Trust as a Game Mechanic

Chris Hazard, PhD

Hazardous Software Inc.
http://hazardoussoftware.com
Trust Us, We Know Trust!

• In econ, CS, psychology, etc.
  – No common definition
  – No common metrics
  – No common criteria or desiderata
• It's, you know, trust!
What Is Reputation?

• Belief that an attribute is a certain way
  – He's untouchable with sniper rifle
  – That game has boring grind

• Targets adverse selection
  – “r either of u a bot?”
    “no. i is speeked englishing.”
    “Look at my pic http://ishoponline.ru/b32tacr9”

• Hindsight, capabilities, signaling, statistics
What Is Trust?

• Econ: belief that another will not exploit
  – Moral hazard
  – “help me kill this giant sewer rat on the crate”
    KABOOM “pwnd noob. got ur loot lol”

• Authentication vs. exploitation
  – soft security
  – “I am Spartacus”

• Forward-looking, strategy, game theory
What Can We Do With This?

- Working hypothesis:
  Humans are actually rational*
  *given limited computational bounds, unfounded beliefs of others, inaccurate capability assessments, inexplicable valuations, and some level of [im]patience

- Valuations, capabilities, and patience can be measured! → reputation

- Patience core of strategy → trustworthiness
Patience (aka Intertemporal Discount Factor)

• Choices
  – $100 today vs $102 next week?
  – $100 today vs $10,000 next year?

• Qualitatively / intuitively related to trust
  – Psychology (e.g., Deutsch '73), Politics (e.g., Addison & Murshed '02), Economics (e.g., Whitmeyer '00)
  – E.g., Thieves → short-term gain for long-term risk & loss of trust (unless large pool of victims)
Discounting

• Uncertain future
  – Affect of delay on reward
  – Influenced by: patience, beliefs, risks, exogenous discount factors & value

• Expected utility =
  – Exponential, dynamically consistent: $\sum \gamma^t u$
  – Hyperbolic, realistic hazard rate: $\sum 1/(1+\gamma t) u$
Defining Trustworthiness

- **Scalar**
- **Strength**: Do something costly ⇒ will do something cheaper
- **Comparison**: Prefer b to c if b would fulfill more costly commitment than c
- **Stability**: Preferences stable if time shifted
Trustworthiness Isomorphic to Discount Factor

• Need valuations

• Compare two agents interacting with third in pure moral hazard situation

• Assumptions
  – Quasilinearity
  – Trustworthiness consistent enough
  – Individually rational

• All else equal, given definitions & assumptions, only factor that affects trustworthiness is discount factor
Measuring Discount Factors: Item Choice

- item 1: $q = 4$, $p = 1$
- item 2: $q = 50$, $p = 10$
- item 3: $q = 200$, $p = 20$
Creeping Sniper's Dilemma

- Mirror Shield
- Ghillie Hat
- Digital Camo
- Ninja Disguised As Tree

Original image from ShadowShield.com
Creeping Sniper's Dilemma

- Single sniper optimal strategy; slow creep out = low risk
  \[ \sigma_t = \frac{w}{(1 + \sqrt{1 - \gamma_{s_1}})} \left( \frac{1 - \sqrt{1 - \gamma_{s_1}}}{\gamma_{s_1}} \right) \]

- Multiple sniper optimal strategy
  - Match quickest visible discount strategy unless too risky
Incentivizing Players to Give Favors

• Rational player
  expected utility = expected benefit of relationship
  – expected cost of relationship
  – cost of favor

• Pure moral hazard

• How to sanction?
  – Tit-for-tat & variations
  – Ability to negate loss by reducing favors offered (derivatives about equal)
  – Figure out discount factor required to yield observed behavior
Combining Observations: Bayesian Inference

\[ f(\gamma) \]

\[ \Delta t = 0 \]
\[ \Delta t = 3 \]
\[ \Delta t = 7 \]
Optimal Level of Patience for Given Scenario

![Graph showing the relationship between total utility and discount factor for different scenarios. The graph includes markers for different discount factors (r=1, r=0.75, r=0.5, r=0.25) and distributions (W~U(950,1050), W~U(550,650), W~U(250,350)).]
What Do We Want Out of a Reputation System?

- **Unambiguous**
  - Player type yields one reputation
- **Monotonic**
  - Better reputation yields higher expected value of relationship
- **Convergent**
  - Reputation should converge quickly near fixed point
- **Accurate**
  - Reputation should converge quickly if large errors/biases exist
Reputation System Dynamics
How Can We Measure Trust in a Game?

• Determine utility, perceived probabilities, and risk aversion for major decisions
  – Model game interactions as “economy” w/ player time
  – Assess for all scenarios including tactical/optimizing (unilateral) and strategic (bi/multilateral)

• Compute decision thresholds based on relationship between player preferences and patience (discount factor)
  – Dynamic evaluation
  – Involve narrative engine, social network, seller ratings, other exogenous info
Trust as an Exploration Mechanic

• Quantitatively ensure “better” game
• Measure valuations & discounting distributions
  – Players' maxent regions
  – A priori playtesting
• In-game decisions:
  – Make sure level of trustworthiness required is below most users' trustworthiness
• Report user reputations
Trust as an Exploitation Mechanic

• Place player in edge situations
  – Ethical boundaries – “what is your price?”
  – ~3 choices good
  – Clear trade-offs

• Use appropriately
  – Player overload
  – “Soap Opera”
  – Players sometimes like stability, e.g., fixed alliances
Challenges to Discount Factor Approach

• Rationality of agents
• Computational complexity: Nash equilibria, combinations of actions, uncertainty
• Disagreements over definition of trust (to include capabilities, reliability?)
• No “ideal” intertemporal discount model in most situations
What Enables Trust Psychologically?

Homophily

Embedding

Corroboration

Image from upcoming Mass Effect 3

Image from WoW Cataclysm

Image from Heavenly Sword
Core Corroboration Caveat

• Provenance hard to assess!

• Trust transitivity & dampening

Hang, Wang, & Singh AAMAS '09
What Trust is NOT

- Social norms & laws
  - Perceived value of relationships
  - Perceived impact of norms on relationships
- Measure of good vs evil
  - Trustworthy henchman & maverick hero
- Keeper of secrets
  - Strong relation, but not necessitated
- Reliability
- Intimacy
Where Isn't Trust the Main Principle?

• MMO Raid
  – Group cohesiveness incentivized by weakness of being alone
  – Mutual dependence
    • Trust “ratcheting”
    • Not much trust but seems like it

• Self-interest vs malice
  – Malice: easier to model in full-information games, harder in partial-information games
  – Faux altruism: strategic relationship building
Trust & Society

- Enforcing/sanctioning often only tools to combat lies
  - Information asymmetry
  - When possible: incentive compatibility & revelation principle
  - Level of trust often req'd for system & market efficiency

- Too trusting with homophily, embedding, corroboration?
  - Trust researchers & high-ranking military leaders
  - Inability to play red in red v blue
Conclusions

- Trust important as games ascend further into social space
- Trust important in narrative
- Can model and compute trust metrics
- Use for game design
Questions?