Yosef S. Razin and Karen M. Feigh

Committing to Interdependence: Implications from Game Theory for Human-Robot Trust
**Trust: HRI vs Game Theory**

**HRI**
- Continuously-valued belief
- Embedded interaction
- Multi-dimensional and layered
- Construct proliferation

**Game Theory (Social Psych)**
- Discrete decisions
- Lab-based (‘toy problems’)
- Strong agreement on single definition
- Defined based on payoff structure

---

**Hoff & Bashir (2015)**
Game Theory for Trust

Requirements for Trustor
- Exposure: \( A_{12} < \{A_{21}, A_{22}\} \)
- Improvement: \( A_{11} > \{A_{21}, A_{22}\} \)

Requirements for Trustee
- Temptation: \( B_{12} > B_{11} \)
- Mutual Gain*: \( B_{11} > B_{21} \)

* Not universally accepted
Interdependence Theory for Trust

Interaction weights of control modalities

- **Reflexive Control (RC):** Unilateral power an actor has over their own outcomes

RC_A = 0.5 * ((A_{11} + A_{12}) - (A_{21} + A_{22}))

RC_B = 0.5 * ((B_{11} + B_{21}) - (B_{12} + B_{22}))
Interdependence Theory for Trust

Interaction weights of control modalities

❖ **Reflexive Control (RC):** Unilateral power an actor has over their own outcomes
❖ **Fate Control (FC):** Unilateral power the other actor has over this actor’s outcomes

\[
FC_A = 0.5 \times ((A_{11} + A_{21}) - (A_{12} + A_{22}))
\]

\[
FC_B = 0.5 \times ((B_{11} + B_{12}) - (B_{21} + B_{22}))
\]
Interdependence Theory for Trust

Interaction weights of control modalities

- **Reflexive Control (RC):** Unilateral power an actor has over their own outcomes
- **Fate Control (FC):** Unilateral power the other actor has over this actor’s outcomes
- **Bilateral Control (BC):** The power an actor has to facilitate or inhibit the other’s outcomes through coordination

\[
BC_A = 0.5 \times ((A_{11} + A_{22}) - (A_{12} + A_{21}))
\]

\[
BC_B = 0.5 \times ((B_{11} + B_{12}) - (B_{21} + B_{22}))
\]
Interdependence Theory for Trust

Requirements for Trustor

| FC_A > 0          | | RC_A |
|-------------------|------------------------|
| BC_A > 0          | | RC_A |

- RC is the player’s commitment (and thus vulnerability)
- Gottman’s proposed Trust Index

\[ T_G = \frac{1}{2} + \frac{RC_A}{2FC_A} \]

- Trust requires the trustee have unilateral or cooperative influence that incentivizes trust and is stronger than the trustor’s own control
- i.e. If I can already do something myself, I don’t need to trust someone else
- This recalls the automation vs self-confidence trust measures that go back to Lee and Morray

Coerced Trust: 

- \( RC_A < 0 \)
- \( T_G = 0.5 \)

Freely Given Trust:

- \( RC_A > 0 \)
- \( 1 < T_G \)
Hypotheses & Method

❖ H1: Gottman’s trust index ($T_G$) will positively predict whether the trustor chooses to trust
❖ H2a: Commitment (RC) is predictive whether the trustor chooses to trust
❖ H2b: Commitment (RC) is predictive whether the trustee chooses to be trustworthy

❖ Game theory data set (Erev, 2002)
❖ 106 Unique Non-Trivial Games that meet trustor requirements
❖ One-off discrete interactions between humans (116 participants)

❖ Compare our models that included interdependence against previous algorithms proposed
❖ Sub-game perfect equilibrium
❖ Inequality aversion
❖ Equality reciprocity
❖ Kindness/Fairness (Charness & Rabin)
❖ “Seven Strategies” (combo of equality, fairness, rationality, altruism)
Results

When Trustors Trust

❖ H1, H2a supported
❖ It was the trustee’s commitment ($\text{RC}_B$) that predicted the trustor’s choice!
❖ Not only do commitment and the trust index predict when the trustor trusted but they were significantly stronger predictors than the previous proposed game theory-based methods

When Trustees are Trustworthy

❖ H2b partially supported. The ‘rational’ sub-game equilibrium dominates, but this IS equivalent to the sign of $\text{RC}_B$

Best Fit Trustor Regression Tree
Discussion

- Interdependence provides a simpler and stronger predictive model than previous ones from game theory.
- By shifting the trust discussion in game theory away from kindness, fairness, altruism, and equality, and replacing it with perceived commitment, control, risk, and vulnerability, how trust is bestowed becomes more applicable to HRI.

Liliana Xu (2018)
Challenges & Future Work

❖ We plan to run a similar set of games between humans and robots/AI
❖ Game theory allowed us to ignore issues of capability, familiarity, likability, and understandability. This ‘purer trust’ may sacrifice validity for reliability - what part of trust is this capturing?
❖ Do discrete decisions translate into continuous interactions? How does this trust play out over time?

Stay tuned for my next talk on our work on what’s going on before the trust decision is made
Thank you!

Questions? Comments?

yrazin@gatech.edu
Condition 3: Basic Requirements

Trust requires both Other's and Joint Control that are positive, concordant, and stronger than the Personal control.

- I.e. If I can already do something myself, I don’t need to trust someone else.

Plugging these into Gottman’s Trust index,

\[ T_G = \frac{1}{2} + \frac{PC_A}{2OC_A} \]

- \(T_G<0\): No Trust
- \(1<T_G\): No Trust
- \(T_G=0.5\): Forced Trust: \(PC_A\) is discordant with \(OC_A\)
- Freely Given Trust: \(PC_A\) is concordant with \(OC_A\)

\(T_G\) loses key information regarding the concordance and relative size of \(JC_A\) vs \(OC_A\)
Extending Interdependence Theory for Trust

- Generalization of Game Theory
  - Drops the rationality assumption
  - Drops the goal of optimizing individual outcomes to various possible social goals (e.g. max. joint outcomes, maximal equity, minimal competition)
- Generalizes “Cost”

TABLE I: Outcome Matrix \( O \) with outcomes of Actor A, \( O_A \), and Actor B, \( O_B \), in a trust-trustworthiness interaction.
Encoding Interdependence

<table>
<thead>
<tr>
<th>Trustworthy</th>
<th>Not Trustworthy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trust</td>
<td>Nice!</td>
</tr>
<tr>
<td>Don’t Trust</td>
<td>Silly, Obstinate</td>
</tr>
</tbody>
</table>
Encoding Interdependence

<table>
<thead>
<tr>
<th>Interaction weights of control modalities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Personal Control (PC): Unilateral power an actor has over their own outcomes</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>OA</th>
<th>Trustworthy</th>
<th>Not Trustworthy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trust</td>
<td>5 + -10/2</td>
<td></td>
</tr>
<tr>
<td>Don’t Trust</td>
<td>-5 + 3/2</td>
<td>-1.5 (PC_A)</td>
</tr>
</tbody>
</table>
Encoding Interdependence

Interaction weights of control modalities

- **Personal Control (PC):** Unilateral power an actor has over their own outcomes
- **Other’s Control (OC):** Unilateral power the other actor has over this actor’s outcomes

<table>
<thead>
<tr>
<th>OA</th>
<th>Trustworthy</th>
<th>Not Trustworthy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trust</td>
<td>5</td>
<td>-10</td>
</tr>
<tr>
<td>Don’t Trust</td>
<td>-5</td>
<td>3</td>
</tr>
</tbody>
</table>

\[ O_A = \frac{-5 + 3}{2} = 3.5 \text{ (OC}_A) \]
Encoding Interdependence

Interaction weights of control modalities

- **Personal Control (PC)**: Unilateral power an actor has over their own outcomes
- **Other’s Control (OC)**: Unilateral power the other actor has over this actor’s outcomes
- **Joint Control (JC)**: The power an actor has to facilitate or inhibit the other’s outcomes through coordination

<table>
<thead>
<tr>
<th>OA</th>
<th>Trustworthy</th>
<th>Not Trustworthy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trust</td>
<td>5</td>
<td>-10</td>
</tr>
<tr>
<td>Don’t Trust</td>
<td>-5</td>
<td>3</td>
</tr>
</tbody>
</table>

\[ \frac{-10}{2} + \frac{3}{2} = \frac{11}{2} \]
# Encoding Interdependence

<table>
<thead>
<tr>
<th>$O_B$</th>
<th>Trustworthy</th>
<th>Not Trustworthy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trust</td>
<td>3</td>
<td>-5</td>
</tr>
<tr>
<td>Don’t Trust</td>
<td>-2</td>
<td>-1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$O_A$</th>
<th>Trustworthy</th>
<th>Not Trustworthy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trust</td>
<td>5</td>
<td>-10</td>
</tr>
<tr>
<td>Don’t Trust</td>
<td>-5</td>
<td>3</td>
</tr>
</tbody>
</table>

- $PC_B = 3.5$ (Concordant)
- $OC_B = 0.5$ (Concordant)
- $JC_B = 4.5$ (Correspond)

- $PC_A = -1.5$ (Discordant)
- $OC_A = 3.5$ (Concordant)
- $JC_A = 11$ (Correspond)
Trust Conditions (adpt. Wagner)

1. Deciding to Trust
2. Setting Thresholds
3. Basic Interaction Requirement
4. Trust is worth the risk
5. Distrust outcomes are independent of the trustee’s action
Condition 1: Deciding to Trust

Comparison Level for the Alternative ($CL_{alt}$)

- Trust is not the same as commitment
- The more attractive alternatives, the less committed
  - There might be someone more trustworthy
- The more committed, the less power/autonomy
- The more attractive the alternative the higher $CL_{alt}$

Proposed Commitment Transformation

\[ PC_A - CL_{alt} = PC'_A \]
Condition 2: Setting a Threshold

The trust probability threshold can be derived from the Nash equilibrium if we are willing to assume that the trustee is rational:

\[ p(TW) > \tau_B = \frac{A_{22} - A_{12}}{A_{11} + A_{22} - A_{12} - A_{21}} = \frac{1}{2} - \frac{PC_A}{2JC_A} \]

The trustee is actually trustworthy if they act trustworthy more often than \( \tau_B \)

During the trust fall, we only have a single-shot test, which is not meaningful

Gottman’s Indices

Trust and untrustworthiness indices are validated, independent of the rationality assumption, independent of each other, and non-statistical

\[ T_G = \frac{A_{11} - A_{22}}{A_{11} + A_{21} - A_{12} - A_{22}} = \frac{1}{2} + \frac{PC_A}{2OC_A} \]

\[ TW_G = \frac{B_{11} - B_{22}}{B_{11} + B_{12} - B_{21} - B_{22}} = \frac{1}{2} + \frac{OC_B}{2PC_B} \]
Condition 3: Basic Requirements

Trust requires both Other's and Joint Control that are positive, concordant, and stronger than the Personal control.

- I.e. If I can already do something myself, I don’t need to trust someone else.

Plugging these into Gottman’s Trust index,

\[ T_G = \frac{1}{2} + \frac{PC_A}{2OC_A} \]

- No Trust
- Forced Trust: \( T_G = 0.5 \)
- Freely Given Trust: \( PC_A \) is concordant with \( OC_A \)
- No Trust

\( T_G < 0 \) \( PC_A \) is discordant with \( OC_A \)

\( 1 < T_G \)

\( T_G \) loses key information regarding the concordance and relative size of \( JC_A \) vs \( OC_A \)
Combining Conditions 4 & 5: When to Trust

- Condition 4: Minimal acceptable payoff vs. risk \((A_{11} - A_{12} < e_1)\)
- Condition 5: Distrust outcomes are independent of the Trustee’s action \((|A_{21} - A_{22}| < e_2)\)
- We derive a new joint bound which suggests that Condition 5 is incorrect, distrust is not independent of the trustee’s action

\[
A_{22} - A_{21} = JC_A - OC_A
\]

If \(OC_A > JC_A\)

\[
A_{21} - A_{22} < \epsilon_2 \Rightarrow \frac{e_1 - e_2}{2} = t_A < PC_A
\]

Else-If \(JC_A > OC_A\)

\[
A_{22} - A_{21} < \epsilon_2 \Rightarrow \frac{e_1 - e_2}{2} = t_A < OC_A
\]
Combining Conditions 4 & 5: When to Trust

- If $OC_A > JC_A$
  - The more one feels powerless, the more regret they suffer and the more they risk hurting their future relationships, so they must make sure the other is trustworthy ($PC_A > t_A$)
  - Validation- Gottman: People who set too high of a joint trust threshold, end up suffering high costs including social isolation and increased risk of disease and even early death

- If $JC_A > OC_A$
  - The more one attributes outcomes to collaboration, the more self-satisfied for avoiding bad situations. The trustor can choose how to set expectations for themselves and their partner ($OC_A > t_A$)
  - Validation- Yamagishi: High trustor’s were better at judging when trusting others was right and exhibit higher emotional and social intelligence. They also start with a higher a priori disposition to trust ($JC_A$), which sometimes comes off as gullible but actually reflects their ability to risk more emotionally...
A New Synthesis of Trustworthiness
Specific Implications for Trust in HCI

- Automation as a tool is simply assessed by its \( OC \)
  - Can be taken on faith or tested for reliability
  - Unique capabilities and high specialization reduce \( CL_{alt} \) and thus raise \( PC \) & \( T_G \)
  - The longer it is used and adapted to, the more committed (Sunk Cost)
- Performance transparency: What is it doing?

- Interaction and Autonomy increase \( JC \)
  - Purpose transparency: Why is it doing?
  - Increases need for trust signaling
    - Anthropomorphism?
      - Decreases initial trust
    - Better resilience
General Implications for Trust in HCI

- Trust interactions can be pre-specified and designed to elicit certain interdependence and power structures
- If the anticipated interaction outcomes can be measured or estimated (see Wagner), the interaction environment can be assessed for suitability of trust/trustworthiness
- Competitive-irrational behavior can be detected from affect (Gottman) and trust-based design modified accordingly
- Systems can judge the trustworthiness of their users and how much their users trust them
Trust Me
Wagner’s 1st: Commitment

Proposed Commitment Transformation

- No effect on \( OC_A \) or \( JC_A \)
- Increasing \( CL_{alt} \), decreases \( PC_A \) by the same amount
- \( T_G \) decreases as the truster’s commitment erodes
- \( TW_G \) increases as the trustee’s commitment erodes
- As the trustor’s commitment wanes, the trustee must spend more time proving themselves

\[
\begin{align*}
PC'_A &= PC_A - CL_{alt} \\
\Delta T_G &= -CL_{alt}/2OC_A \\
\Delta TW_G &= -OC_B/2CL_{alt} \\
\Delta \tau_B &= CL_{alt}/2JC_A
\end{align*}
\]
Betrayal!

- Understanding that one’s gain is the other’s lost
  (competitive mindset/understanding the world as zero-sum)
- Kelley and Thibaut proposed such a transformation with actors trying to maximize their relative gain.
- Gottman showed that when actors use such transformations, they begin to act irrationally.

\[
\begin{align*}
PC'_A &= PC_A - OC_B \\
OC'_A &= OC_A - PC_B \\
JC'_A &= JC_A - JC_B \\
PC'_B &= PC_B - OC_A \\
OC'_B &= OC_B - PC_A \\
JC'_B &= JC_B - JC_A
\end{align*}
\]

- As \( JC_A \to 0 \) and \( CL_{alt} \) increases, \( PC_A \to 0, PC'_A \to -OC_B \) & \( JC'_A \to -JC_B \)
- Get’s worse, if Actor A exploits his own power by increasing \( PC'_A \), decreasing \( OC_B \)
- Ultimately, leads to increasing non-correspondence, causing both actors to become more adversarial