NACS 645 – Why do we cooperate?

-

Valentin Guigon

DEPARTMENT OF PSYCHOLOGY

SLD Lab

PROGRAM IN NEUROSCIENCE & COGNITIVE SCIENCE

Cooperation, competition, coordination

Prisoners vs civilians, rigged boats, *The Dark Knight*

Mexican standoff, *Reservoir dogs*

	C	*D*
C	-1,-1	-1,0
D	0,-1	-1,-1

	C	*D*
C	1,1	-1,0
D	0,-1	-1,-1

Cooperation, competition, coordination

Prisoners' dilemma

P1	C	D
C	-1,-1	-4,0
D	0,-4	-3,-3

No matter what P2 plays, it is best to defect (1 Nash Equilibrium)

Mixed motive: cooperation & competition

Side of the road

	Left	Right
Left	1,1	0,0
Right	0,0	1,1

If P2 goes to left, P1 better go left; same for the right (2 NE)

Pure coordination: exactly aligned interests

Battle of the sexes

P1	B	F
B	2,1	0,0
F	0,0	1,2

The best response is to select the action that the other wants (2 NE)

Mixed motive: cooperation & competition

Matching pennies

	Heads	Tails
Heads	1,-1	-1,1
Tails	-1,1	1,-1

P1 ends up in a circle (0 NE)

Pure competition: exactly opposed interests

Cooperation, competition, coordination

Prisoners' dilemma

P1		C	D
	C	-1,-1	-4,0
	D	0,-4	-3,-3

No matter what P2 plays, it is best to defect (1 Nash Equilibrium)

Mixed motive: cooperation & competition

Battle of the sexes

P1		B	F
	B	2,1	0,0
	F	0,0	1,2

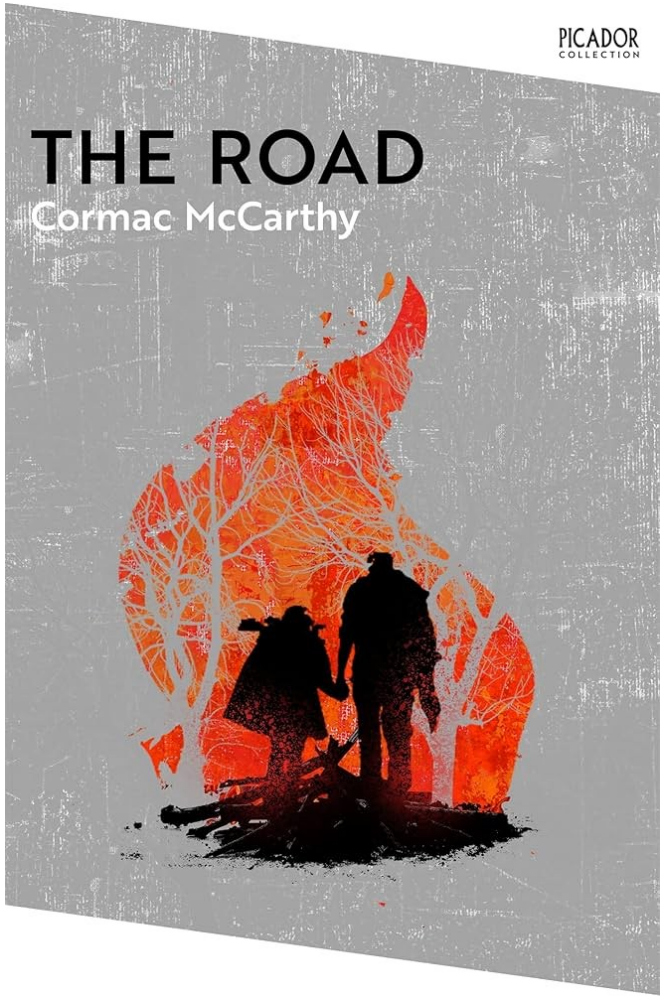
The best response is to select the action that the other wants (2 NE)

Mixed motive: cooperation & competition

- Cooperation happens in situations with conflicting interests of actors: actors opt for an action suboptimal for themselves but superior for the collective.
- The population does best if individuals cooperate (*higher social welfare*), but for each individual there is a temptation to defect (*higher individual utility*).
- One individual pays a cost for another to receive a benefit.
- Problem in the PD: why should you reduce your own fitness to increase that of a competitor in the struggle for survival?

Choosing actions

- What will other players do?
- What should I do in response?



Classical rationality

Homo economicus

- Rational agents
- Perfectly informed
- Maximize their utility

Rational choice theory

- Coherent (ordered) preferences
- Rational thinking leads to choices aligned with preferences
- Group behavior reflect the aggregate of individual behaviors (efficient allocation)

Mill, Pareto, Samuelson, etc.

Game theory

Assumptions:

- Each agent has its own description of states of the world
- Each agent has a utility function (preferences, uncertainty profile)
- Each agent maximizes expected utility (decision-theoretic rationality)

Ingredients to describe decision-making:

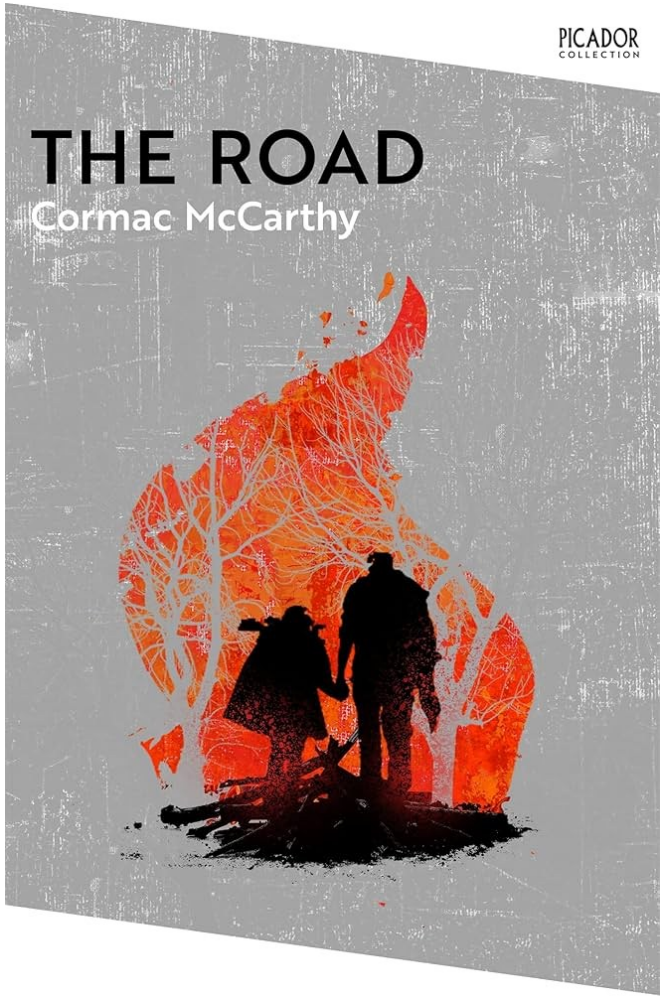
- Set of Players (people, governments, companies)
- Set of Actions (bid, strike, vote, cooperate, defect)
- Set of Payoffs (monetary preferences, social preferences)

Choosing actions

- What will other players do?
- What should I do in response?

Given prior assumptions (each agent has its own description of states of the world, a utility function and aims at maximizing expected utility):

Each player ***best responds*** to the others (*Nash equilibrium*)



Nash equilibrium

Side of the road

	Left	*Right*
Left	1,1	0,0
Right	0,0	1,1

An equilibrium:

- A list of actions (action set)
- With which each player's action maximizes his/her payoff given the actions of the others
- In a consistent/stable pattern (profile)

Implications:

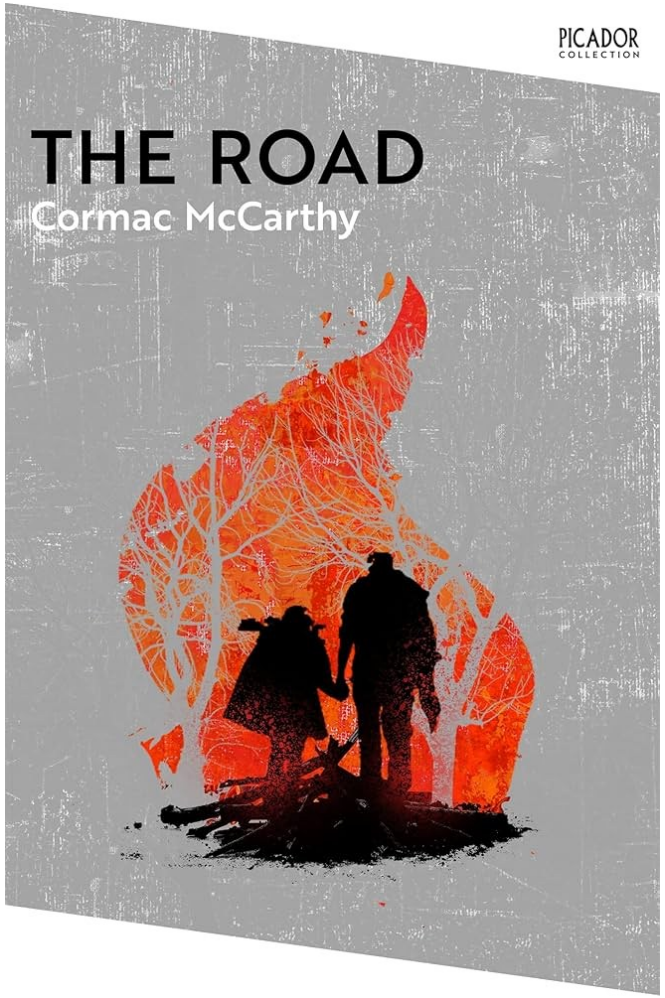
- Nobody has an incentive to *deviate* from their action if an equilibrium profile is played
- Someone has an incentive to *deviate* from a profile of actions that do *not* form an equilibrium

Under correct assumptions, equilibria should be expected to be played once participants understand the game (non-equilibria should vanish over time)

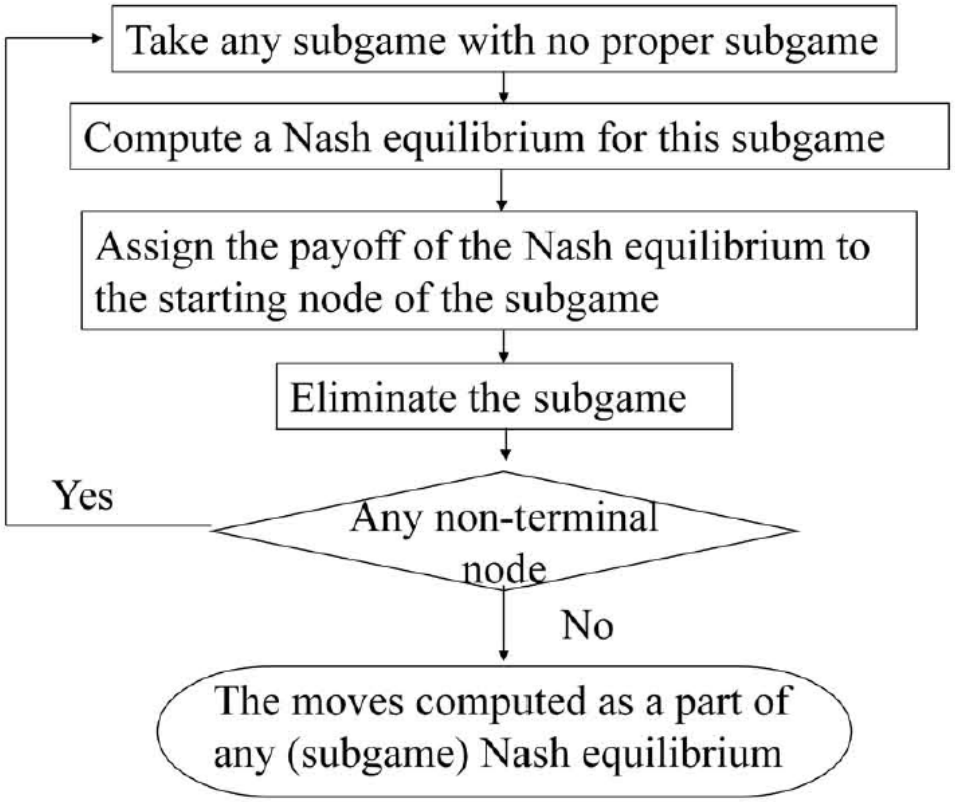
Choosing the best action

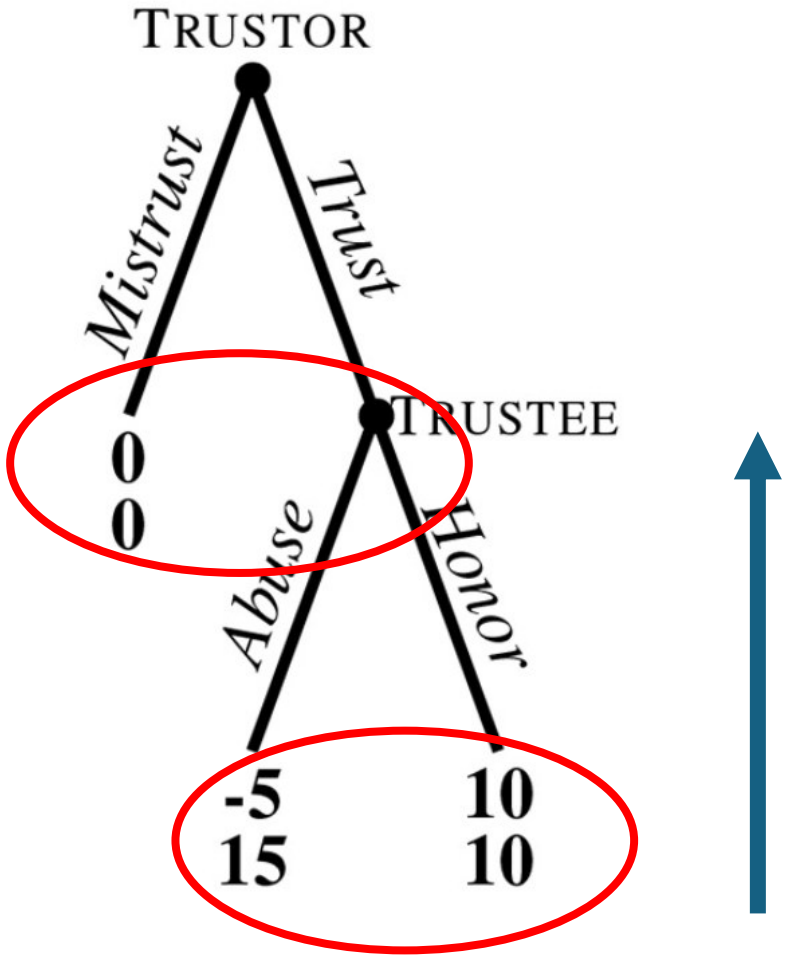
- If we knew what everyone else was going to do, it would be easy to pick the action
- Idea: look for stable action profiles (actions that nobody has incentives to deviate from)
- The « pure strategy » Nash equilibrium is a set of actions, one for each agent, such that each action is the best response to the actions of others

- In *The road*, the world has ended and cannibalism is widespread. We can assume the best action is the *untrustworthy* one
- In the *State of nature*, we may need others and we want to avoid risks. We can assume the best action is the *trustworthy* one



Choosing actions – backward induction

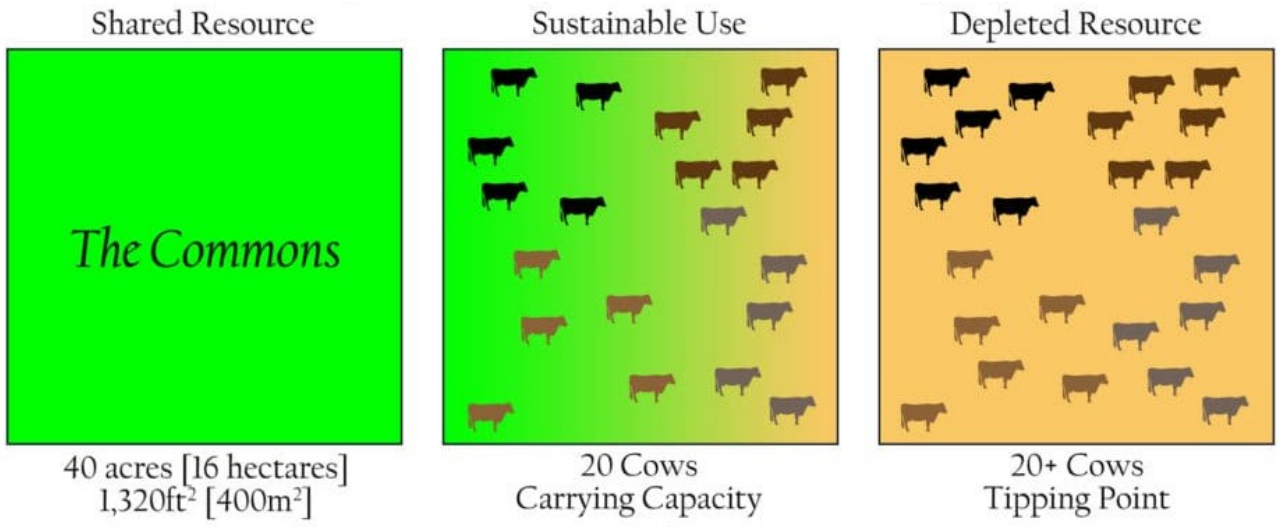




If we can prevent the defection at the last trial n (and all n_{-i} trials), we can incentivize cooperation (e.g., infinitely and indefinitely repetead games)

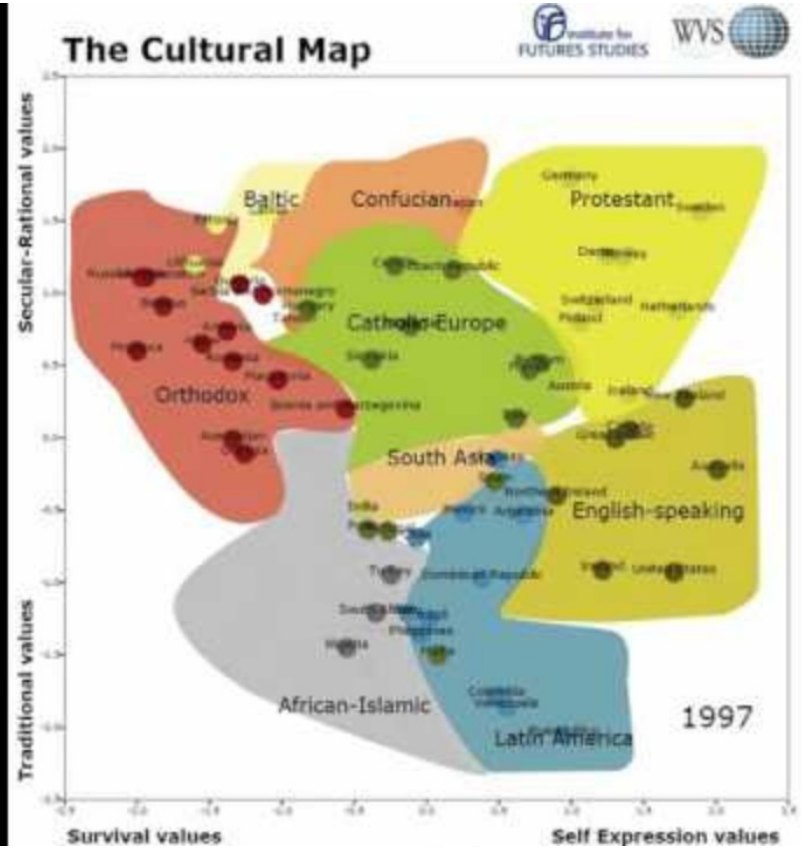
Cooperation at the population-level

Joshua Greene, Moral Tribes (2013). *Penguin Press*



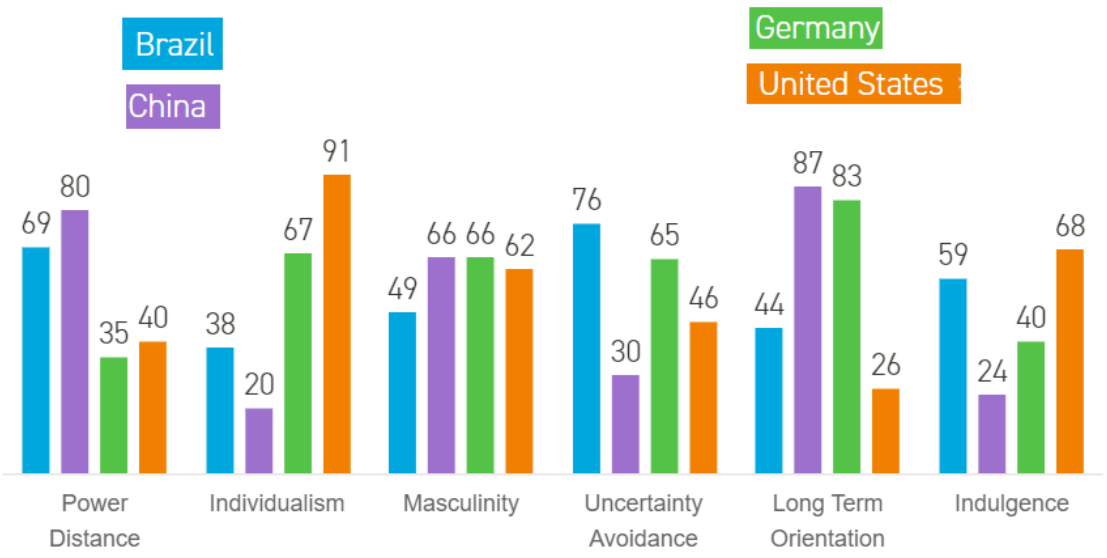
- Hardin: We need *mutual coercion mutually agreed upon*
- Greene: *Human morality evolved* to align individual self-interest with collective welfare
  - Shared norms, emotions, reputations make selfish behavior costly. Guilt, shame, pride, and gratitude act as internal "moral technologies"
- Rand, Nowak: Evolution gives rise to people who are truly altruistic and cooperate (*Intuitive reciprocation*)

Inglehart-Welzel culture map



1. *Traditional vs secular–rational values*: traditional values emphasize religion, parent–child ties, deference to authority, absolute standards; secular-rational values emphasize more acceptance of divorce, abortion, euthanasia, etc.
2. *Survival vs self-expression values:* survival values emphasize economic/physical security, ethnocentrism, low trust/tolerance; self-expression values emphasize subjectively defined well-being, tolerance of diversity, participatory decision-making, environmental protection.

Hofstede's cultural dimensions theory



- **Power distance**: describes the extent to which less powerful members of a society accept and expect unequal distribution of power.
- **Individualism vs collectivism**: measures whether people in a culture prioritise personal goals over group goals, or vice versa.
- **Masculinity vs femininity**: Masculinity refers to cultures that value competitiveness, achievement, and material success, while femininity values cooperation, care, and quality of life.
- **Uncertainty avoidance**: measures how comfortable a culture is with ambiguity, change, and the unknown.
- **Long-term avoidance**: reflects whether a culture prioritises future rewards over immediate results.
- **Indulgence vs restraint**: looks at how freely societies allow people to gratify their desires and enjoy life.

Can we predict that we will observe cooperation
to the relatively same extent,
in distinct societies,
and despite their distinction in values,
if we assume the same set of mechanisms for the evolution of cooperation?

In other words, does the set of 5 mechanisms for the evolution of cooperation guarantee cooperation everywhere?

NACS 645 – The origins of social cognition

-

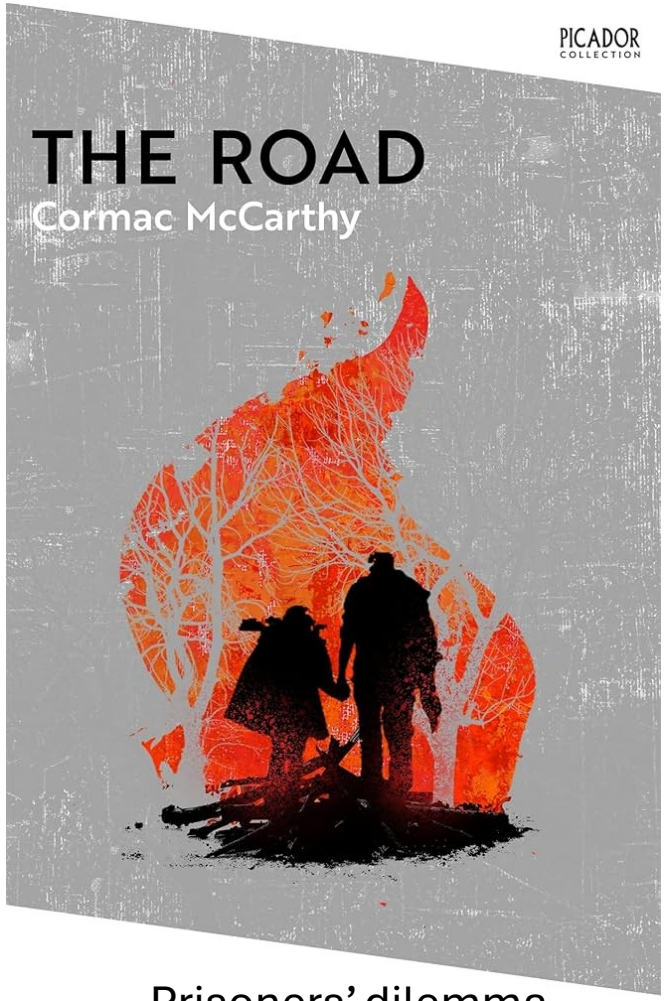
Valentin Guigon

DEPARTMENT OF PSYCHOLOGY

SLD Lab

PROGRAM IN NEUROSCIENCE & COGNITIVE SCIENCE

Choosing actions – Game theory



Game Theory assumptions:

- Each agent has its own description of states of the world (beliefs)
- Each agent has a utility function (preferences, uncertainty profile)
- Each agent maximizes expected utility (decision-theoretic rationality)

Given these assumptions, each player ***best responds*** to the others (*Nash equilibrium*)

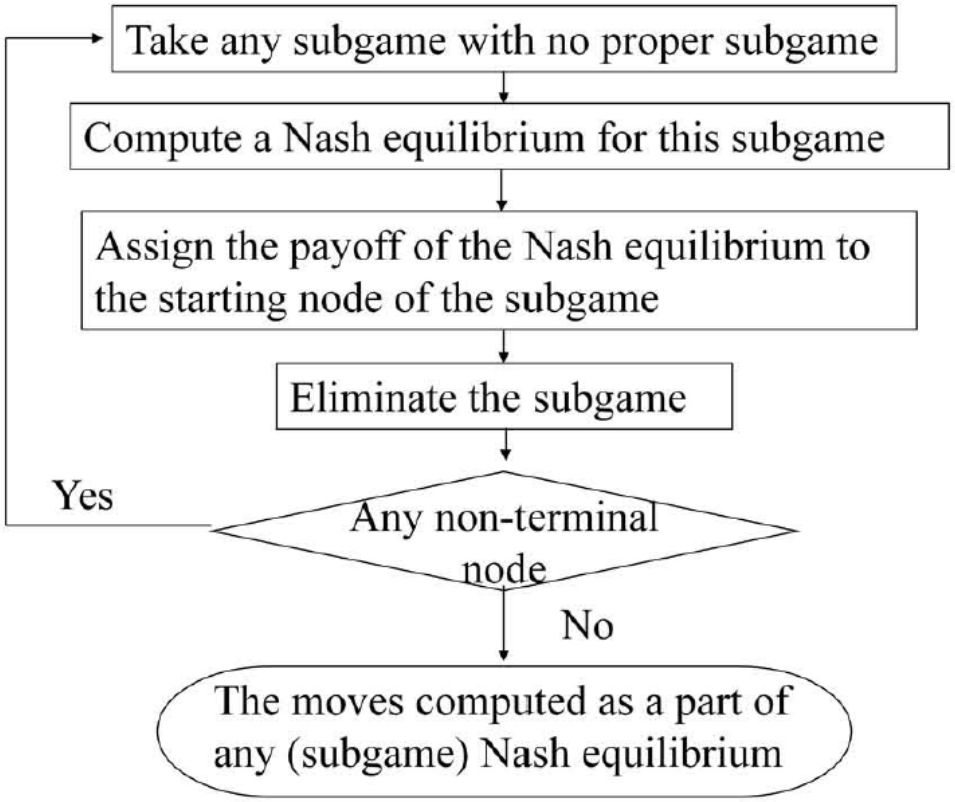
Nash equilibrium:

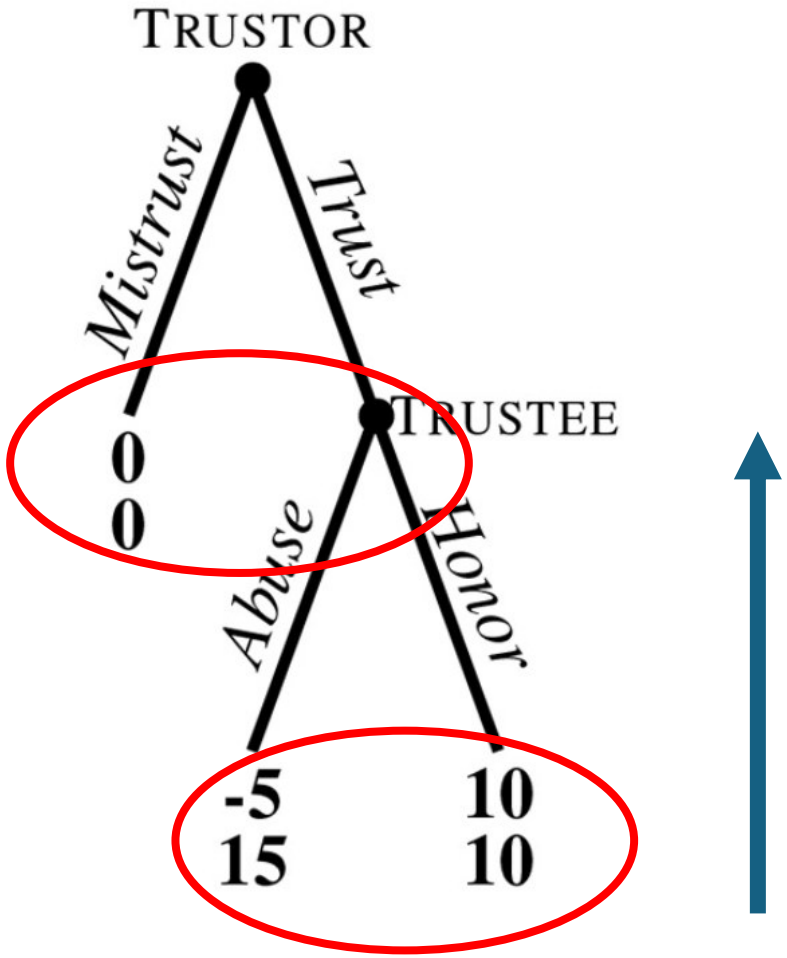
- A set of actions (action profible that is stable) with which each player's action maximizes his/her payoff given the actions of the others
- Each player's strategy is optimal given the strategies of all other players, so no player can benefit by unilaterally changing their action.
- Over time, as players learn the game, equilibria tend to emerge, while non-equilibrium profiles should disappear.

Prisoners' dilemma

	C	*D*
C	-1,-1	-4,0
D	0,-4	-3,-3

How to find the best action – backward induction





If we can prevent the defection at the last trial n (and all n_{-i} trials), we can incentivize cooperation (e.g., infinitely and indefinitely repetead games)

Game Theory and Nuclear deterrence

The TSEconomist Student Magazine

Game Theory & Nuclear Strategy

Game theory tends not to be the most popular class for many M1 Economics students. Mathematical formalism and rigorous logical arguments tend to scare or bore rather than engage. What makes game theory truly exciting, however, is its wide range of applications from evolution biology to political science. The discussion of military strategy during the Cold War was the crucial catalyst that brought game theory onto the stage in the first place. This article will give a historical overview of the development of nuclear strategy during the Cold War era and show parallels to game theoretic applications.

Philip Hanspach, 2016, Game Theory & Nuclear Strategy, available at: *The Tseconomist*

How To Win A Nuclear Standoff

President Trump and Kim Jong Un's saber-rattling is dangerous, but not irrational

By Oliver Roeder
Filed under The Trump Administration
Published Sep. 6, 2017

Oliver Roeder, 2017, How to win a nuclear standoff, available at: *FiveThirtyEight*

Pure and mixed games

Stag Hunt

	Stag	*Hare*
Stag	8,8	0,7
Hare	7,0	5,5

1 Nash equilibrium

Side of the road

	Left	*Right*
Left	1,1	0,0
Right	0,0	1,1

2 Nash equilibria

Pure coordination

Prisoners' dilemma

	C	*D*
C	-1,-1	-4,0
D	0,-4	-3,-3

1 Nash equilibrium

Battle of the sexes

	B	*F*
B	2,1	0,0
F	0,0	1,2

2 Nash equilibria

Mixed motive: coop / competition

Matching pennies

	Heads	*Tails*
Heads	1,-1	-1,1
Tails	-1,1	1,-1

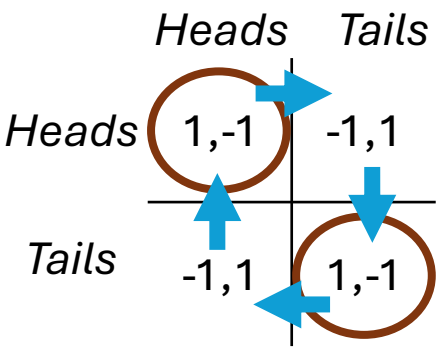
0 Nash equilibrium

Pure competition

Mixed strategies

- In **strategic environments with mixed or competitive motives, deterministic strategies are exploitable**:
  predictability creates information asymmetry that can be exploited
- Idea:
  - a) confuse the opponent by introducing randomness,
  - b) avoiding patterns by randomizing completely
- Let's define a strategy (s_i) for the agent (i) as any probability distribution over the set of actions (A_i)
  - Pure strategy: only one action played with $p = 1$
  - Mixed strategy: more than one action played with $p > 0$
- In matching pennies, optimal strategy is to play each action a_i with equal proba $p = \frac{1}{|A_i|}$

Matching pennies



0 Nash equilibrium

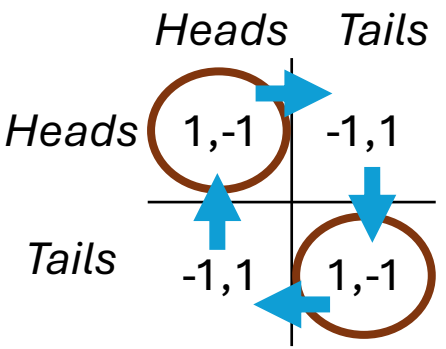
Pure competition

Mixed strategies – exploiting others

Based on the works of Matthew O. Jackson, Kevin Leyton-Brown & Yoav Shoham, Game Theory, on *Stanford online*.

- In matching pennies, optimal strategy is to play each action a_i with equal proba $p = \frac{1}{|A_i|}$
  - Any player deviating from $p = \frac{1}{|A_i|}$ becomes exploitable:
    if P2 plays H,T with [.55,.45], P1 best response is H with $\mathrm{p} = 1$
  - Any player using $p = \frac{1}{|A_i|}$ is unexploitable:
    [.45,.55] vs [.5,.5] would yield utility $u_1 = u_2 = 0$
- Yet, humans struggle to generate truly random sequences, making them exploitable

Matching pennies



0 Nash equilibrium

Pure competition

- Players may instead rely on pseudo-random or deceptive strategies, faking patterns to induce false beliefs in opponents. Efficacy depends on types of games and opponents:
  - It can succeed only if player remains unpredictable, e.g., due to uncertainties
  - GTO players are guaranteed immunity from exploitation and may detect patterns
  - Non-GTO players can overfit perceived regularities

Interpreting mixed strategies

Matching pennies

	Heads	*Tails*
Heads	1,-1	-1,1
Tails	-1,1	1,-1

Battle of the sexes

	B	*F*
B	2,1	0,0
F	0,0	1,2

Different interpretations of mixed strategies

- Randomize to confuse the opponent
  (e.g., matching pennies)
- Randomize when uncertain about the other's action
  (e.g., battle of the sexes)
- Concise description of what happens in repeated play
  (probability distributions over sets of actions)
- Describes population dynamics
  (e.g., 2 agents drawn from a population, all with deterministic strategies. Mixed strategies gives the probability of getting each pure strategy available)

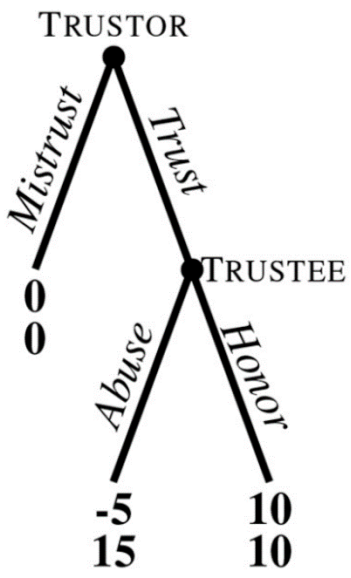
Perfect vs Imperfect information games

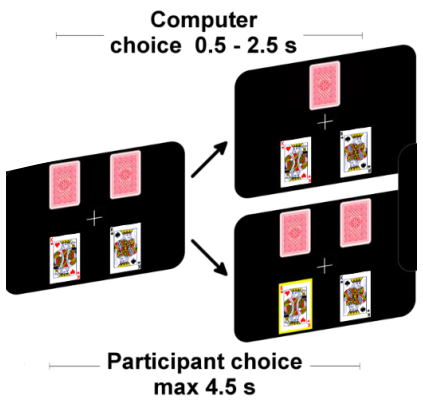
Perfect information games

When all players have access to all the information of a game.
(e.g., Trust game, Chess, Tic-tac-toe)

Easy to solve:

- Every player knows all past actions and the full state of the game
- Backward induction ensures a pure strategy Nash equilibrium
- Randomization is unnecessary, because decision is transparent





Imperfect information games

When some actions of states are hidden from at least one player
(e.g., MP, Rock-Paper-Scissors, Poker)

Harder to solve:

- Players do not observe some past moves or states (e.g., simultaneous moves, hidden cards)
- Players' may better be probabilistic
- Pure-strategy equilibria may not exist, but mixed-strategy equilibria do

Incomplete information games

What if players are **uncertain about the *structure* of the game** itself?
→ number of players, available actions, payoffs for each action profile

- Each player faces several *possible games* they *believe* could be the real one
- All players share a **common prior** over possible games, but receive **private signals** to **update their beliefs about the true game**

Interpretation**:**
- Some players know information that others do not
- Each player's *epistemic type* encodes *what they know* and *what they believe others know*
- **Optimal actions depend on beliefs about others'** epistemic types

- Deception manipulates others' beliefs, exploiting information asymmetries
- Predicting and interpreting others' moves requires modeling their *beliefs*, not just their *payoffs – i.e., mentalizing*

	$I_{2,1}$	$I_{2,2}$
$I_{1,1}$	MP: $\begin{pmatrix} 2,0 & 0,2 \\ 0,2 & 2,0 \end{pmatrix}$, $p = 0.3$	PD: $\begin{pmatrix} 2,2 & 0,3 \\ 3,0 & 1,1 \end{pmatrix}$, $p = 0.1$
$I_{1,2}$	Coord: $\begin{pmatrix} 2,2 & 0,0 \\ 0,0 & 1,1 \end{pmatrix}$, $p = 0.2$	BoS: $\begin{pmatrix} 2,1 & 0,0 \\ 0,0 & 1,2 \end{pmatrix}$, $p = 0.4$

Incomplete information games

Task setup

- There are **4 possible games**: *Matching Pennies (MP), Prisoner's Dilemma (PD), Coordination (Coord),* and *Battle of the Sexes (BoS)*
- **Nature** randomly selects which one is played (not publicly revealed)
- Each player receive a private signal

Information structure: let's say all players received private info

- **P1** knows they are in the **bottom row** → game is either *Coord* or *BoS*
- **P2** knows they are in the **right column** → the game is either *PD* or *BoS*

Interpretation

- Each player faces **uncertainty about the true game** and must reason about the opponent's beliefs and likely behavior
- Both share a **common prior** over which of the four games Nature chose

Inference logic

- If game is *Coord*, P2 will play *MP* or *Coord*; otherwise P2 will play *PD* or *BoS*. By observing P2, P1 can update beliefs about the true game
- **e.g., if P2 doesn't play pure strategy, the game is *BoS***

	$I_{2,1}$	$I_{2,2}$
$I_{1,1}$	MP: $\begin{bmatrix} 2,0 & 0,2 \\ 0,2 & 2,0 \end{bmatrix}$ $p = 0.3$	PD: $\begin{bmatrix} 2,2 & 0,3 \\ 3,0 & 1,1 \end{bmatrix}$ $p = 0.1$
$I_{1,2}$	Coord: $\begin{bmatrix} 2,2 & 0,0 \\ 0,0 & 1,1 \end{bmatrix}$ $p = 0.2$	BoS: $\begin{bmatrix} 2,1 & 0,0 \\ 0,0 & 1,2 \end{bmatrix}$ $p = 0.4$

Mentalizing, central to strategic interactions

Tomasello, 2008. *MIT Press*.
Tomasello, 2020. *Episteme*.
Sperber et al., 2010. *Mind & Language*.

Cooperative communication

- Aligning beliefs and actions requires mutual **transparency** of minds
- Language evolved for coordination and shared understanding (Tomasello, 2008, 2020): sharing mental states, referencing *what is*, building ground truth in common reality
- But transparency allows **exploitation**: free riders gain if they remain undetected

Need for truthfulness

- Stable cooperation requires a) most communication be **truthful**; and b) truth is the default expectation
- Yet, this enables strategic deception and exploiting others
- Too many free riders → collapse of cooperation

Dual pressures

- Mentalizing makes cooperation possible and deception feasible
- Lying is effortful because it requires simulating others' minds
- Communication evolved under **dual pressures**: **cooperation** through truth-sharing **vs. competition** through manipulation

Epistemic vigilance

- To protect cooperation, evolution favored mechanisms for detecting dishonesty (Sperber, 2010)
- Cognitive systems that assess reliability and sincerity of communicated information, acting as counterweights to gullibility

The development of ToM

Bettle & Rosati, 2021. *Language learning and Development.*
Fujita, Devine, Hughes, 2022. *Cognitive Development.*
Rakoczy, 2022. *Nature Reviews Psychology.*

(shared with other primates)		*(human-unique?)*	**function**
Simple gaze following	→	Joint attention; sustained attention to others	• Recognize positive intentions • Share attention with others
Goal and intention attribution	→	Attributing cooperative or shared intentions	• Identify cooperative opportunities • Understand helpful communication • Maintenance so all parties benefit
Level 1 Perspective taking (presence/absence)	→	Level 2 perspective-taking ('how' it looks)	• 'Bird's-eye view' of joint acts • Facilitate coordination
Attribute knowledge and ignorance	→	Attribute false beliefs	• Facilitate coordination • Avoid miscommunication • Reputation management

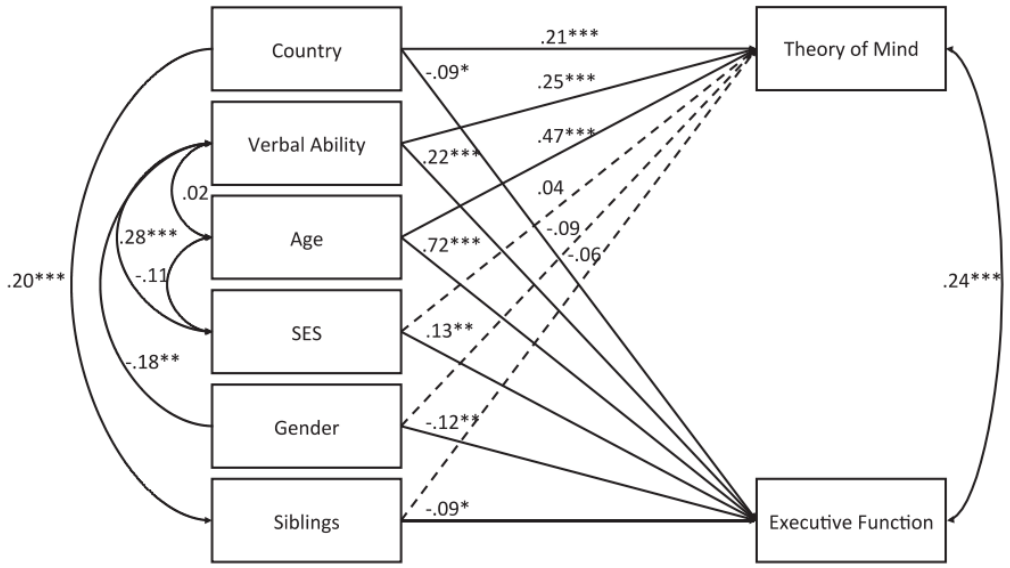
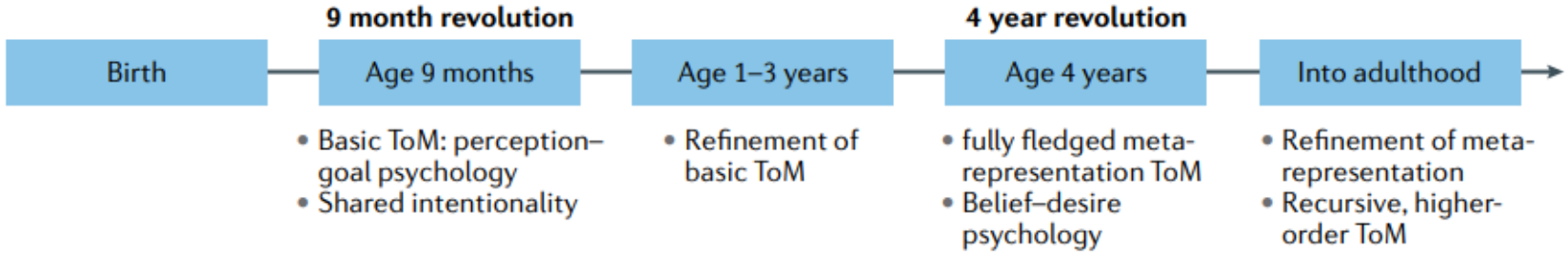


Fig. 1. A structural equation model of cross-cultural differences in theory of mind and executive function with covariates. Country: 0 = Japan, 1 = UK. Gender: 0 = girls, 1 = boys. SES = Socioeconomic Status. *Note.* * $p < .05$. ** $p < .01$. *** $p < .001$.



NACS 645 – Moral technologies

-

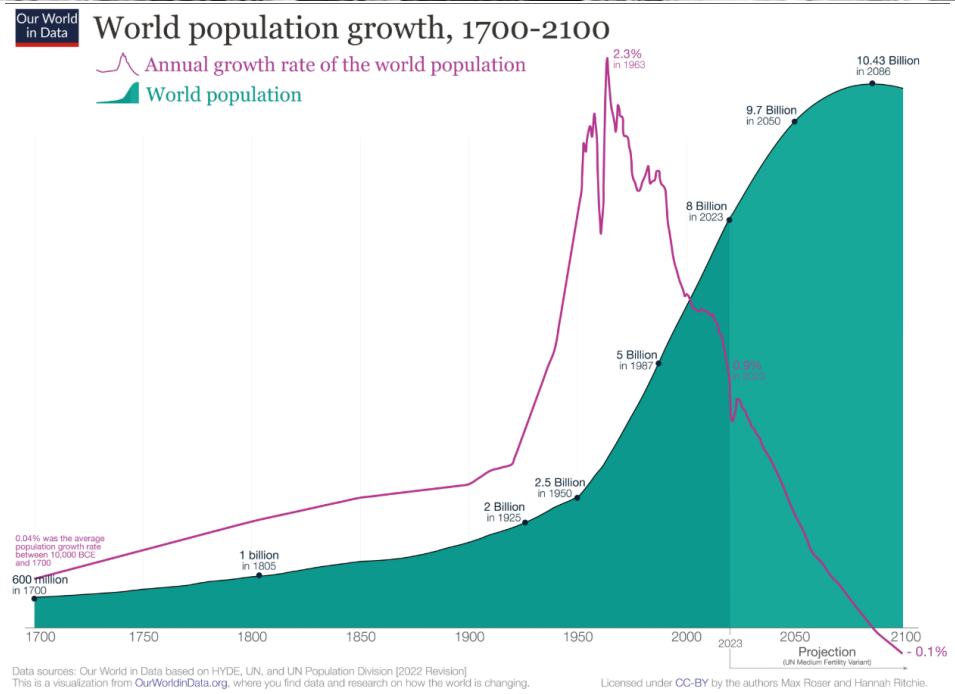
Valentin Guigon

DEPARTMENT OF PSYCHOLOGY

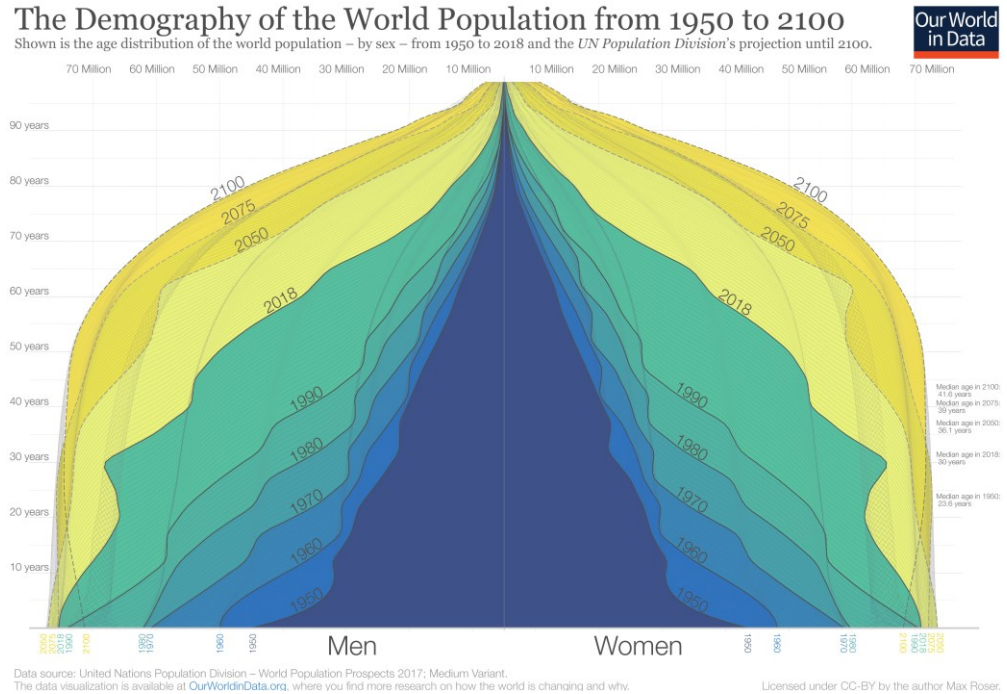
SLD Lab

PROGRAM IN NEUROSCIENCE & COGNITIVE SCIENCE

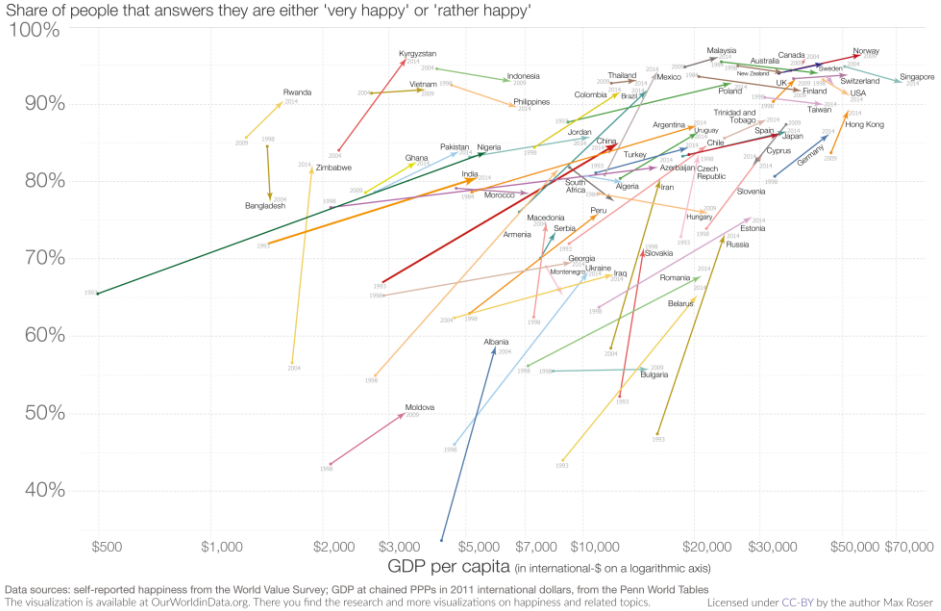
Population explosion



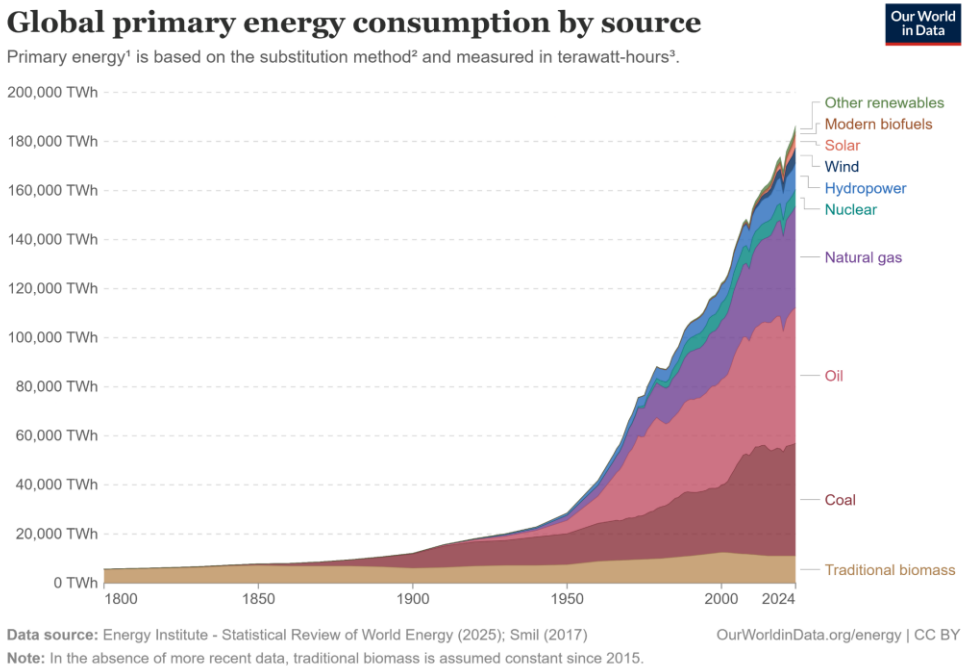
Higher life expectancy



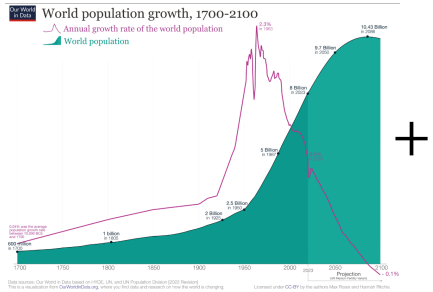
Progressing income and happiness

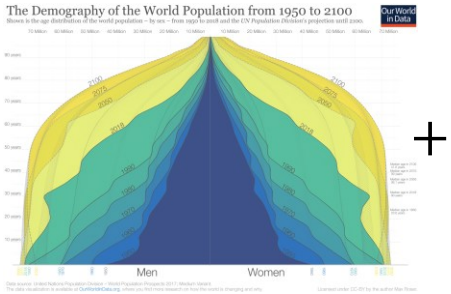


Stronger energy consumption

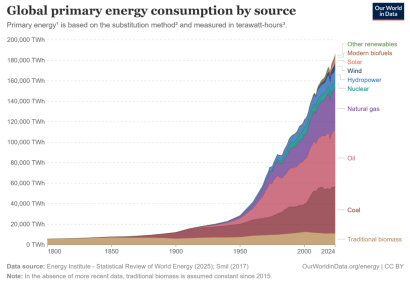


Common Good maximization problem

$Good = \omega_1 \cdot$ 

$+\omega_2 \cdot$ 

$+\omega_3 \cdot$

$+\omega_4 \cdot$ 

$+\varepsilon$

Hardin:

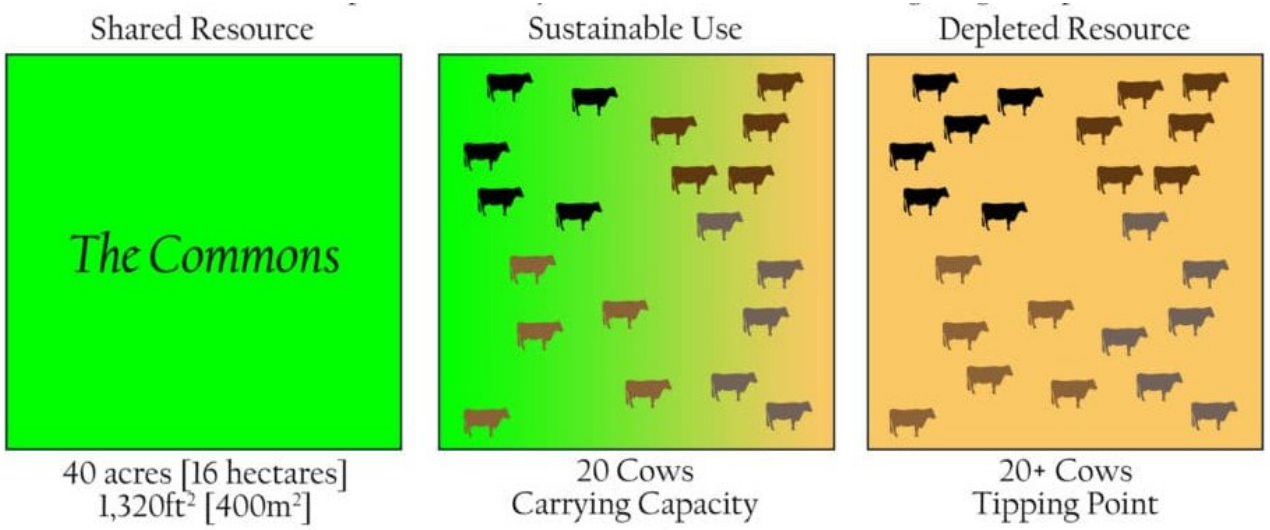
- In a finite world, *maximizing population* requires driving ***work calories per person toward zero***: no art, no gastronomy, no leisure
- What we gain in numbers, we lose in quality of life
- ***Technical solution*** *(*infinite energy supply) **only shifts** *energy acquisition problem* to *energy dissipation problem*
- **We can't jointly maximize two variables** (e.g., population, welfare): *mathematically and materially incompatible*
- Hence, the task is not to **maximize**, but to **optimize**, within constraints, the *Good* we choose

- **Technologies lift constraints** and rise ceilings, **but do not remove the maximization problem**
- **Optimizing the *Good*** in a finite world therefore demands a ***non-technical solution***
- It also creates a resource allocation problem

Tragedy of the commons

Pasture open to all (i.e., commons):

- Each herdsman can **increase private utility** by adding more cattle
- Finite resource, but unrestricted access



Adapted from: Ric Stephens, University of Oregon College of Design, School of Planning, Public Policy, and Management.

Utility Calculation:

- **Positive component: +1** (private gain from the extra cow)
- **Negative component: −ε** (fractional share of the collective loss from overgrazing)

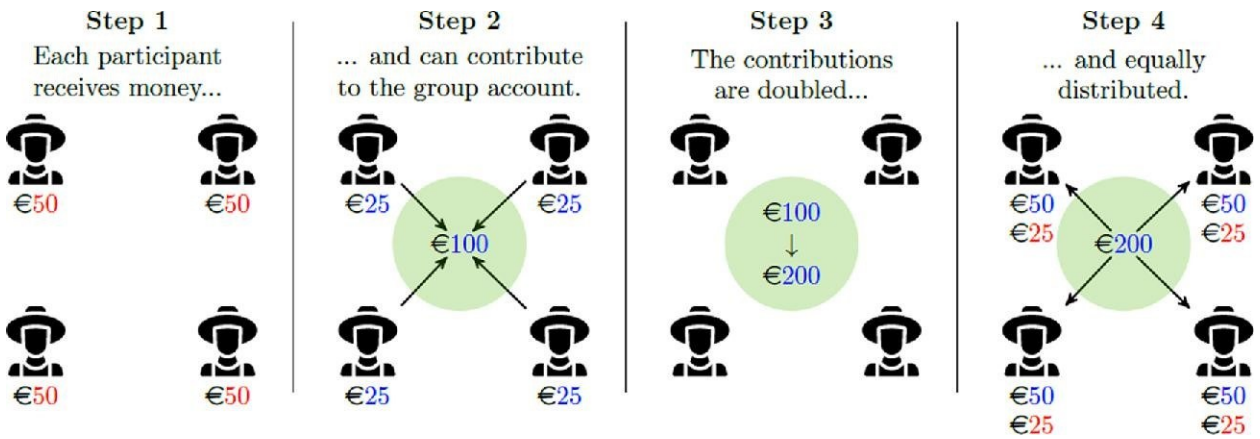
The Tragedy:

- Since $\varepsilon \ll 1$, rational choice for each is to add another cow…
  until pasture is degraded beyond recovery
- **Individually rational actions aggregates into collective ruin**

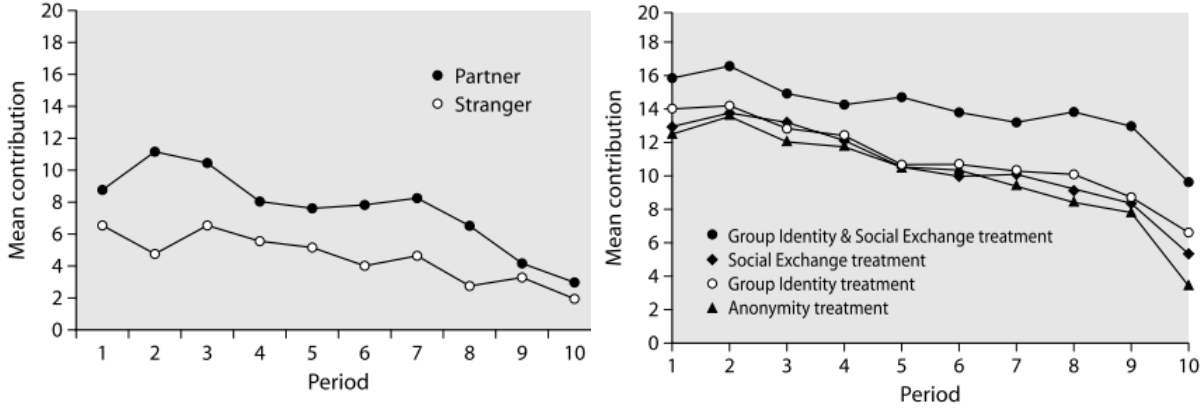
Implications:

- The commons is a **non-zero-sum game of mixed cooperative and competitive motives**
- In a finite system, **individual liberty and collective welfare cannot both be maximized**
- Technologies may lift constraints but cannot alter the incentive structure that rewards self-interest
- Avoiding tragedy requires **collective rules, limits, and moral restraint**

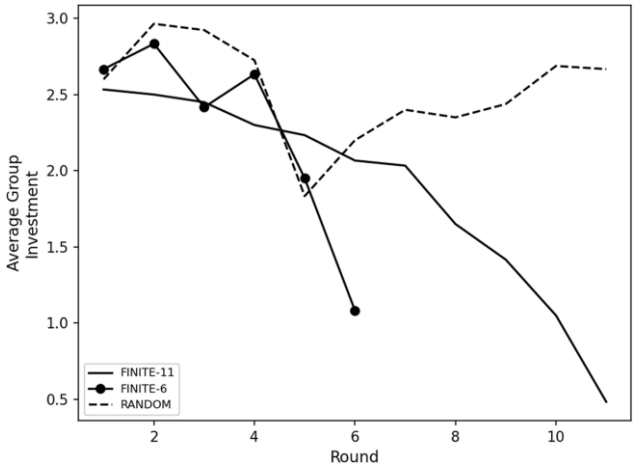
Public goods games



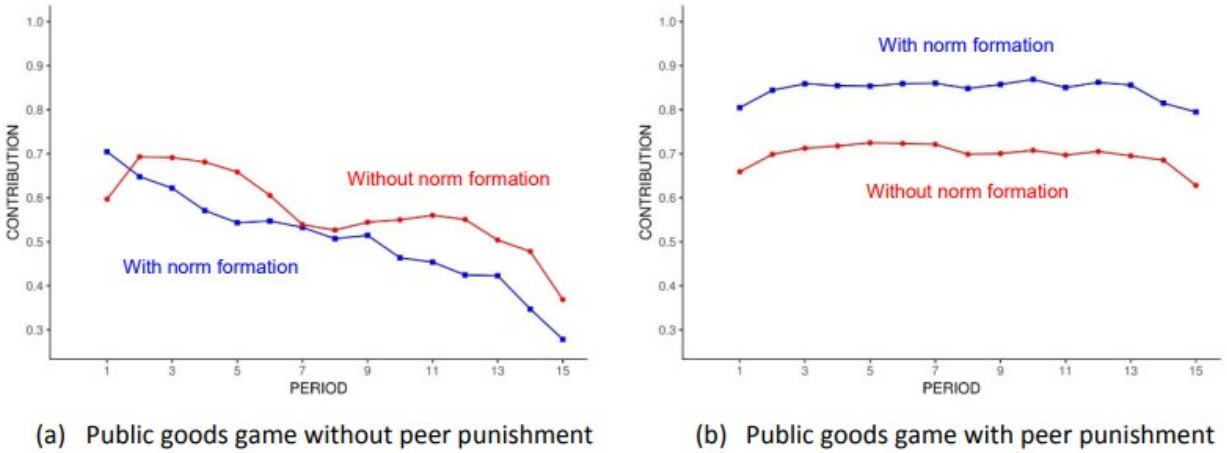
Adapted from Rommel et al., 2022. *Q Open*.



Adapted from Gächter & Herrmann, 2006. Cooperation in primates and humans: mechanisms and evolution



Adapted from Anderson et al., 2024. *Journal of the Economic Science Association*



Adapted from Fehr & Schurtenberger, 2018. *Nature Human Behavior.*

Greene & Young, 2020. The Cognitive Neuroscience of Moral Judgment and Decision-Making
Greene, 2014. *Penguin Press*.
Greene, 2015. *Cognition*.

The moral cognition

Morality is a collection of biological and cultural solutions that promote and sustain cooperation, not special organ or processes.

Moral cognition integrates valuation, reasoning, control, emotions to regulate cooperative behavior.

Moral emotions operate as *psychological carrots and sticks*.

e.g., *Guilt* → self-directed punishment; *Gratitude* → reward toward others

- Moral emotions (empathy, guilt, righteous anger, compassion) act as internalized regulators:
  They encode explicit moral motives
  ("I should," "That's wrong")
- Non-moral social emotions (gossip, embarrassment, vengefulness, in-group favoritism) act as indirect or local regulators:
  They manage reputation, kinship, coalitions, but are not moral obligations

Neural architecture:

- No dedicated "moral" module; relies on brain systems applied to cooperative problems :
  - Value and motivation – vmPFC, striatum
  - Emotion and affective learning – amygdala, insula
  - Mental-state representation – TPJ, mPFC
  - Cognitive control – dlPFC, ACC
  - Simulation and imagination – hippocampus, DMN

Implications:

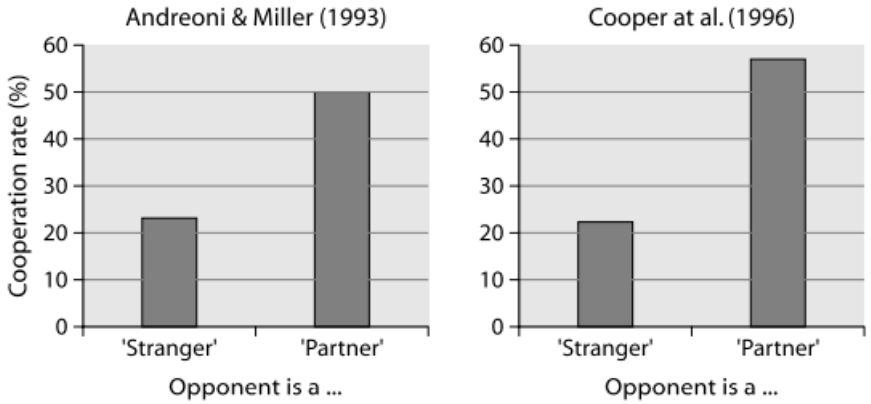
Within this framework:

- The *moral* character of a behavior/process comes from its *cooperative* function
- The "moral brain" is the brain's distributed network tackling cooperative problems through its ordinary systems for value, emotion, control, and social cognition

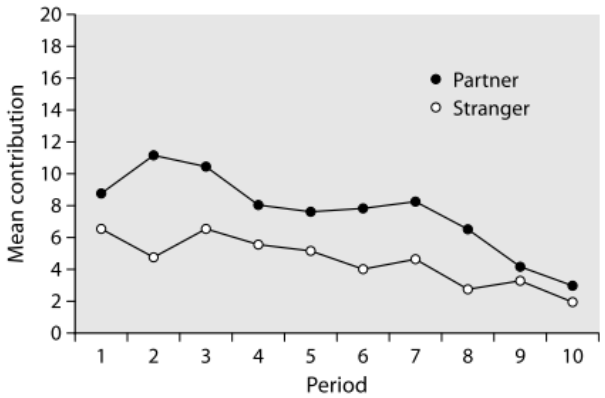
Maintaining cooperation

Gächter & Herrmann, 2006. *Cooperation in primates and humans: mechanisms and evolution.*
Fehr & Schurtenberger, 2018. *Nature Human Behavior.*

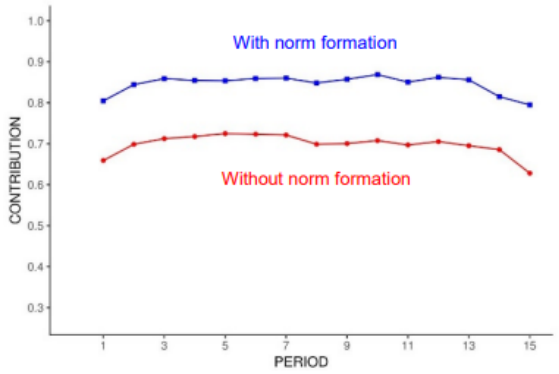
Despite overgrazing, tax evasion, or failed agreements on environmental protection, humans achieve high levels of cooperation – even among strangers



Prisoner's Dilemma with constant (Partner) and randomly-changing (Stranger) opponents



Public good game with constant (Partner) and randomly-changing (Strangers) groups



(b) Public goods game with peer punishment

1. Evolutionary Pathways for Cooperation
(selection mechanisms for cooperation; cf. Rand & Nowak, 2013)

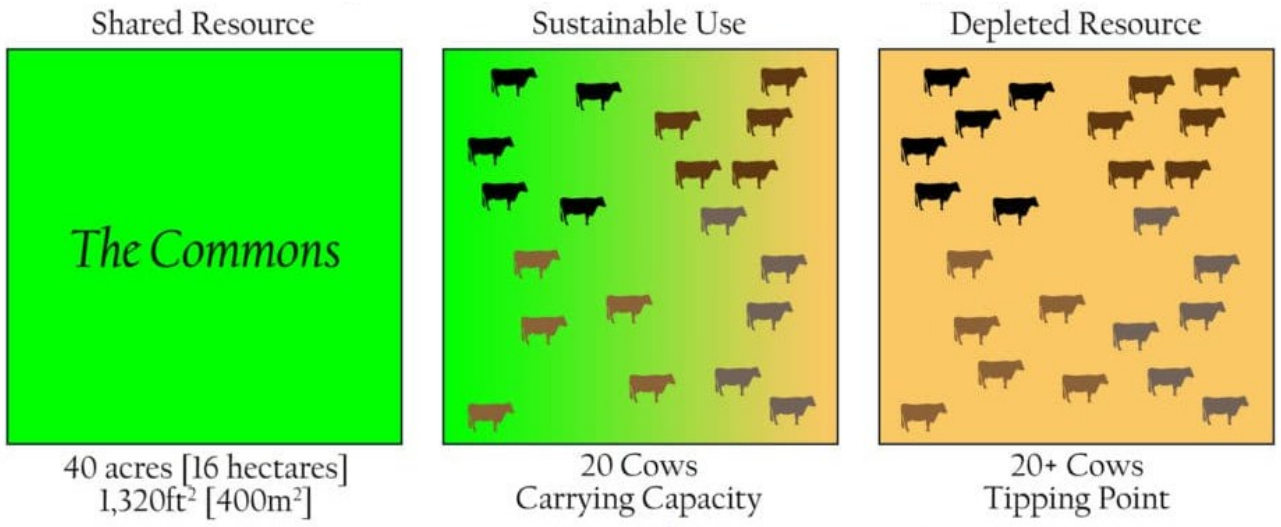
- **Kin selection**
- **Direct reciprocity**
- **Indirect reciprocity**
- **Spatial selection**
- **Group selection**

2. Proximate and Institutional Reinforcers
(mechanisms for cooperation within societies)

- **Altruistic punishment** – Costly sanctioning of free riders
- **Moral emotions** – Anger, guilt, shame, and empathy as immediate motivators for punishment and compliance
- **Social norms** – *Known standards of behavior based on shared beliefs about how individuals ought to behave in a specific context*
- **Rules and institutions** – Formal codifications of norms with enforcement mechanisms

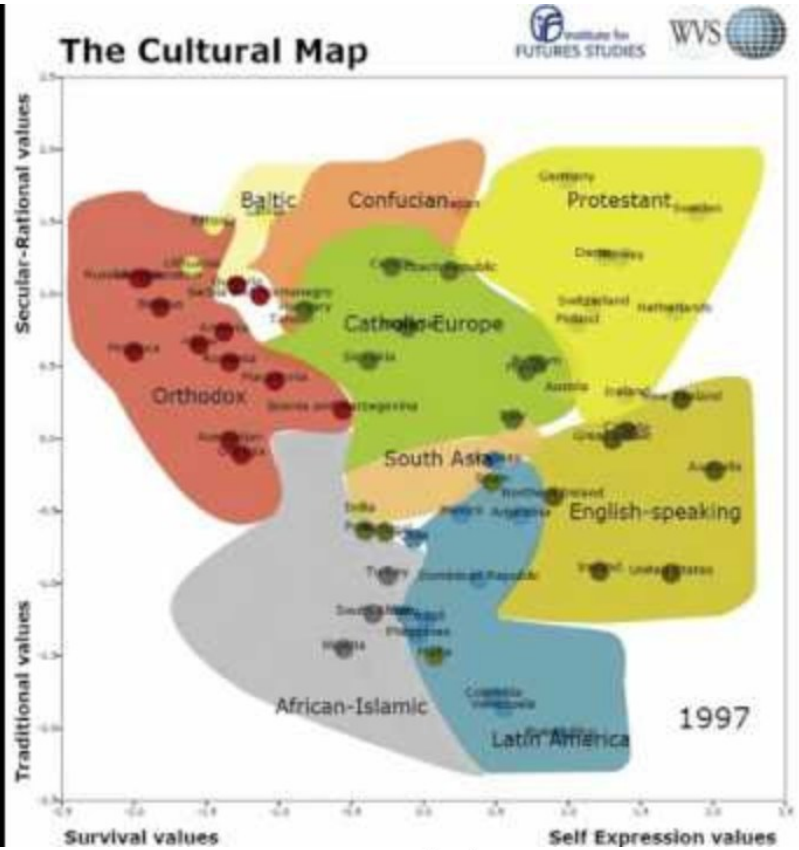
Cooperation at the population-level

Joshua Greene, Moral Tribes (2013). *Penguin Press*



- Hardin: We need *mutual coercion mutually agreed upon*
- Rand, Nowak: Evolution gives produces genuine cooperators – intuitive reciprocation and selective mechanisms sustain coop.
- Greene; Curry: *Human morality evolved* to align individual self-interest with collective welfare
  - Shared norms, emotions, reputations make selfish behavior costly. Guilt, shame, pride, and gratitude act as internal "moral technologies"
  - Emotions regulate local cooperation; reason, norms and institutions extend cooperation to the group level

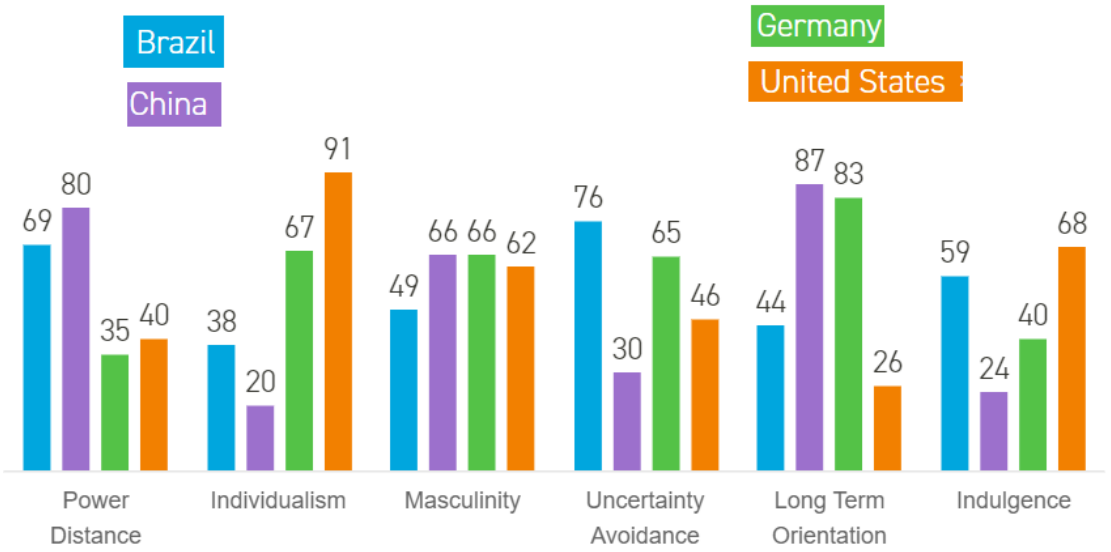
Cultural differences



Inglehart-Welzel culture map

1. *Traditional vs secular–rational values*:
   traditional values emphasize religion, parent–child ties, deference to authority; secular-rational values emphasize more acceptance of divorce, abortion, euthanasia, etc.

2. *Survival vs self-expression values:*
   survival values emphasize economic/physical security, ethnocentrism, low trust/tolerance; self-expression values emphasize subjectively defined well-being, tolerance of diversity, participatory decision-making, etc.



Hofstede's cultural dimensions theory

- **Power distance**: describes the extent to which less powerful members of a society accept and expect unequal distribution of power.
- **Individualism vs collectivism**: measures whether people in a culture prioritise personal goals over group goals, or vice versa.
- **Masculinity vs femininity**: Masculinity refers to cultures that value competitiveness, achievement, and material success, while femininity values cooperation, care, and quality of life.
- **Uncertainty avoidance**: measures how comfortable a culture is with ambiguity, change, and the unknown.
- **Long-term avoidance**: reflects whether a culture prioritises future rewards over immediate results.
- **Indulgence vs restraint**: looks at how freely societies allow people to gratify their desires and enjoy life.

Kitayama & Uskul, 2011. *Annual Review of Psychology.*
Chiao et al., 2013. Psychological Inquiry.

Culture, Mind, and the Brain

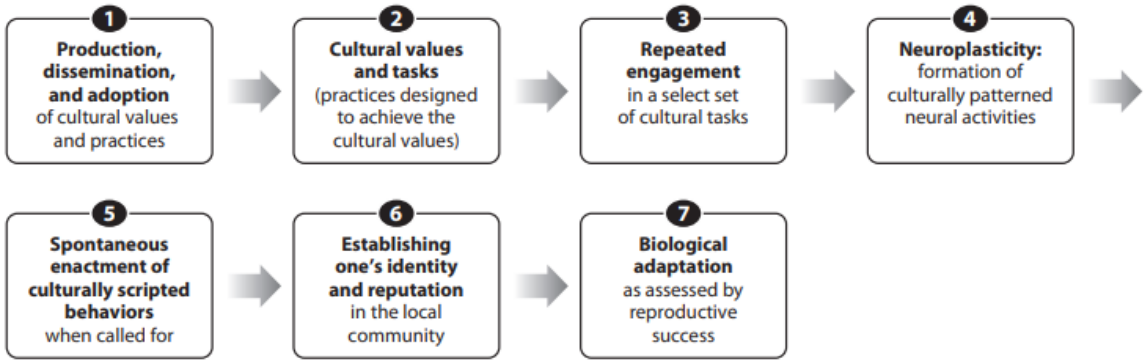


Figure 1

A neuro-culture interaction model. Values and practices of culture are produced, disseminated, and adopted as a function of a variety of collective-level factors. Individuals choose some select set of available cultural practices as their own cultural tasks. They then actively engage in them so as to realize their culture's primary values such as independence and interdependence in their own idiosyncratic ways. Repeated engagement in the cultural tasks results in culturally patterned brain activities, which in turn enable the individuals to spontaneously and seamlessly enact the culturally scripted behaviors when such behaviors are called for by situational norms. The ability of the individuals to perform the culturally scripted behaviors when normatively required to do so enhances their own identity and reputation as a decent member of the cultural tradition and, eventually, their ability to achieve biological adaptation as assessed by reproductive fitness.

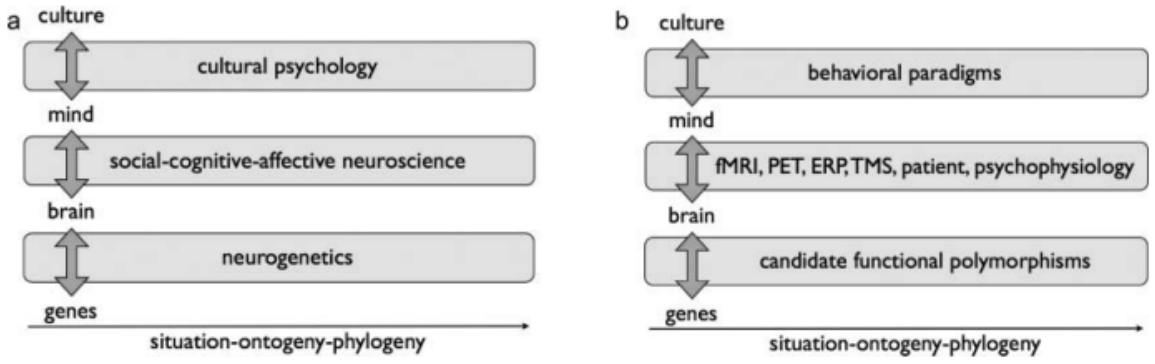


Figure 1. Framework of cultural neuroscience (Chiao, 2009, 2011; Chiao & Ambady, 2007). *Note.* (a–b) Research in cultural neuroscience integrates theory and methods from cultural psychology, social-cognitive-affective neuroscience and neurogenetics across multiple time scales, specifically situation, ontogeny and phylogeny. *Source.* Adapted from Chiao and Ambady (2007) and Chiao (2009).

If we assume the same set of mechanisms for the evolution of cooperation,
should we observe cooperation to the (relatively) same extent in distinct societies?

NACS 645 – The neural need to infer others

-

Valentin Guigon

DEPARTMENT OF PSYCHOLOGY

SLD Lab

PROGRAM IN NEUROSCIENCE & COGNITIVE SCIENCE

Two-player strategic games

Stag Hunt

	Stag	*Hare*
Stag	8,8	0,7
Hare	7,0	5,5

1 Nash equilibrium

Side of the road

	Left	*Right*
Left	1,1	0,0
Right	0,0	1,1

2 Nash equilibria

Pure coordination

Prisoners' dilemma

	C	*D*
C	-1,-1	-4,0
D	0,-4	-3,-3

1 Nash equilibrium

Battle of the sexes

	B	*F*
B	2,1	0,0
F	0,0	1,2

2 Nash equilibria

Mixed motive: coop / competition

Matching pennies

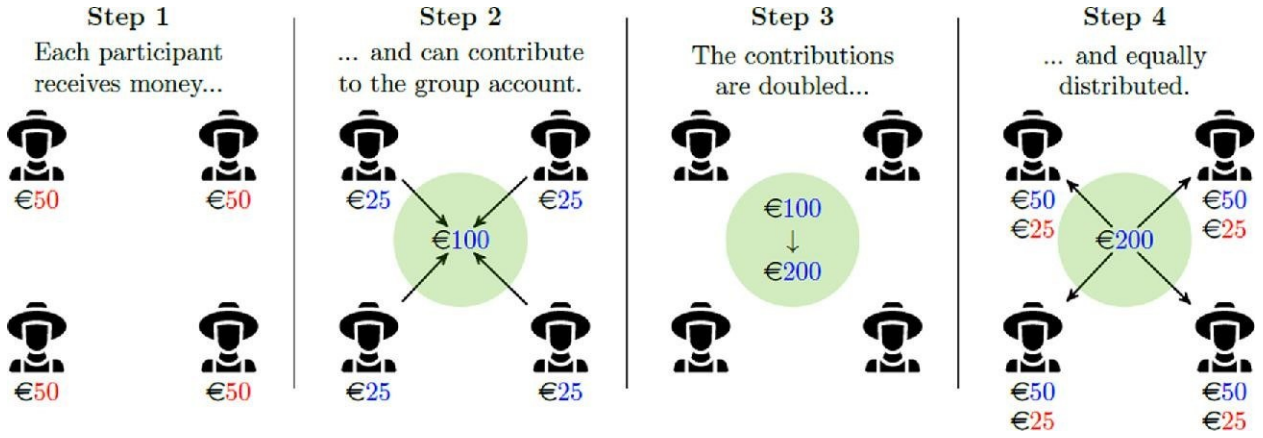
	Heads	*Tails*
Heads	1,-1	-1,1
Tails	-1,1	1,-1

0 Nash equilibrium

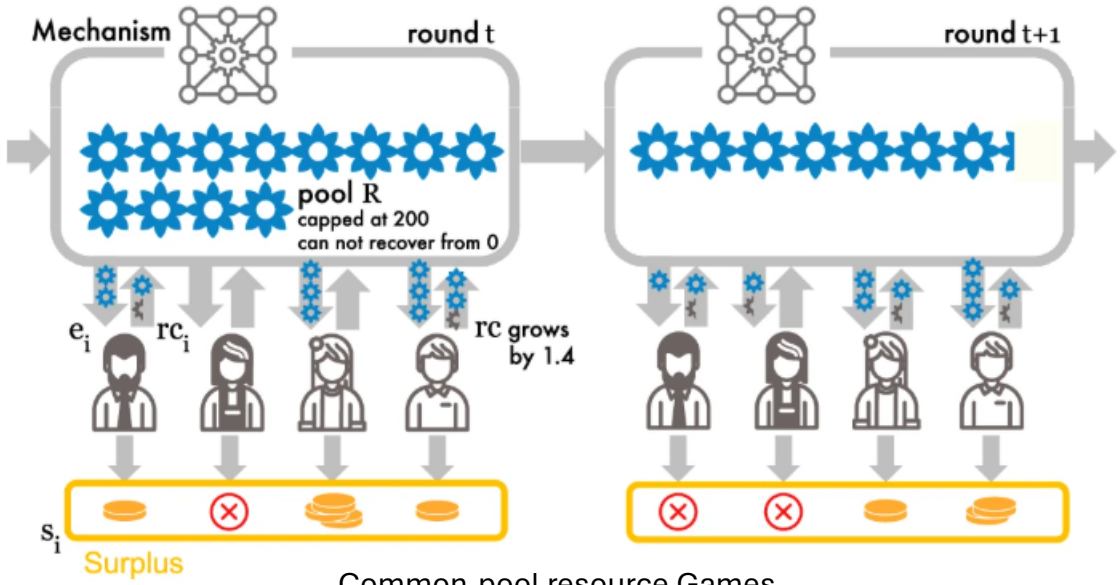
Pure competition

- In **strategic environments with mixed or competitive motives, deterministic strategies are exploitable**: predictability creates information asymmetry that can be exploited
- **Cooperation emerges** when it is **incentivized**, and is supported by **direct reciprocity**: **Mutual monitoring** and **repeated interactions** allow conditional cooperation (Best Responses are identifiable)
- Expectations and emotions may act as proximate regulators

N-player strategic games



Public Goods Game.
Adapted from Rommel et al., 2022. *Q Open*.



Common-pool resource Games.
Adapted from Koster et al., 2025. *Nature Communications*.

- In n-player strategic games with mixed motives, not only **deterministic strategies are exploitable**, but **games are hardly tractable**, and **often incomplete** (Bayesian)
  There is a high incentive for free-riding, which is responsible for cooperation collapse.
- **Cooperation emerges** when it is **incentivized**, and is **supported by indirect reciprocity**.
  Cooperation is sustained when **shared norms** appear and when **altruistic punishment** is allowed (external regulators); otherwise free-riders might trigger cascades of defection.

Cooperation at scale

Societal coordination

- In large, anonymous populations, direct and indirect reciprocity are **insufficient**
- Cooperation depends on **shared norms, social identity, and institutional enforcement**

Norms and expectations

- Norms are **shared rules sustained by mutual expectations and enforcement**. Abiding requires:
  - (i) **one believes most others follow the rule**
  - (ii) **one believes most others *approve* of the rule and will sanction non-compliance**
- Large-scale cooperation hinges on perceived compliance and enforcement
  Without, probability estimates of others' cooperation decrease, lowering the *subjective value* of cooperating

Institutions and identity

- **Social identity** extends mutual trust to symbolic communities
- **Institutions** (laws, contracts, markets, formal punishment systems) replace interpersonal reciprocity with credible enforcement. They act by:
  - **reshaping payoffs** (incentives for cooperation)
  - **altering expectations** (perceived probability that others will cooperate)
- Institutional trust built on norms and shared identity
  Erosion of legitimacy through unfairness, corruption, or unequal enforcement

Van Bavel et al., 2022. *Advances in experimental social psychology.*

Pathways to cooperation

Rand & Nowak, 2013. *TICS*.
Fehr & Schurtenberger, 2018. *Nature Human Behavior*.
Greene & Young, 2020. The Cog. Neuro. of Moral Judgment and Dec.-Making
Curry et al., 2022. *Review of Philosophy and Psychology*

Evolutionary Pathways for Cooperation
(selection mechanisms for cooperation; cf. Rand & Nowak, 2013)

- **Kin selection**
- **Direct reciprocity**
- **Indirect reciprocity**
- **Spatial selection**
- **Group selection**

Non-moral social emotions act as indirect or local regulators: manage reputation, kinship, coalitions, but are not moral obligations.

Cultural and Institutional Reinforcers
(external mechanisms for cooperation within societies)

- **Altruistic punishment** – Costly sanctioning of free riders
- **Social norms** – *Known standards of behavior based on shared beliefs about how individuals ought to behave in a specific context*
- **Rules and institutions** – Formal codifications of norms with enforcement mechanisms

Moral machinery
(evolved psychological and motivational mechanisms for internally regulated cooperation)

- Morality as a collection of biological & cultural solutions that promote and sustain cooperation.
- Moral cognition integrates valuation, reasoning, control, emotions to regulate cooperative behaviors.
- Moral emotions act as internalized regulators, by operating as *psychological carrots and sticks*.
  e.g., Guilt → self-directed punishment; *Gratitude* → reward toward others

Overview of morality-as-cooperation.

	Label	Problem/Opportunity	Solution
1	Family	Kin selection	Kin Altruism
2	Group	Coordination	Mutualism
3	Reciprocity	Social Dilemma	Reciprocal Altruism
4	Heroism	Conflict Resolution (Contest)	Hawkish Displays
5	Deference	Conflict Resolution (Contest)	Dove-ish Displays
6	Fairness	Conflict Resolution (Bargaining)	Division
7	Property	Conflict Resolution (Possession)	Ownership

Table 2 Twenty-one moral molecules

	Mutualism	Exchange	Hawk	Dove	Division	Possession
Kinship	Fraternity	Blood Revenge	Family Pride	Filial Piety	Gavelkind	Primo-geniture
Mutualism		Friendship	Patriotism	Tribute	Diplomacy	Common ownership
Exchange			Honour	Confession	Turn-taking	Restitution
Hawk				Modesty	Mercy	Munificence
Dove					Arbitration	Mendicance
Division						Queuing

M-a-C predits 127 moral & 127 immoral (up to seventh-order combination) with a periodic-table system, to solve as many cooperation problems.

We can rely on many external and internalized devices to set up expectations, but reasons to mentalize others don’t disappear.

Mutual monitoring, moral norms, institutions, and the mere individual willingness to follow rules don’t guarantee one will meet their expected utility.

In situation of mixed motives, of imperfect information (hidden actions, states or rewards), and/or incomplete (Bayesian) information, one needs to infer causality relationships between actions and intents.

This involves estimating actions, rewards, identities, personalities, preferences, mental states.

Mentalizing, central to strategic interactions

Tomasello, 2008. *MIT Press*.
Tomasello, 2020. *Episteme*.
Sperber et al., 2010. *Mind & Language*.

Cooperative communication

- Aligning beliefs and actions requires mutual **transparency** of minds
- **Language evolved for coordination and shared understanding** (Tomasello, 2008, 2020): sharing mental states, referencing *what is*, building ground truth in common reality
- **But transparency allows exploitation**: free riders gain if they remain undetected

Need for truthfulness

- Stable cooperation requires a) **most communication be truthful**; and b) **truth is the default expectation**
- Yet, this enables strategic deception and exploiting others
- Too many free riders → collapse of cooperation

Dual pressures

- Mentalizing makes cooperation possible and deception feasible
- Lying is effortful because it requires simulating others' minds
- **Communication evolved** under **dual pressures**: **cooperation** through truth-sharing **vs. competition** through manipulation

Epistemic vigilance

- **To protect cooperation, evolution also favored mechanisms for detecting dishonesty** (Sperber, 2010)
- Cognitive systems that assess reliability and sincerity of communicated information, acting as counterweights to gullibility

The development of ToM

Bettle & Rosati, 2021. *Language learning and Development.*
Fujita, Devine, Hughes, 2022. *Cognitive Development.*
Rakoczy, 2022. *Nature Reviews Psychology.*

(shared with other primates)	*(human-unique?)*	**function**
Simple gaze following	Joint attention; sustained attention to others	• Recognize positive intentions • Share attention with others
Goal and intention attribution	Attributing cooperative or shared intentions	• Identify cooperative opportunities • Understand helpful communication • Maintenance so all parties benefit
Level 1 Perspective taking (presence/absence)	Level 2 perspective-taking ('how' it looks)	• 'Bird's-eye view' of joint acts • Facilitate coordination
Attribute knowledge and ignorance	Attribute false beliefs	• Facilitate coordination • Avoid miscommunication • Reputation management

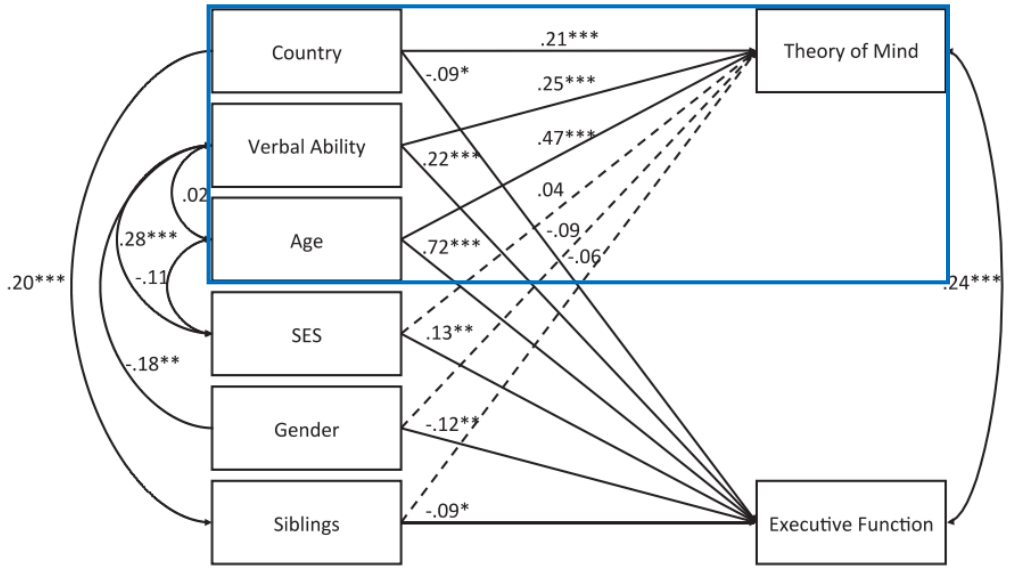
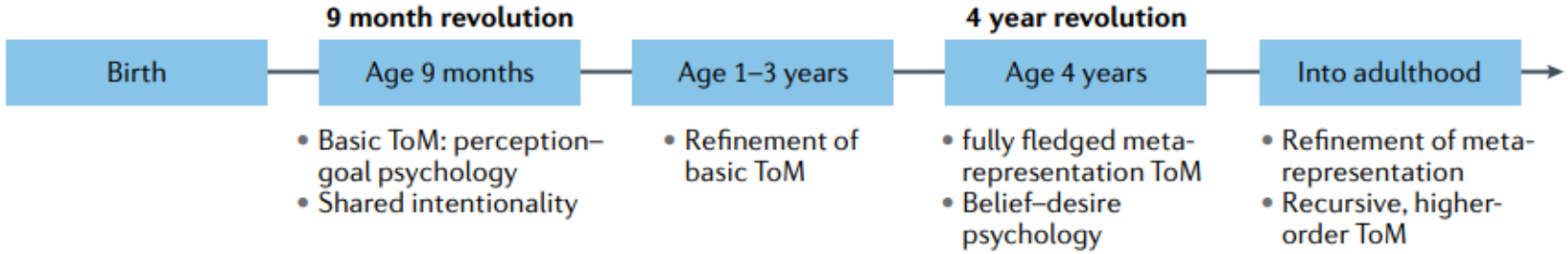
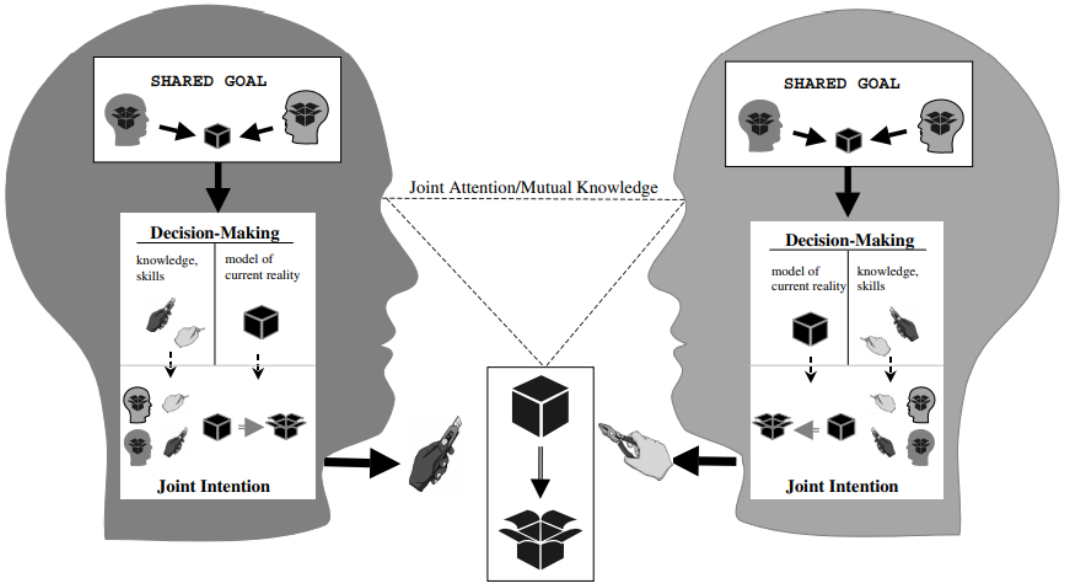


Fig. 1. A structural equation model of cross-cultural differences in theory of mind and executive function with covariates. Country: 0 = Japan, 1 = UK. Gender: 0 = girls, 1 = boys. SES = Socioeconomic Status. *Note.* * $p < .05$. ** $p < .01$. *** $p < .001$.



Understanding and sharing intentions

Tomasello et al., 2005. *Behavioral and brain sciences.*



Human's conception of a collaborative activity (shared goal & joint intention)

Ape general ontogenetic pathway of understanding

Understand others as: a) animate, b) goal-directed and c) intentional agents

Human-specific ontogenetic pathway of motivation

Motivation to share: a) emotions, b) experience and c) activities with others

Human sociality as **intersecting** both pathways: understanding intentional actions & motivation to share states. Intersection between 9mo and 14mo.

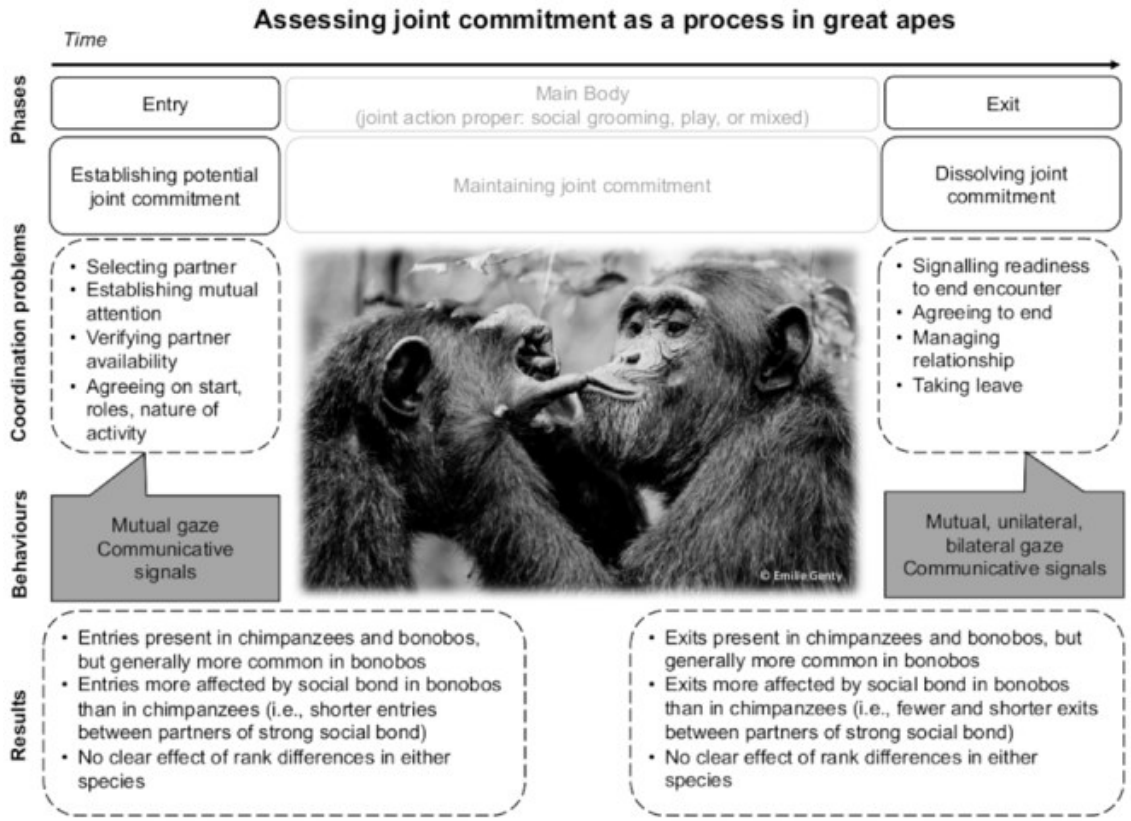
Tomasello et al. proposed that:

- The key distinction of human cognition is the capacity for **shared intentionality**: **participating in collaborative activities through mutually recognized goals and intentions**
- This participation **requires intention reading, cultural learning, and a motivation to share psychological states**
- **As a result, enables representations such as language, norms and collective beliefs**
- Primates grasp individual intentions. Humans integrate individual intentions into *joint commitments*

Heesen et al., 2021. *iScience*.

Joint commitment

Shared sense of obligation between co-actors in a common activity



- **Great apes show proto-commitment behaviors during joint actions** (e.g., play, grooming)
  They coordinate entry, maintenance, and exit phases, using gaze, gestures, and pauses to negotiate participation
- These process-level commitments suggest a **graded evolutionary continuity** with humans
  Bonobos show phases more moderated by friendship than chimpanzees ("face management" –like pattern)
- **Humans display normative, institutionalized forms of commitment** (conventions, roles, agreements) that allow stable, long-term collaboration and cultural accumulation

- Coordination capacities are not strictly uniquely human, but may explain the shift from coordination for mutual benefit to cooperation under shared obligation, and the capacities for complex mentalizing