Measuring and Manipulating Player Trust through Choice and Game Mechanics

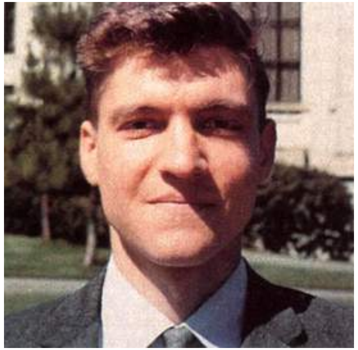
Christopher J. Hazard, PhD
CEO, Hazardous Software Inc. / CTO, Szl.it

ARTIFICIAL INTELLIGENCE SUMMIT
GAME DEVELOPERS CONFERENCE®
MOSCONE CENTER · SAN FRANCISCO, CA
MARCH 2-6, 2015 · EXPO: MARCH 4-6, 2015


COAL PORTER
SHOVEL BOY
GRAVELLER
COAL BEARER
TANK CLEANER
MATCH DIPPER
CROSSING SWEEP
SEWER BRICKER
PIT MONKEY

from zap2it.com

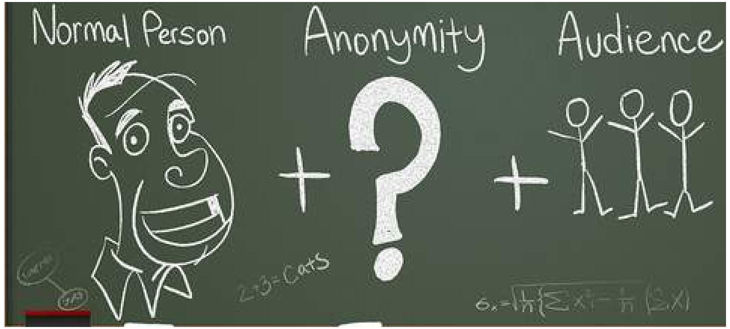
from Seattle Weekly



from trutv.com

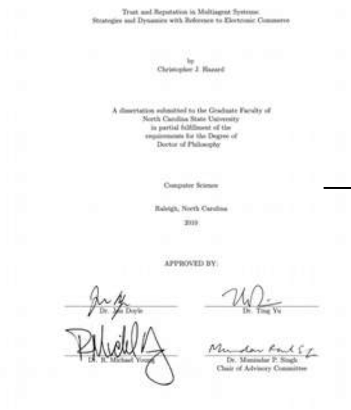
by tinyfroglet, cc

from supermanhomepage.com



from penny-arcade.com

~~From~~ whence the results came...



Trust and Reputation in Multiagent Systems:
Strategies and Dynamics with Reference to Electronic Commerce

by
Christopher J. Hazard

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Computer Science

Raleigh, North Carolina

[illegible]

APPROVED BY:

Dr. Jon Doyle

Dr. Ting Yu

Dr. R. Michael Young

Dr. Munindar P. Singh
Chair of Advisory Committee

This Box Intentionally Left Blank

This Talk

Reputation

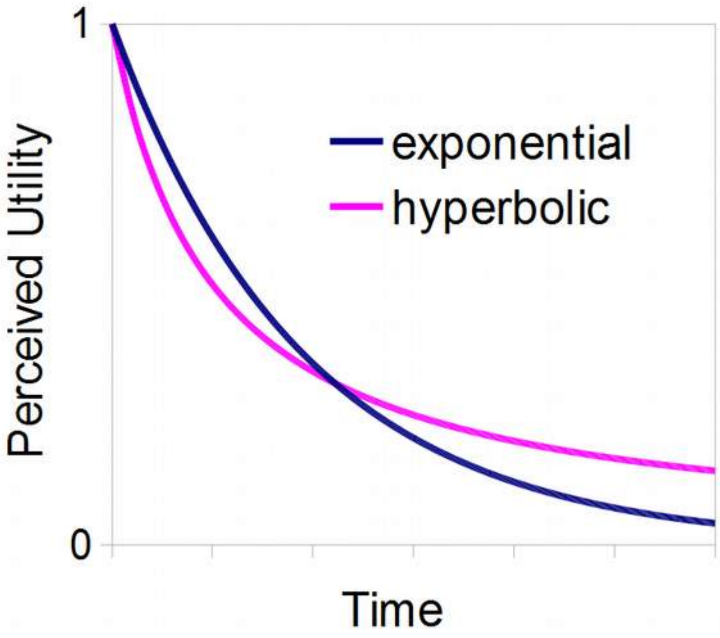
- Belief about attribute
- Hindsight, capabilities, statistics
- Concern: adverse selection

Trust

- Belief will not exploit
- Foresight, strategy, game theory
- Concern: moral hazard

- Humans are 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
- Model new situations → trustworthiness

Discounting



- Uncertain future
  - 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: $\sum 1/(1+\gamma t)\, u$

Discounting Everywhere

- Stochastic search

$$\pi(a) = \frac{\exp\left(\frac{Q(a)}{n(a)}/\tau\right)}{\sum_{b \in A(s)} \exp\left(\frac{Q(b)}{n(b)}/\tau\right)}$$

- Amortization

$$\mathrm{NPV}(i, N) = \sum_{t=0}^{N} \frac{R_t}{(1+i)^t}$$

- Bellman Equation

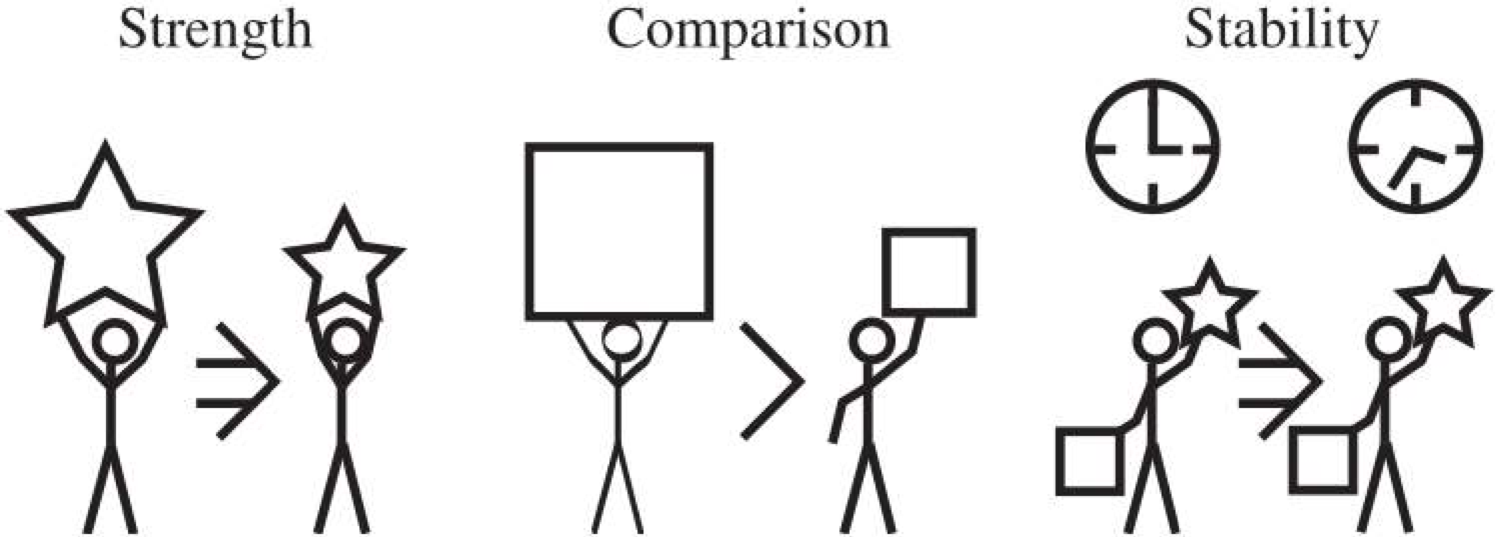
$$V(x_0) = \max_{\{a_t\}_{t=0}^{\infty}} \sum_{t=0}^{\infty} \beta^t F(x_t, a_t),$$

- Reinforcement Learning
- Markov Decision Processes & POMDPs

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_{t+1} + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)]$$

- ***Normalize discount rate wrt time

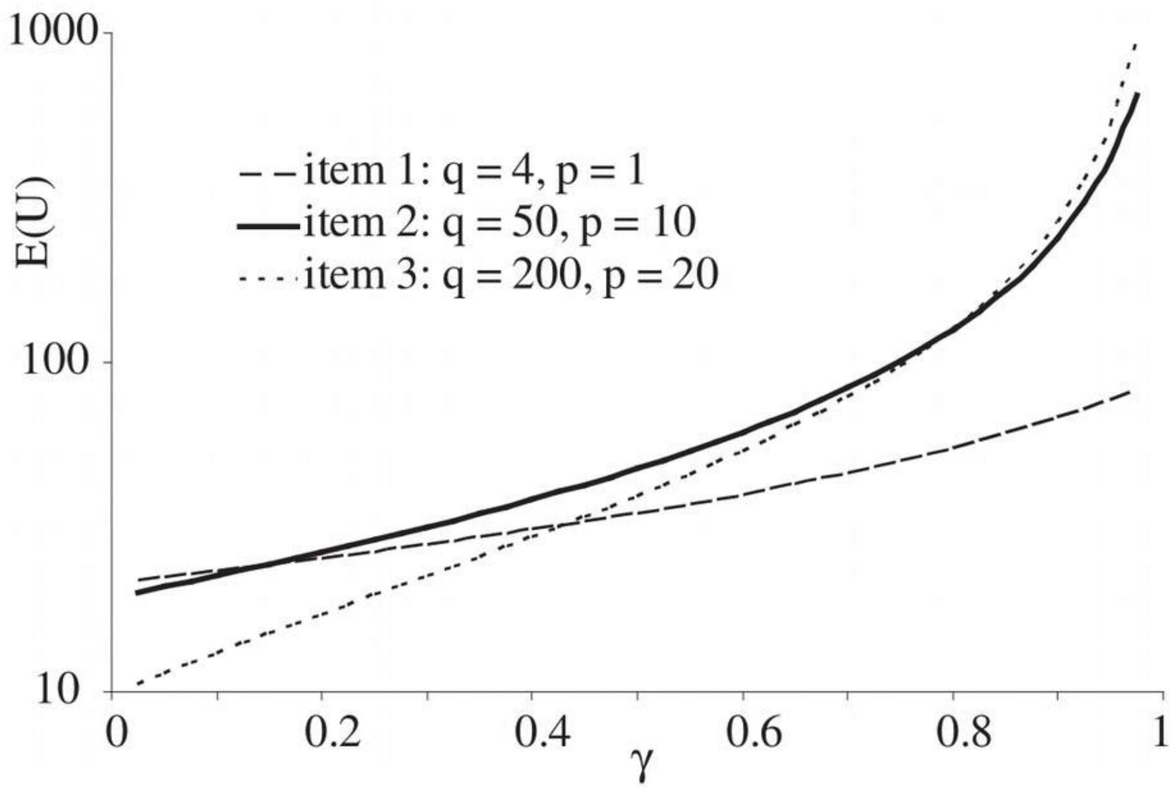
Defining Comparable Trustworthiness



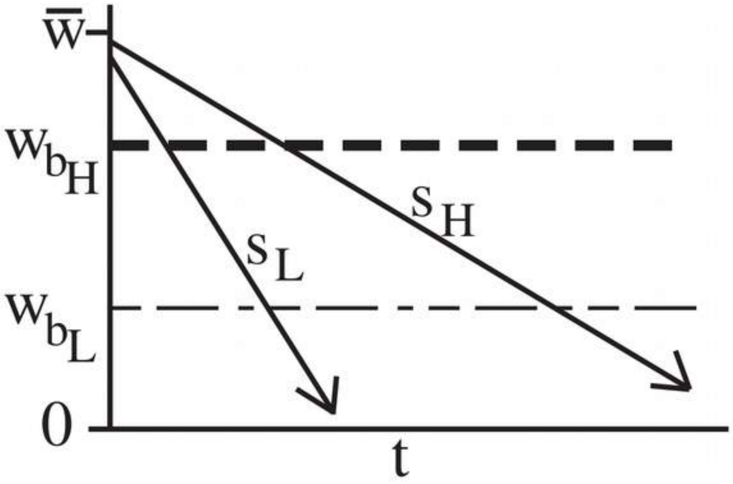
Trustworthiness Isomorphic to Discount Factor

- Compare two agents interacting with third in pure moral hazard situation
- Assumptions
  - Consistent valuations
  - Quasilinearity
  - Trustworthiness sufficiently consistent
  - Individually rational
- All else equal, given definitions & assumptions, only factor that affects trustworthiness is discount factor

Measuring discount factor by choice



Creeping Sniper's Dilemma



- Single sniper optimal strategy; slow creep out = low risk

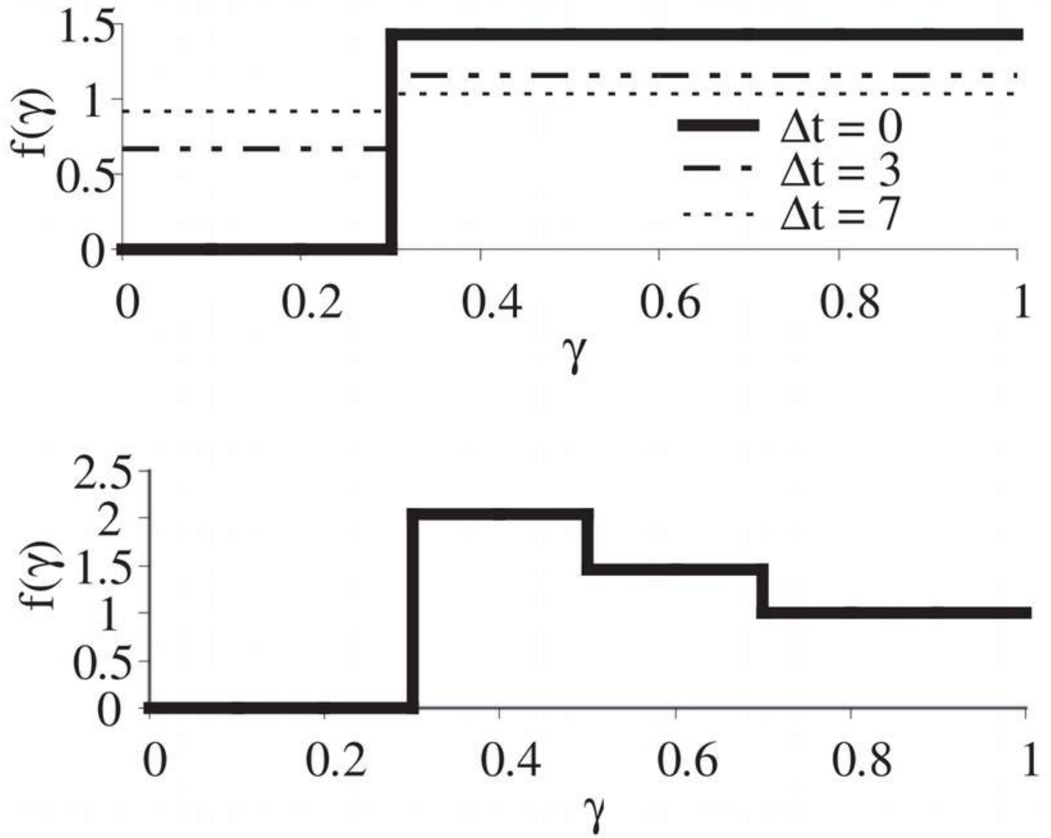
$$\sigma_t = \frac{\overline{w}}{(1+\sqrt{1-\gamma_{s_1}})}\left(\frac{1-\sqrt{1-\gamma_{s_1}}}{\gamma_{s_1}}\right)^t$$

- Multiple sniper optimal strategy
  - Match quickest visible discount strategy unless too risky

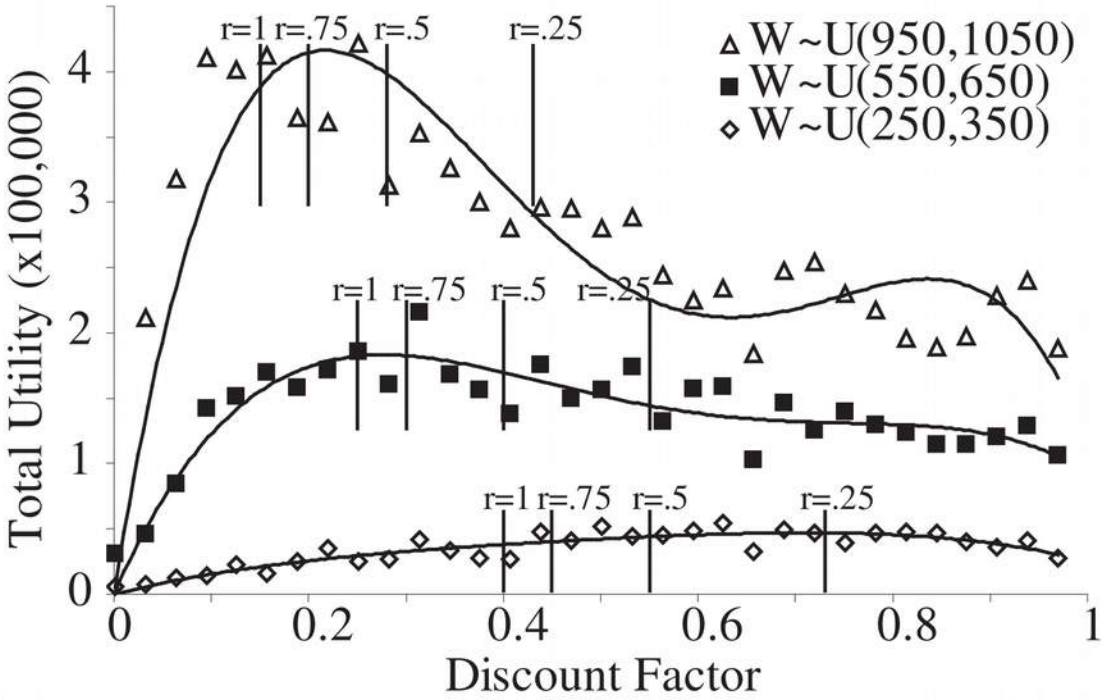
Negotiating

- Rubenstein Negotiation
  - $v_1 = (1-\gamma_2)/(1-\gamma_1\gamma_2)$
  - Inequalities if rationality not guaranteed
  - Player & NPC interaction inequalities
- Impatience $\frac{w_b - \sigma_{T-1}}{w_b - \sigma_T} < \gamma_b \leq \frac{w_b - \sigma_T}{w_b - \sigma_{T+1}}$
- NPC disagreements with player over choices

Combining Observations: Bayesian Inference



Optimal Level of Patience for Given Scenario



Trust Exploration

- Measure valuations, discount factor, beliefs, maxent regions
- NPCs of different trustworthiness
- Reputations

Trust Exploitation

- Push player's ethics buttons: “what is your price?”
- Stability & comfort vs conflict
- Trickery

Psychological Heuristics of Trust

Homophily

Image from WoW Cataclysm

Mass Effect 3

Embedding

Corroboration

Image from Heavenly Sword

Trust & Society

- Enforcing/sanctioning to combat lies
  - Incentive compatibility & revelation principle wrt information asymmetry
  - Level of trust req'd for system & efficiency
- Too trusting with homophily, embedding, corroboration?
  - Common inability to play “red player”

Direct Applications (Conclusions)

- NPC decisions: favors, purchases, alliances
- Measuring player patience
- Adversary willingness to look ahead related to organizational trust (e.g., big bad)
- NPC subordinates following player commands based on trustworthiness (explicit or implicit)

For further info

hazardoussoftware.com
cjhazard@hazardoussoftware.com