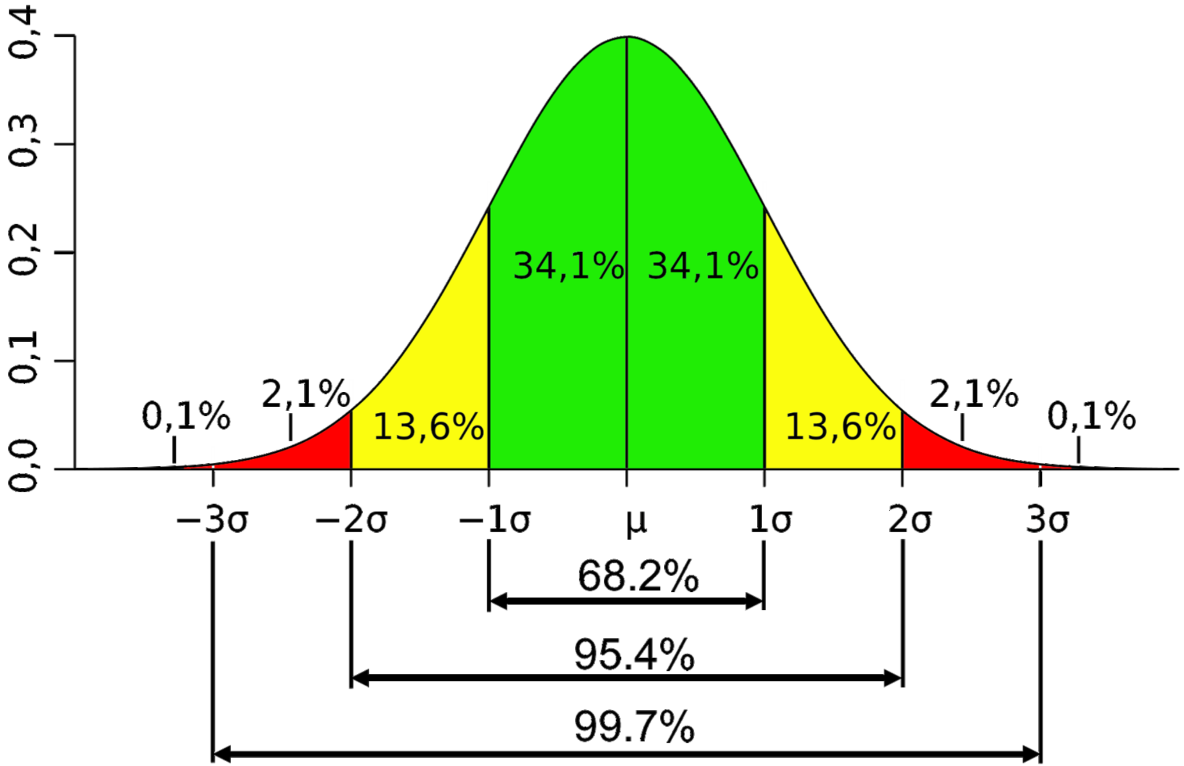
Results

- Main question is whether NGTE behavior falls within cultural diversities.
- To be considered part of the diversity of the other groups:

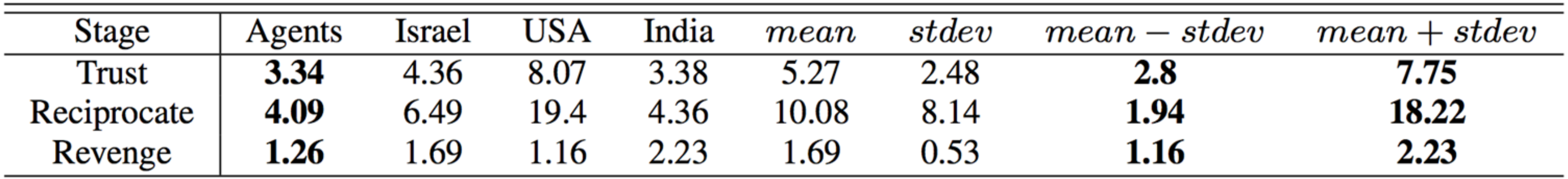
$$\mathcal{B} = \{B_1, B_2, B_3, ...\}$$

$$avg(A) \in avg(\cup\mathcal{B}) \pm stdev(\{avg(B_1), avg(B_2), avg(B_3), ...\}).$$

Population within 1 std → 68.2%



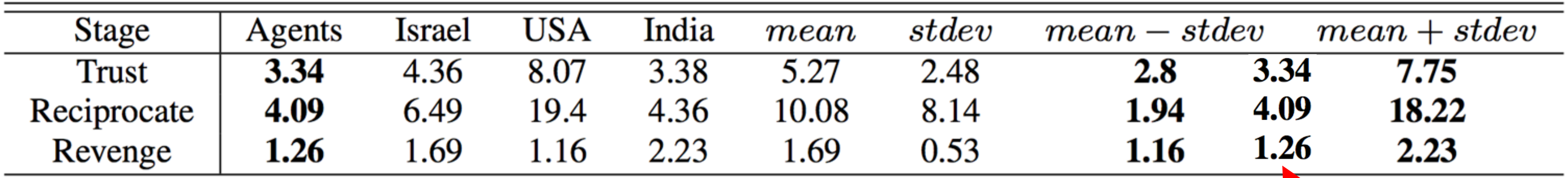
Number of chips per Stage



Stage	Agents	Israel	USA	India	$mean$	$stdev$	$mean - stdev$	$mean + stdev$
Trust	**3.34**	4.36	8.07	3.38	5.27	2.48	**2.8**	**7.75**
Reciprocate	**4.09**	6.49	19.4	4.36	10.08	8.14	**1.94**	**18.22**
Revenge	**1.26**	1.69	1.16	2.23	1.69	0.53	**1.16**	**2.23**

- The activity of the agents falls within one standard deviation of the average of the three human cultures.
- This indicates that autonomous agents built by NGTE can indeed be treated within cultural diversities.

Number of chips per Stage

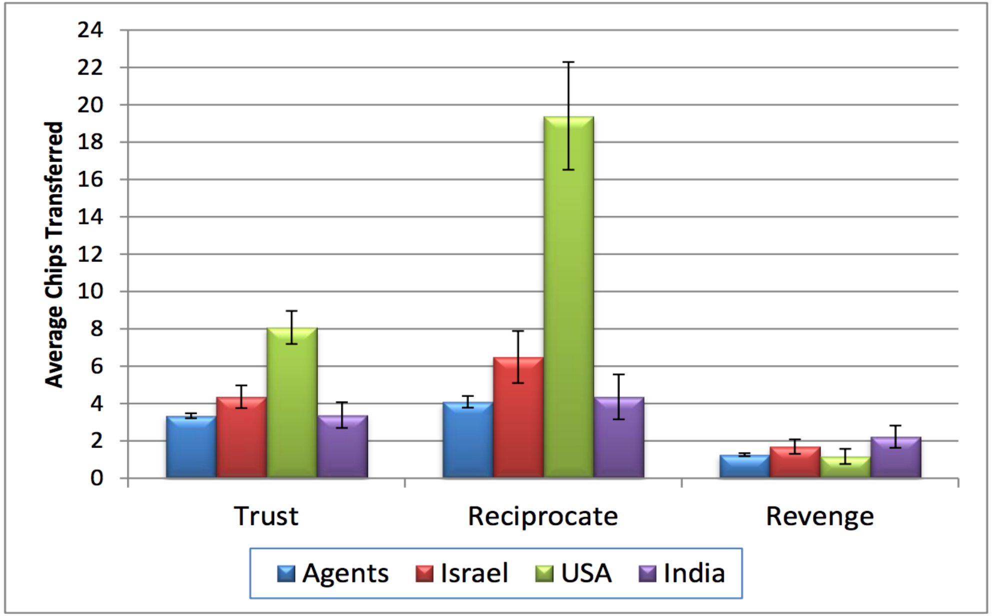


Stage	Agents	Israel	USA	India	*mean*	*stdev*	*mean − stdev*		*mean + stdev*
Trust	**3.34**	4.36	8.07	3.38	5.27	2.48	**2.8**	**3.34**	**7.75**
Reciprocate	**4.09**	6.49	19.4	4.36	10.08	8.14	**1.94**	**4.09**	**18.22**
Revenge	**1.26**	1.69	1.16	2.23	1.69	0.53	**1.16**	**1.26**	**2.23**

Agents average

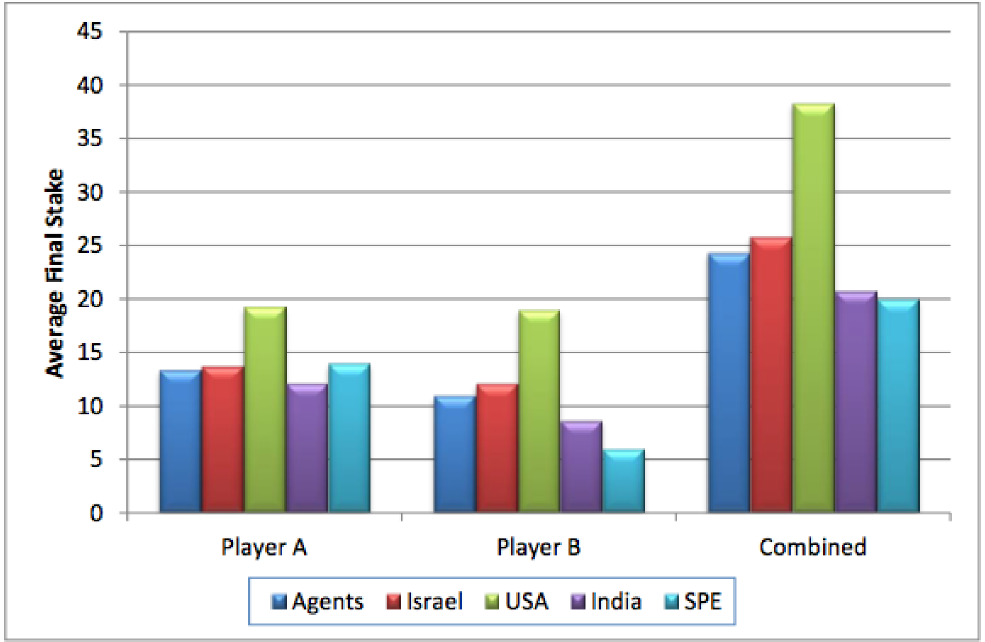
Chips given by stage

- On average, Player B reciprocated more than Player A trusted, with both humans and autonomous agents
- It is fair to assume that the agent designers thought it might be beneficial to display trust as their agent might be rewarded.
- On the other hand, is clearly not an optimal behavior which is present in all the four groups.

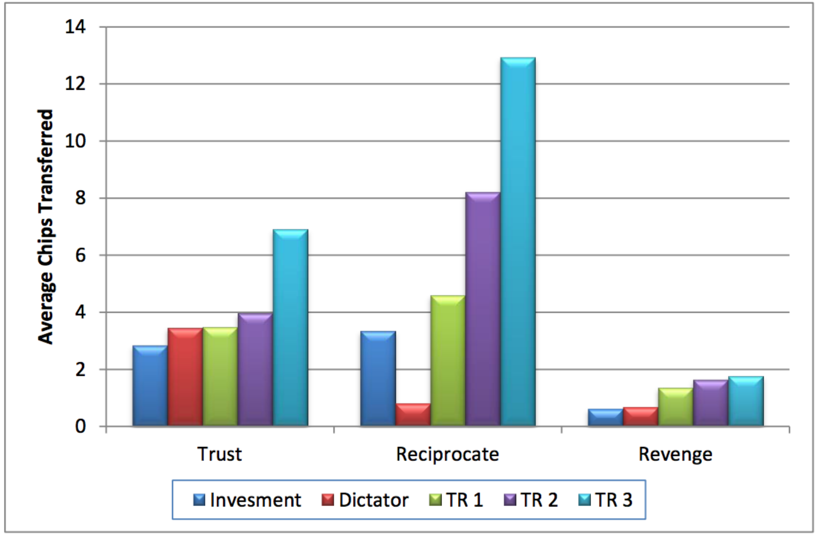


Overall profit

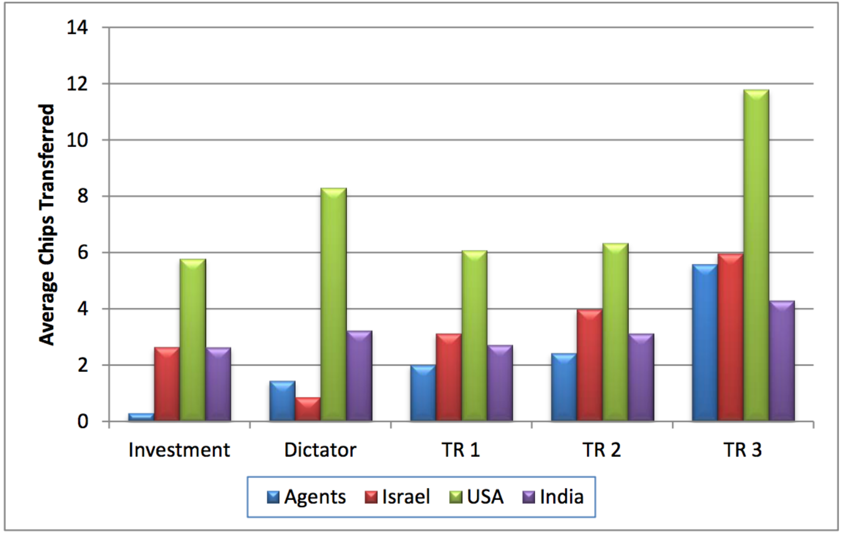
- SPE does not achieve the highest outcome.
- Autonomous achieved a similar score of their own culture.



Overall performance



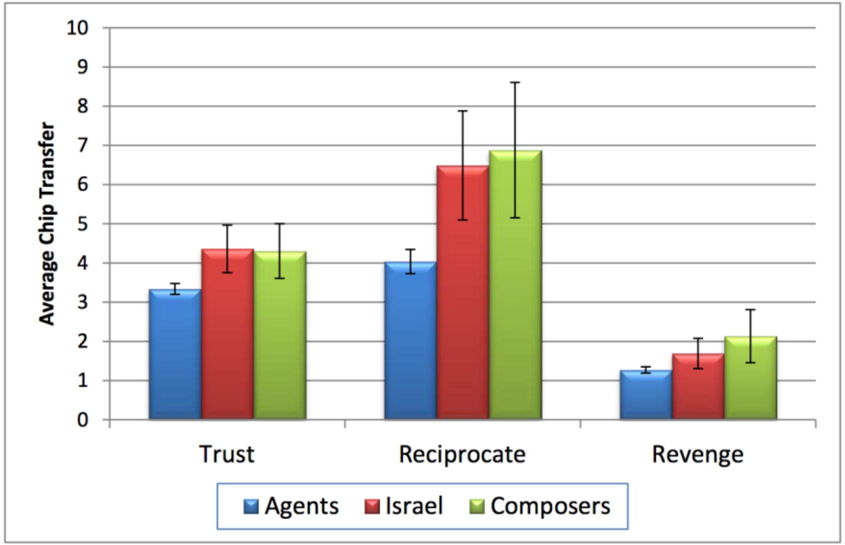
Average chips transferred per stage in each setting.



Average chips transferred per game setting.

Composers behavior

- Composers did indeed take revenge on average 62% less than the Israeli group (Investment and Dictator)
- However, composers took revenge 70% more than their own agents.
- Therefore, the only impact that building agents had on the Composers was the reduction of human error.
- No statistical difference in behavior to other subjects.



Results Summary

- Expert agents that interact with NGTE agents can use the same models developed for modeling cultural diversities within humans, for modeling the NGTE agents.
- NGTE agents' behavior was closer to that of the subgame-perfect equilibrium.
- NGTE agents were less prone to human error.
- Composing the agents had no impact on human behavior aside of possibly reducing error rate.

Limitations & Future Work

- Were there hidden motivations for the subjects behaviors?
- NGTE agents behavior was within the diversity of different human cultures.
- Compared human to human, and human-agent. What about agent to agent?

Conclusions

- Humans and NGTE agents did not follow the subgame-perfect equilibrium when playing the game.
- Average action performed by NGTE agents was within one standard deviation of the average action of the three human cultures.
- Taking revenge is attributed to emotional human behavior or the search for justice.

Our Conclusions

- Games as Trust-Revenge Game can provide substantial psychological information.
- This paper presents promising game techniques for games such as Poker.
- NGTE modelling was biased since it was not a diverse group but only comprised of Israeli people.
- Subgame perfect equilibrium (SPE) is a refinement of Nash equilibrium used in dynamic games.