Containing multitudes:

moral enhancement, game theory and the stability of society

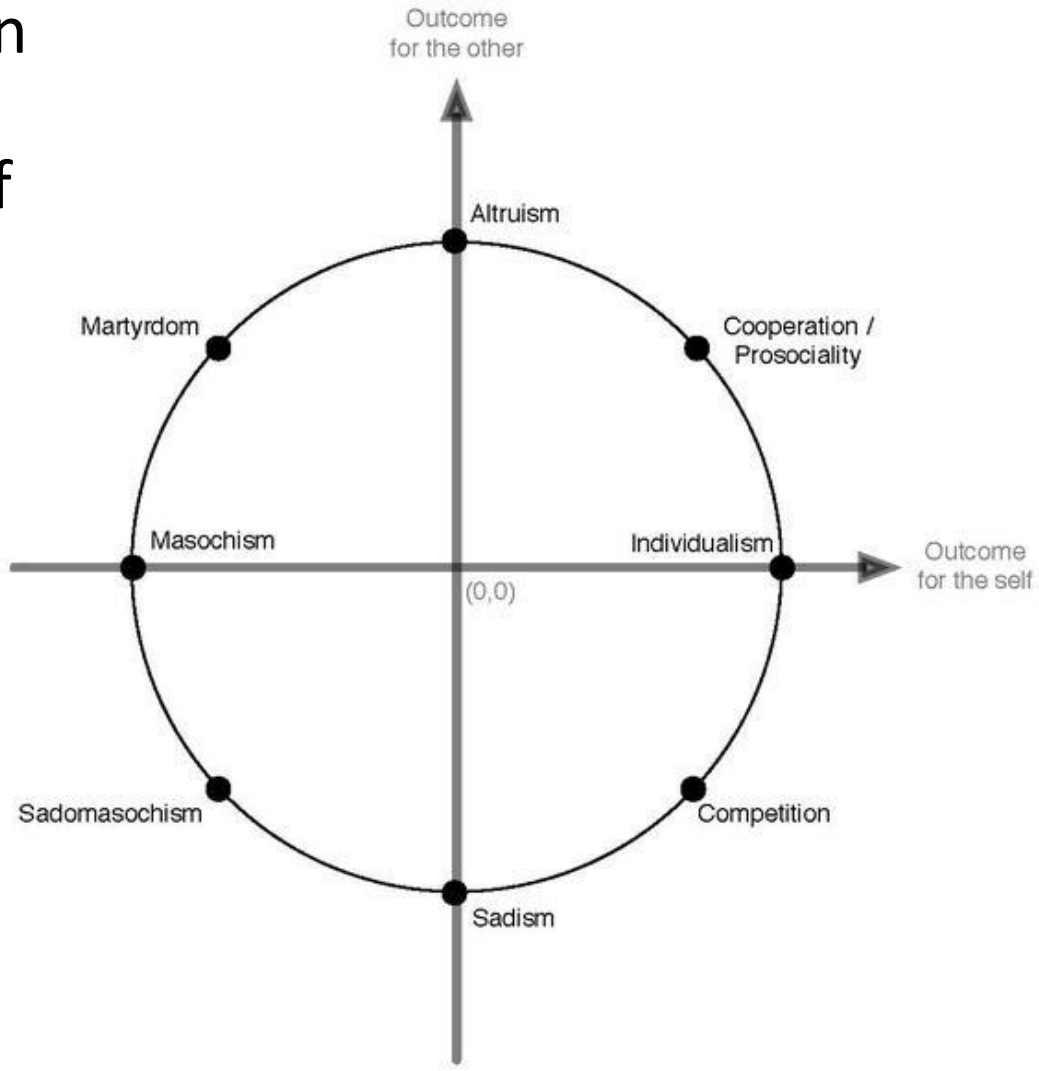
Dr Anders Sandberg
João Lourenço de Araujo Fabiano
Oxford Uehiro Centre for Practical Ethics
University of Oxford

”Why can’t we all just get along?”

- Biomedical moral enhancement: because exhortations and policing may not be enough
- Adjust evolved abilities and orientation to be more in line with current world
- But if people are allowed to freely choose what they want, will this be stable?
- ABM of society where agents update their cooperation preferences

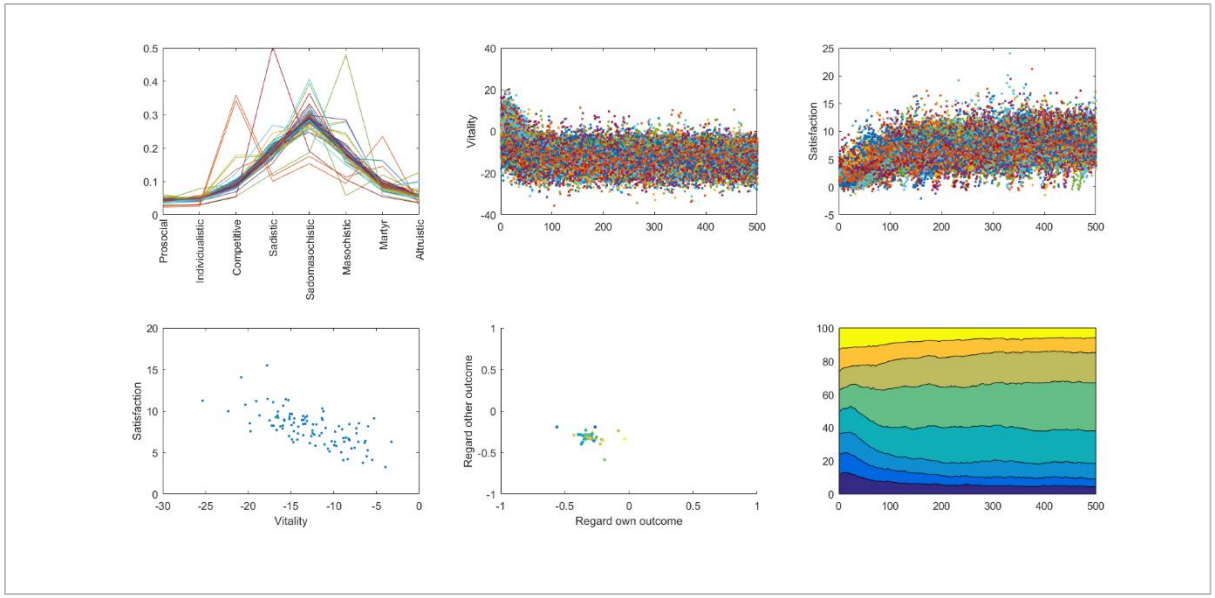
Social value orientation

- The satisfaction with an outcome depends on own and other's payoff
- Prosocial (+/+)
  Individualistic (+/0)
  Competitive (+/-)
  Sadistic (0/-)
  Sadomasochistic (-/-)
  Masochistic (-/0)
  Martyr (-/+)
  Altruistic (0/+)
  MaxDiff
  MinDiff

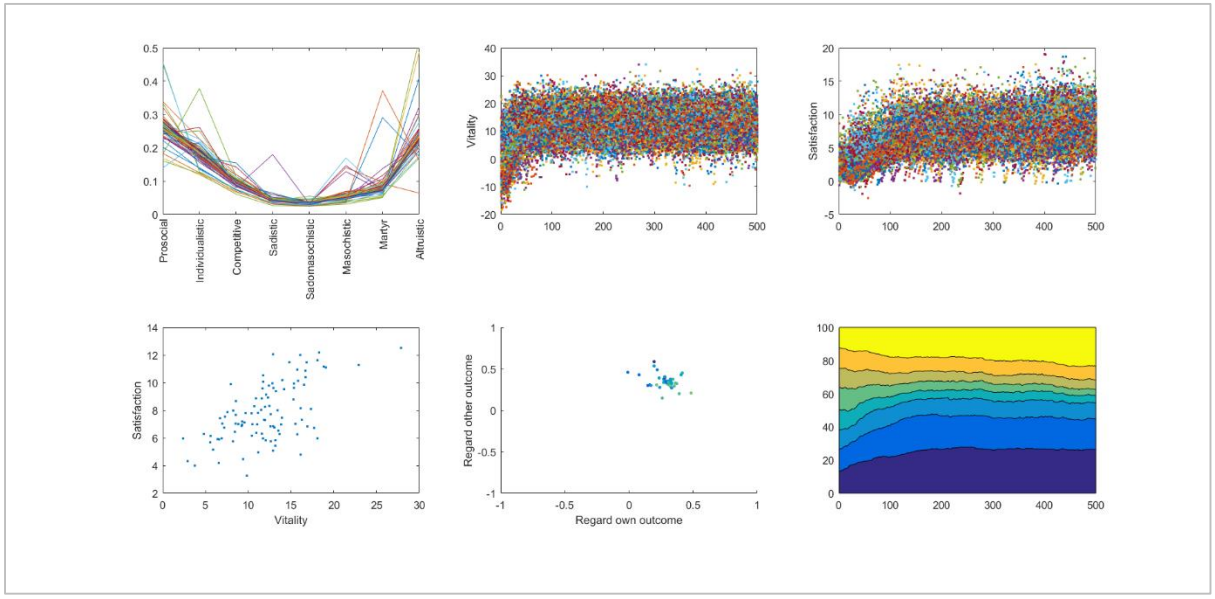


Model

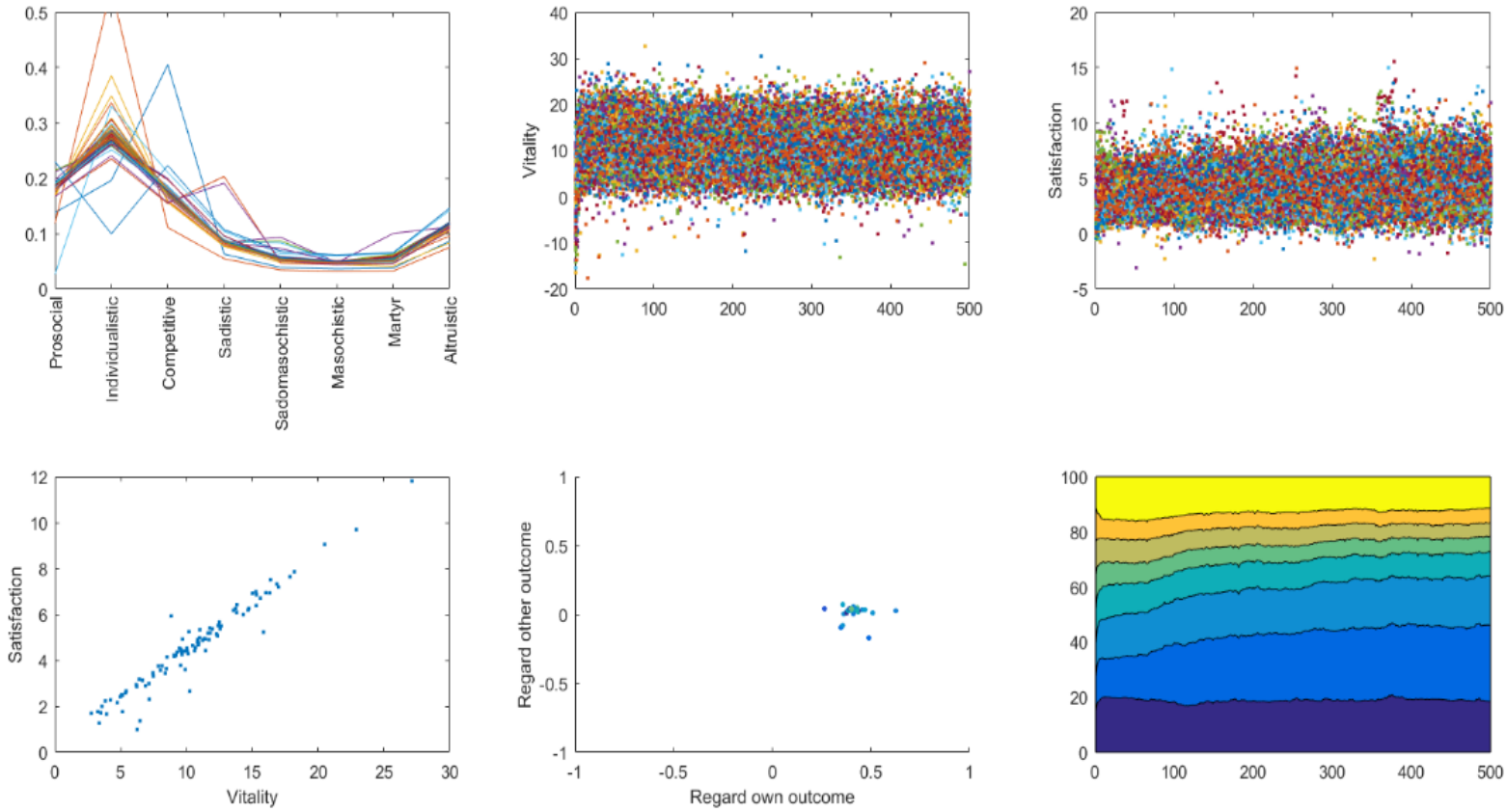
- Agents have normalized weight vectors of SVOs
- **Social success**: Each epoch they play 10 games against randomly selected other agents
  - Either random 2x2 games, PD or other dilemma games
  - Mixed strategy Nash equilibrium calculated using Lemke-Howson algorithm
  - Objective payoffs ("vitality") are converted to subjective payoffs ("satisfaction") based on agent's weights.
- **Moral enhancement**: At end of epoch agents compare satisfaction to another random agent. If they are less satisfied than other, they will change weights in their direction
  - Also mutation probability
  - Select unequally: "celebrity", select based on similarity to self
- Agents below vitality "poverty" threshold removed and replaced with copied agents.



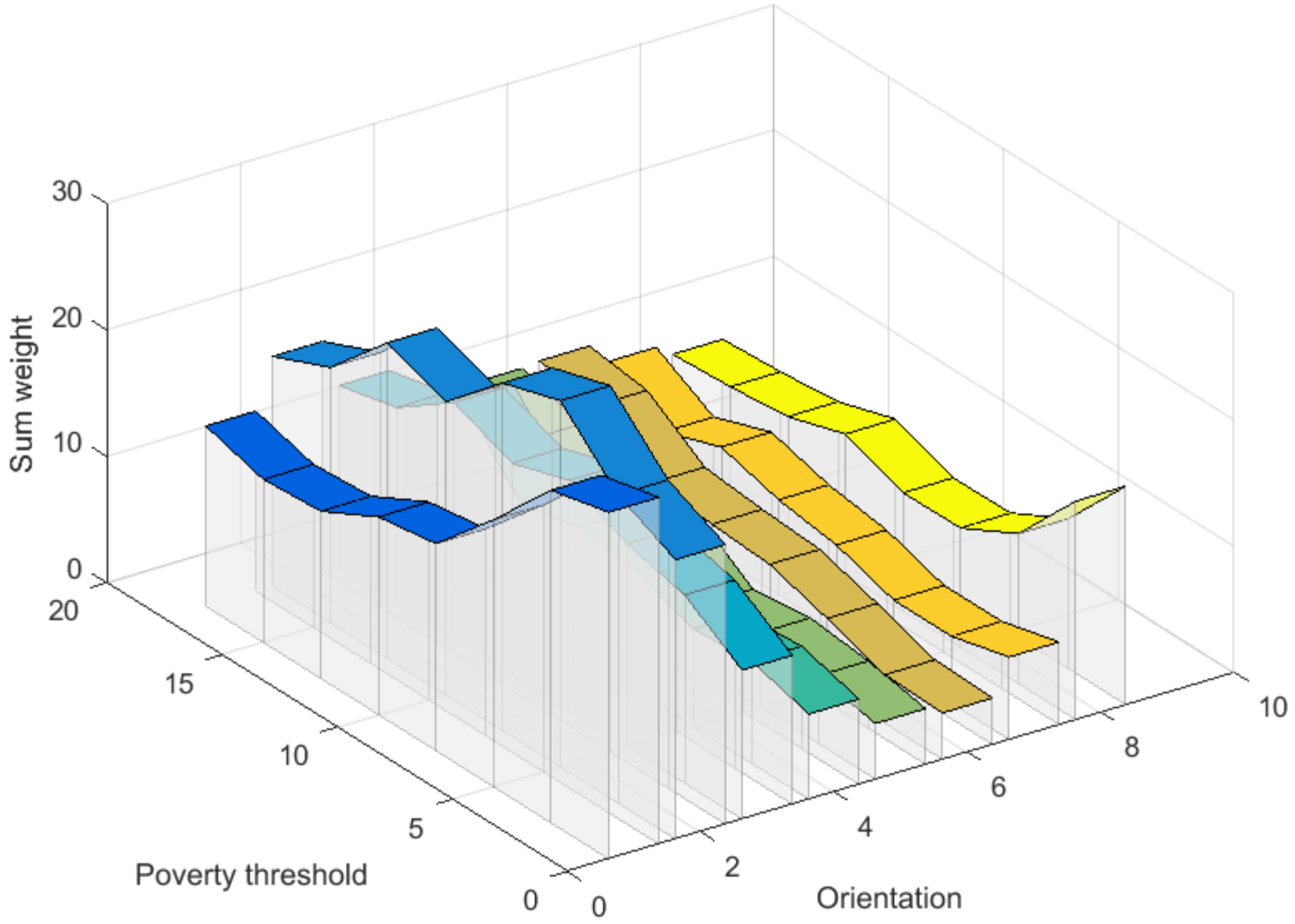
End state in the "nasty" attractor state after 500 epochs. Upper left: plot of weightings for the 100 individuals. Upper middle: time evolution of vitality scores. Upper right: time evolution of satisfaction scores. Lower left: vitality and satisfaction at time 500. Lower middle: population plotted by average weighting towards self, other (color denotes satisfaction). Lower right: total weights for the different strategies across time.



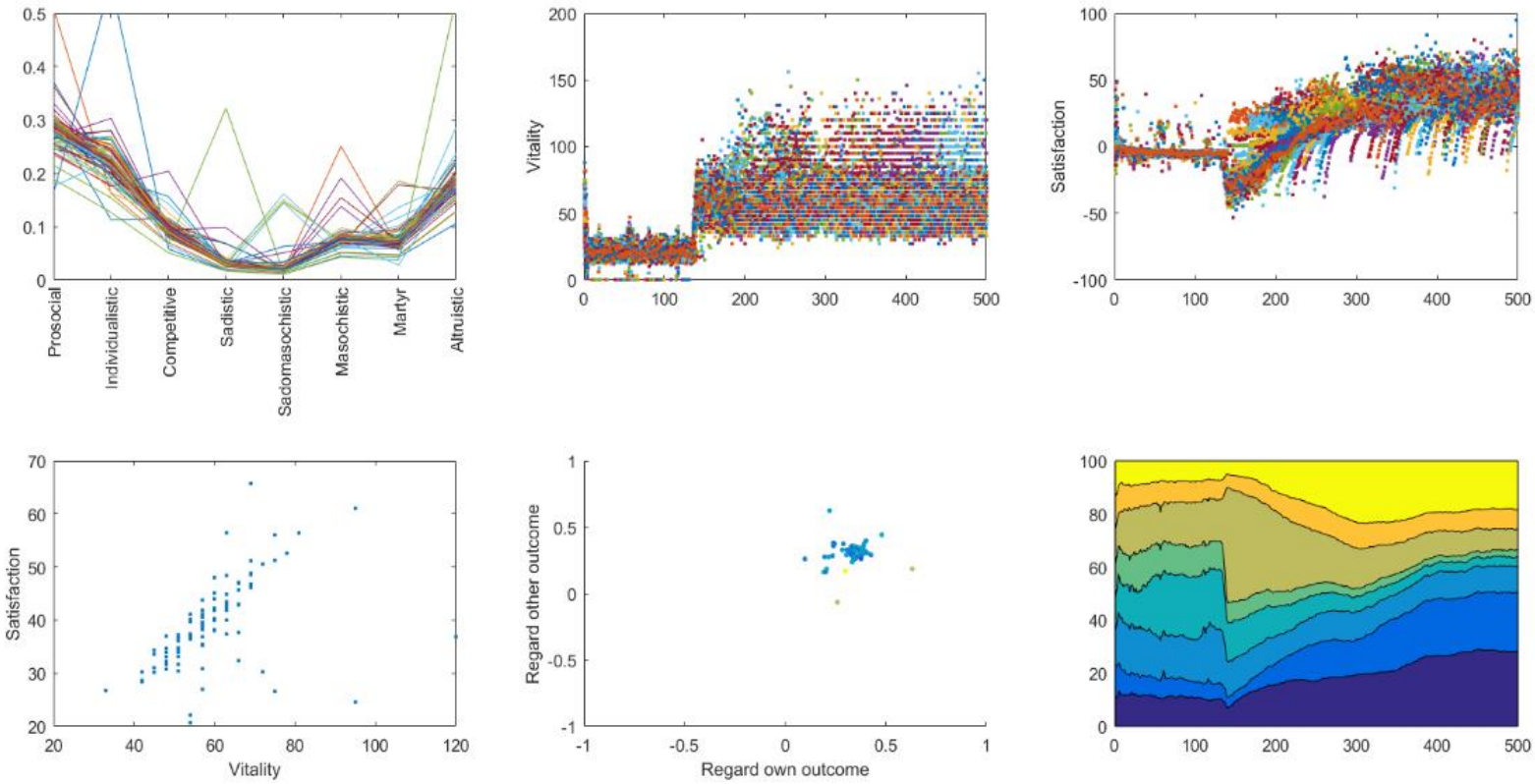
End state in "nice" attractor.



Evolution of game with poverty threshold=5.



Total weight of different orientations in the population as a function of poverty threshold. Average of 10 simulations.



Evolution of prisoners' dilemma game.
Poverty threshold 5.

Outlook

- Merely adding moral enhancement does not destabilize things
  - as long as there is a vitality threshold the prosocial/individualistic attractor wins
  - May be issue with balance Individualist (+/0)/Prosocial(+/+) SVOs
- Satisfaction can diminish as a result of individual satisfaction-seeking
- To investigate:
  - altruistic punishment
  - cooperation within/between groups