Game theoretic models of moral behaviour

Sarita Rosenstock

Game Theory is like decision theory but PVP

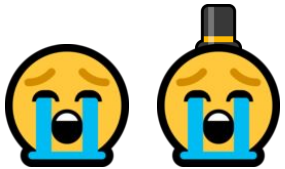
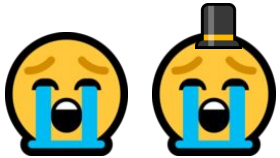
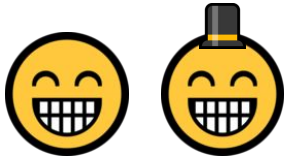
Exp(☂️) = P(☀️)*U(😒) + P(🌧️)*U(🙂)

Examples

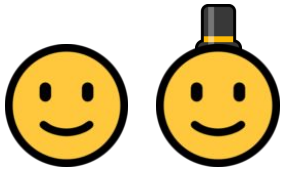
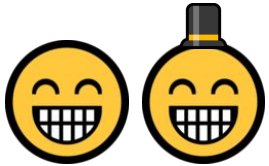
Prisoner’s Dilemma

	Alban (🎩) cooperates	Alban defects
I cooperate	😐 🎩😐	🎩😭 🎩😁
I defect	😁 🎩😭	😟 🎩😟

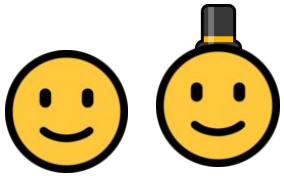
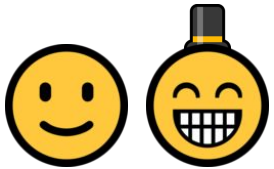
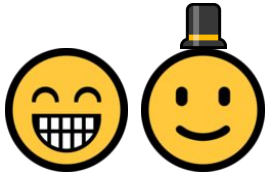
Pure Coordination

	Alban goes right	Alban goes left
I go right		
I go left		

Stag Hunt

	Alban hunts rabbit	Alban hunts stag
I hunt rabbit		
I hunt stag		

Hawk-Dove (AKA Chicken)

	Alban is deferential	Alban is aggressive
I'm deferential		
I'm aggressive		

Analysing Games

Characterising Strategy pairs

A pair of strategies is a Nash equilibrium if neither player can benefit by *unilaterally* changing strategy.

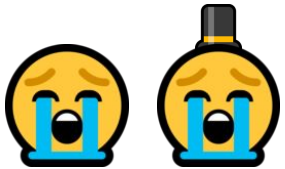
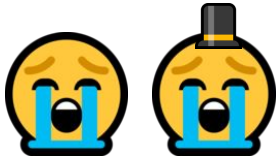
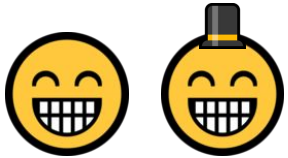
Other properties of strategy pairs:

- Total utility
- “Fairness”

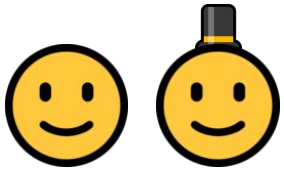
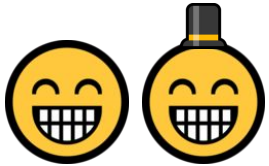
Prisoner’s Dilemma

	Alban (🎩) cooperates	Alban defects
I cooperate	😐 🎩😐	🎩😭 🎩😁
I defect	😁 🎩😭	😟 🎩😟

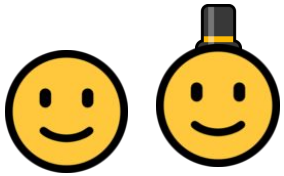
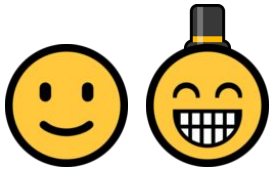
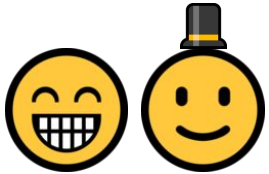
Pure Coordination

	Alban goes right	Alban goes left
I go right		
I go left		

Stag Hunt

	Alban hunts rabbit	Alban hunts stag
I hunt rabbit		
I hunt stag		

Hawk-Dove (AKA Chicken)

	Alban is deferential	Alban is aggressive
I'm deferential		
I'm aggressive		

Irrational Behaviour?

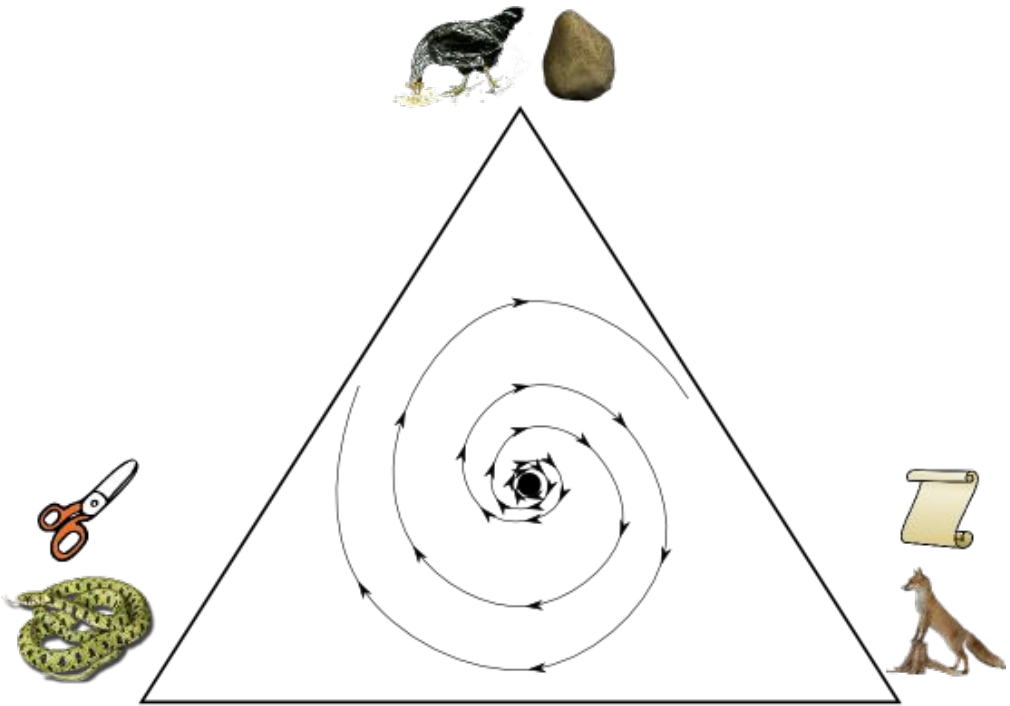
The Ultimatum Game

Player 1: choose how to divide the pie.

Player 2: accept or reject offer.

Evolving Strategies

Adding Dynamical Evolution



Evolutionarily Stable Strategy (ESS)

A pair of strategies (A, B) is ***evolutionarily stable*** if

1. Exp(A, A) > Exp(B, A), **or**
2. Exp(A, A) = Exp(A, B) and E(A,B) > E(B,B)

Idea: ESSs are robust against invasion by an alternate strategy.

Replicator Dynamics

State = (portion of pop playing strategy S_1, S_2, …)

Fitness

$$f_i(x_1, \ldots, x_i, \ldots) = \sum_j (\text{prob of interacting with } j\text{-player}) * U_i(i, j)$$

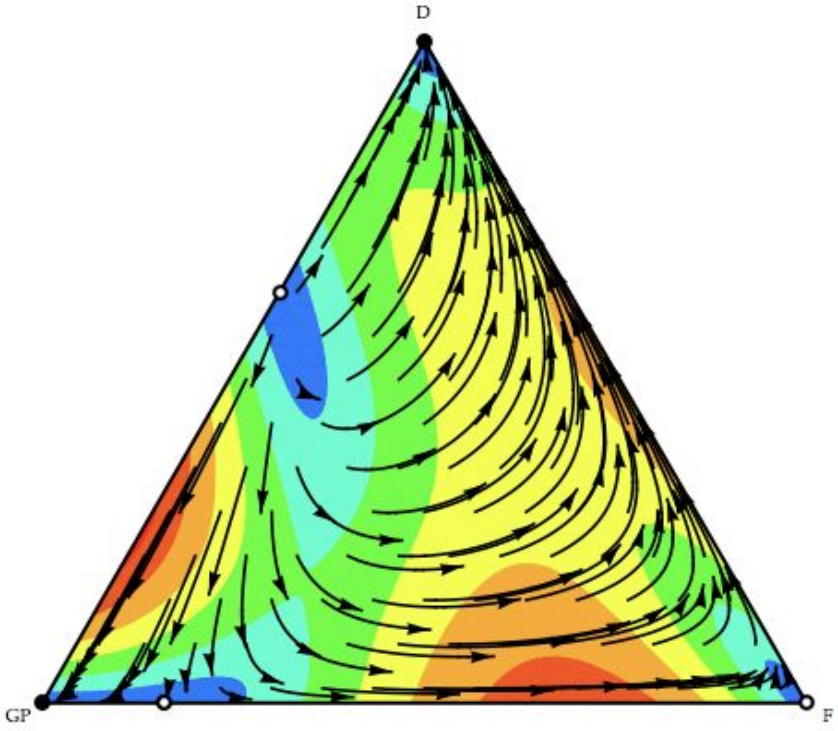
Average fitness

$$F = \sum f_i(\mathbf{x}) * x_i$$

Dynamical Equation

$$\dot{x}_i = x_i * (f_i(\mathbf{x}) - F)$$

Basins of Attraction



How can ethical behaviour evolve?

Iterated Prisoner’s Dilemma

Parameters

n = chance of repeat encounter

ε = probability of error

c = cost of apology

p = chance apology is believed

Strategies

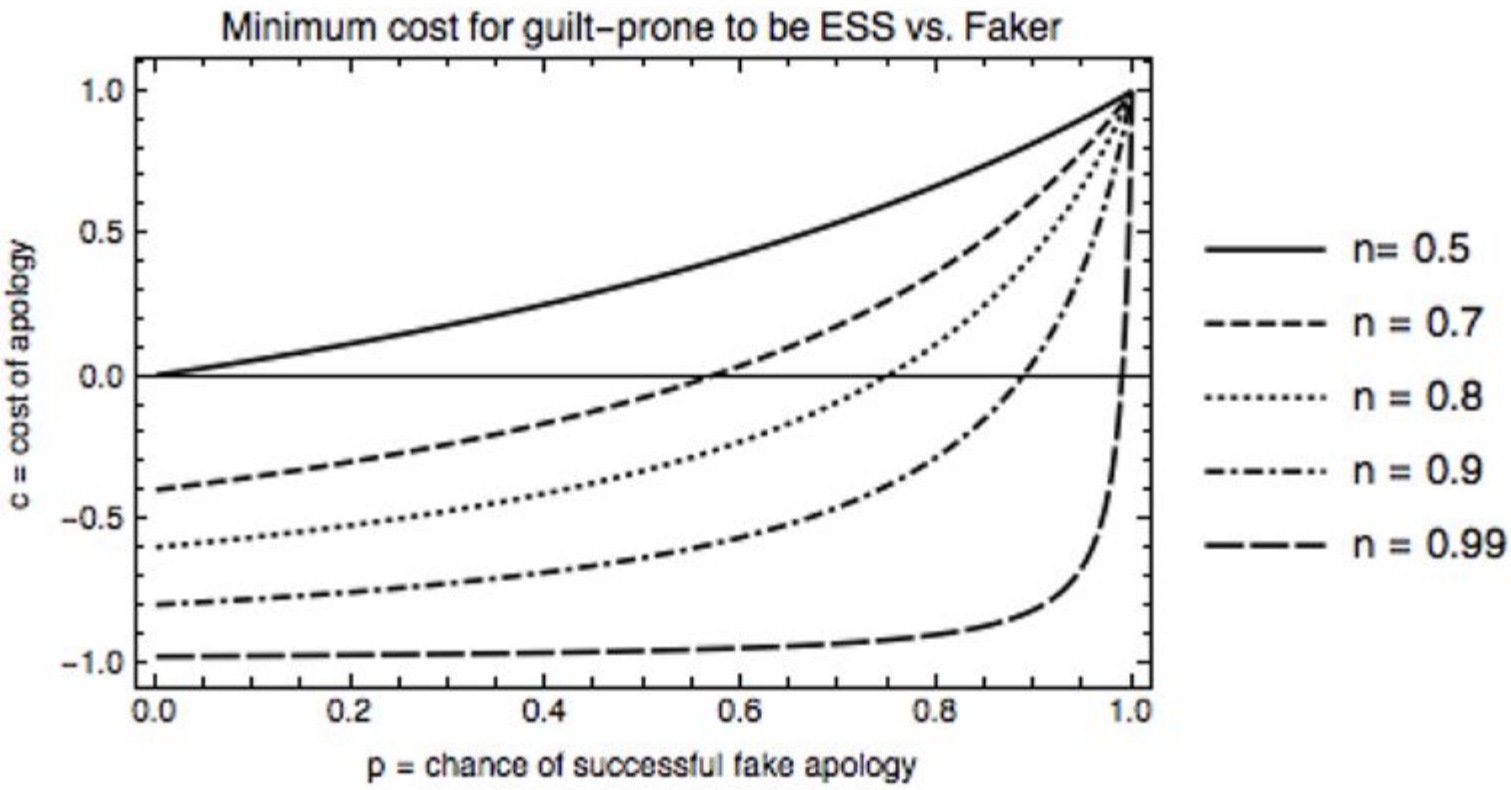
Always Cooperate

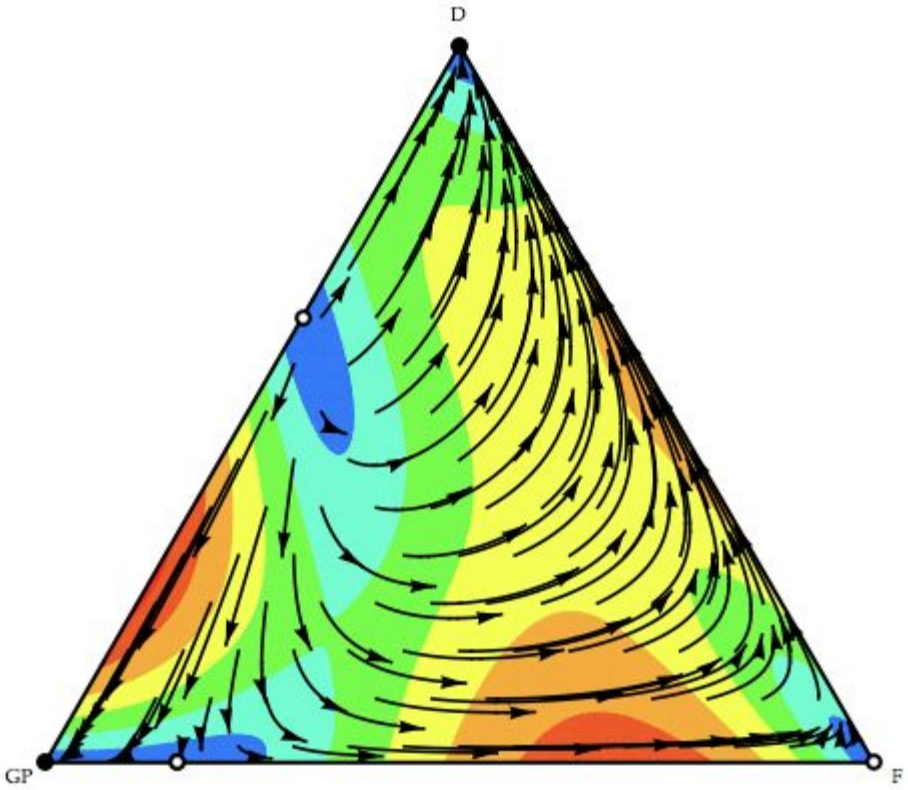
Always Defect

(Guilt Prone) Grim Trigger

(Guilt Prone) Tit-for-tat

Faker (defector who apologises)


Minimum cost for guilt−prone to be ESS vs. Faker
1.0
0.5
0.0
−0.5
−1.0
c = cost of apology
0.0
0.2
0.4
0.6
0.8
1.0
p = chance of successful fake apology
n= 0.5
n = 0.7
n = 0.8
n = 0.9
n = 0.99

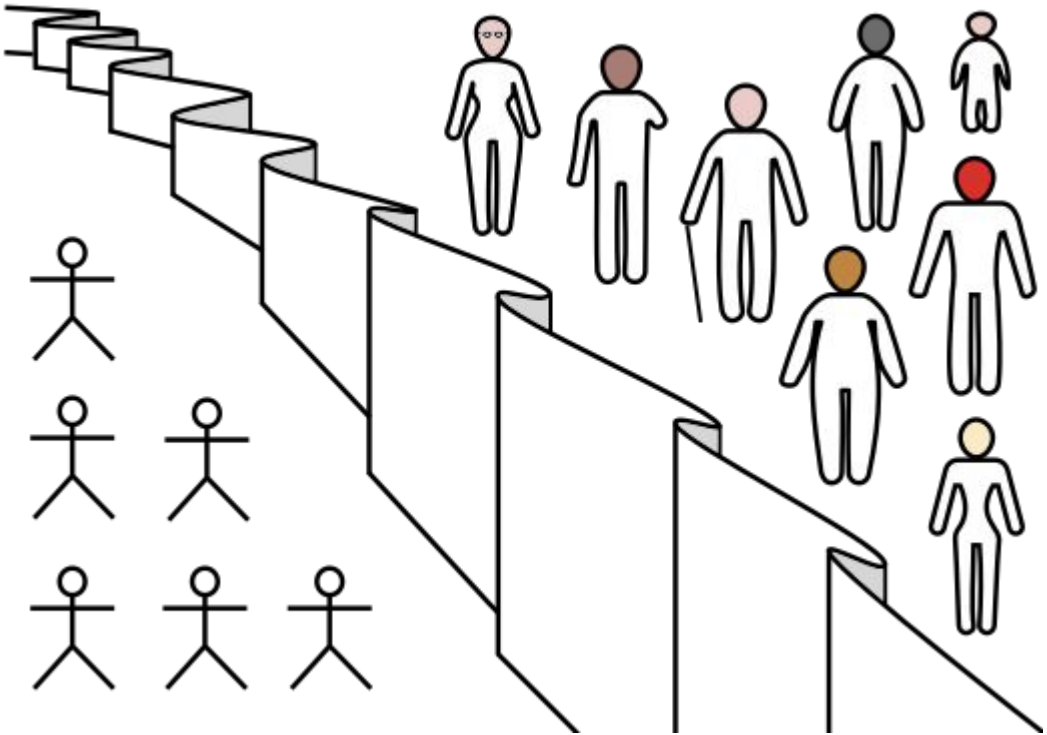

D
GP
F

What if you actually care about ethics?

Change the Utility Function

Design games to maximise ethical objectives

Veil of Ignorance



How does game theory relate to ethics?

- Functionalism
  - Moral norms can be adopted intentionally to push us towards more cooperative, mutually beneficial strategies and avoid pitfalls of rational self-interest
- Bargaining Theory & Contractarianism
  - Game theory helps us model a bargaining process for the fair aggregation of preferences
- Recovery
  - EGT shows how moral norms can naturally evolve as heuristic solutions to collective action problems

Questions?

Combinatorics Question

How many “qualitatively different” 2x2 games are there?

Bonus Material

https://ncase.me/trust/