Does moral play equilibrate?

Jörgen W. Weibull
Stockholm School of Economics

May 4, 2020

1 Introduction

- Results in evolutionary game theory, as well as recent experimental results suggest that we should expect some Kantian moral motivation, alongside pure self-interest, in a range of strategic interactions.

- But does Nash equilibrium always exist if players are partly morally motivated?

- The answer is negative, and the reason is that morality may destroy linearity with respect to randomization, so the "right thing to do" may sometimes be not to randomize.

- This talk presents an analysis of the existence of (pure or mixed) symmetric Nash equilibria in finite and symmetric two-player games when played by partially morally motivated players.

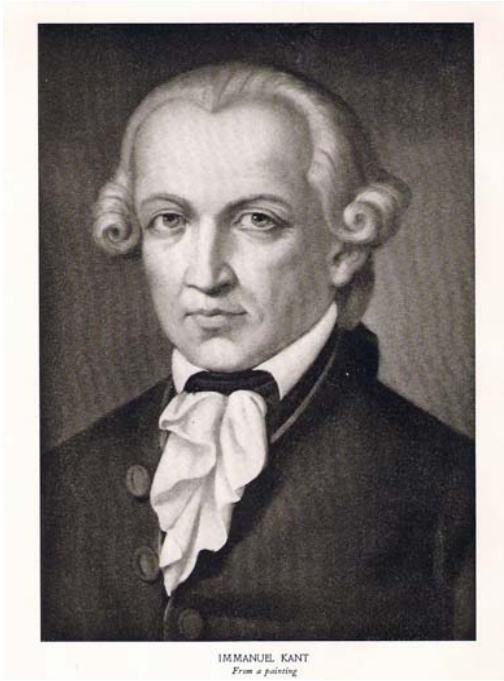
- The analysis is carried out for complete information between equally moral players, and for incomplete information about one's opponent's degree of morality.

- Necessary and sufficient conditions for the existence of equilibrium are given, and the results are illustrated by examples and counter-examples.

Question 1: *What is meant by "Kantian moral motivation" and "partly morally motivated players"?*

Question 2: *Is there any empirical evidence for moral motivation?*

Question 3: *If players are partly morally motivated, do Nash equilibria always exist?*



IMMANUEL KANT
From a painting

"Act only according to that maxim whereby you can, at the same time, will that it should become a universal law." [Immanuel Kant, *Groundwork of the Metaphysics of Morals*, 1785]

1.1 Evolutionarily stable preferences

[Alger & Weibull: "Homo moralis: Preference evolution under incomplete information and assortative matching", *Econometrica* 2013]

[Alger & Weibull: "Homo moralis: Preference evolution under incomplete information and assortative matching", *Games and Economic Behavior* 2016]

- Maynard Smith and Price (Nature, 1973) defined *evolutionary stability* as a property of (pure or mixed) *strategies* in symmetric and finite games

- Let $\pi(x, y)$ be the payoff to a player using mixed strategy $x \in \Delta$ against an opponent using strategy $y \in \Delta$

Definition 1.1 *A strategy $x \in \Delta$ is* **evolutionarily stable** *if, for any other strategy $y \in \Delta$ there exists an $\bar{\varepsilon} \in (0,1)$ such that*

$$\pi\left(x,(1-\varepsilon)\,x+\varepsilon y\right)>\pi\left(y,(1-\varepsilon)\,x+\varepsilon y\right)\quad\forall\varepsilon\in(0,\bar{\varepsilon})$$

Now consider:

- A symmetric and finite game in "material" payoffs, just as in Maynard Smith & Price's model
- Random matching (assumed uniform in MS&P) that may be assortative
- Each individual is endowed with a continuous utility function, $u:\Delta^2\rightarrow\mathbb{R}$

- Each individual's utility function, the individual's *type*, is his or her private information

- Each individual strives to maximize the expected value of his or her personal utility, given the matching protocol and the type distribution in the population

Definition 1.2 *A type* u *is* **evolutionarily stable against type** v *if* $\exists\ \bar{\varepsilon} > 0$ *such that individuals of type* u *earn a higher material payoff than individuals of type* v *in* **all** *(Bayesian) Nash equilibria in all population states* $s = (u, v, \varepsilon)$ *with* $\varepsilon \in (0, \bar{\varepsilon})$*.*

Definition 1.3 *A type* u *is* **evolutionarily unstable** *if there exists a type* v *such that* $\forall\ \bar{\varepsilon} > 0$ *there exists a population state* $s = (u, v, \varepsilon)$ *with* $\varepsilon \in (0, \bar{\varepsilon})$ *and* **some** *(Bayesian) Nash equilibrium in which type* v *earns a higher material payoff than type* u*.*

- A particular family of utility functions:

Definition 1.4 Homo moralis *preferences of degree of morality* $\kappa \in [0,1]$ *are given by the utility function,* $u_\kappa : \Delta^2 \to \mathbb{R}$ *defined by*

$$u_\kappa(x,y) = (1-\kappa)\cdot\pi(x,y) + \kappa\cdot\pi(x,x).$$

- A convex combination of pure *selfishness* (Adam Smith) and *Kantian morality.*

- Main result (glossing over what exactly is meant by "behaviorally distinct"):

Theorem 1.1 *The HM type u_κ, with κ equal to the index of assortativity, is evolutionarily stable against all behaviorally distinct types v. All types u that are behaviorally distinct from u_κ are evolutionarily unstable.*

- Proof topological, based on the *upper hemi-continuity* of the Nash equilibrium correspondence at $\varepsilon = 0$

Question 2: *Is there any empirical evidence for Homo moralis preferences?*

1.2 Experimental evidence

1.2.1 Experiment 1

[Miettinen, Kosfeld, Fehr & Weibull: "Revealed preferences in a sequential prisoners' dilemma: A horse race between six utility functions", *Journal of Economic Behavior and Organization*, 2020]

- Laboratory experiment with 98 master students in Zürich

- One sequential prisoners' dilemma interaction

- Main result:

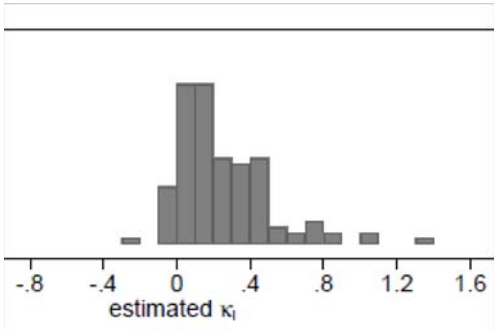
TABLE 4: Analysis with 4 homogeneous groups.

model	**hit rate**	**Selten-Krischker score**
Homo oeconomicus	0.28	0.16
Inequity aversion	0.53	0.28
Conditional welfare	0.76	0.26
Reciprocity	0.76	0.26
Altruism	0.40	0.02
Homo moralis	0.70	0.33

1.2.2 Experiment 2

[van Leeuwen, Alger & Weibull: "Estimating social preferences and Kantian morality in strategic interactions", WP, 2020]

- Laboratory experiment with 136 master students in Tilburg. Three classes of simple sequential two-player games, each class represented by 6 monetary specifications



Question 3: *If players are partly morally motivated, in line with Homo moralis, do Nash equilibria always exist?*

- This question will now be analyzed, for finite and symmetric two-player games, in two information settings:
  - Complete information between equally moral players
  - Incomplete information among moral players drawn from a (morally) heterogeneous population

2 Equilibria among Homines morales

[Bomze, Schachinger & Weibull: "Does moral play equilibrate?", *Economic Theory* 2020]

2.1 Notation and preliminaries

- Let $S = \{1, ..., m\}$ be the set of pure strategies, and let Δ be the associated unit simplex of mixed strategies,

$$\Delta = \left\{ x \in \mathbb{R}^m_+ : e^T x = \sum_{i=1}^{m} x_i = 1 \right\},$$

where $e = \sum_{i=1}^{m} e_i$ and each e_i is the i:th unity (column) vector, and the superscript T denotes transpose

- Let A be an $m \times m$-matrix with "material" payoffs, and write $\pi : \Delta^2 \to \mathbb{R}$ for the "material" payoff function, with the convention that $\pi(x, y)$ is the "material" payoff to a player using mixed strategy $x \in \Delta$ against an opponent using mixed strategy y, where x and y are (column) vectors in Δ:

$$\pi(x, y) = x^T A y$$

- Let $\theta \in \Theta = [0, 1]$ be a *player type*, and consider the associated payoff function $u_\theta : \Delta^2 \to \mathbb{R}$, defined by

$$u_\theta(x, y) = (1 - \theta) x^T A y + \theta x^T A x, \tag{1}$$

  This is the utility function of a *Homo moralis* with degree of morality θ

- The game being symmetric, $B = A^T$ is the matrix of material payoffs to the column player

- "Material" welfare is defined as the sum of both players' material payoffs: $x^T \left(A + A^T\right) y$

- When both players use the same strategy, let $W : \Delta \to \mathbb{R}$ be defined by

$$W(x) = 2x^T Ax$$

- Thus:

$$u_\theta(x, y) = (1 - \theta) x^T Ay + \frac{\theta}{2} W(x)$$

- Let $\beta_\theta : \Delta \rightrightarrows \Delta$ be the best-reply correspondence of *Homo moralis* of degree θ:

$$\beta_\theta(y) = \arg\max_{x \in \Delta} u_\theta(x, y) \qquad \forall y \in \Delta .$$

- By Weierstrass' maximum theorem, $\beta_\theta(y) \subseteq \Delta$ is a non-empty and compact set for every $\theta \in [0,1]$ and $y \in \Delta$

- However, as will be seen shortly, this set is not always convex

- Write

$$X_\theta = \{x \in \Delta : x \in \beta_\theta(x)\}$$

  Pairs $(x,x) \in \Delta^2$ with $x \in X_\theta$ are thus the *symmetric Nash equilibria* between two *Homines morales* of the same degree of morality θ

- By Berge's maximum theorem, β_θ is upper hemi-continuous with respect to y and θ. As we saw above it is also compact valued. However, as will be seen shortly, it is not always convex valued

2.2 Equally moral players

Proposition 2.1 *The set X_θ is non-empty and compact if $\theta \in \{0, 1\}$. The same is true for all $\theta \in [0, 1]$ if W is concave.*

- The first claim follows from standard arguments. The second claim follows immediately from the Glicksberg, Fan, Kakutani fixed-point theorem.

- A sufficient condition for W to be concave is that the symmetric matrix $A + A^T$ is negative semidefinite

- Proposition below 2.5 provides necessary and sufficient conditions for W to be concave (or strictly convex)

Example 2.1 *Consider the coordination game*

$$A = \begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix}$$

for $a > b > 0$. *Clearly* $X_0 = \{e_1, e_2, x^*\}$, *where*

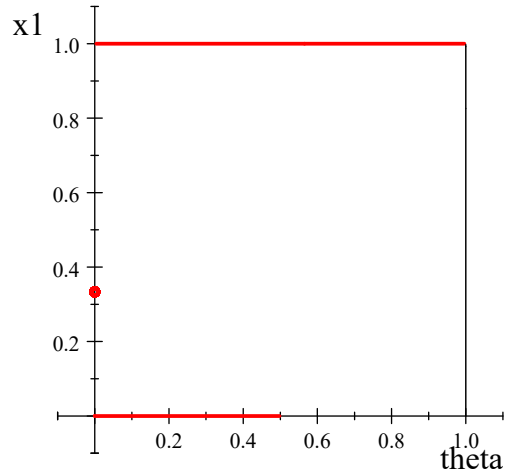
$$x^* = \begin{pmatrix} b/(a+b) \\ a/(a+b) \end{pmatrix}$$

We note that W *is strictly convex:* $W(x) = 2ax_1^2 + 2bx_2^2$. *Hence,* $u_\theta(x, y)$ *is strictly convex in* x *for any given* $\theta > 0$ *and* $y \in \Delta$, *and* $\beta_\theta(y) \subseteq \{e_1, e_2\}$ *for all* $\theta > 0$

Clearly $e_1 \in \beta_\theta(e_1)$ $\forall \theta \in [0, 1]$

It is easily verified that $e_2 \in \beta_\theta(e_2)$ *iff* $\theta \leq b/a$.

At $\theta = b/a$, $\beta_\theta(e_2)$ *is a binary set. For all other* θ, *both* $\beta_\theta(e_1)$ *and* $\beta_\theta(e_2)$ *are singletons:*



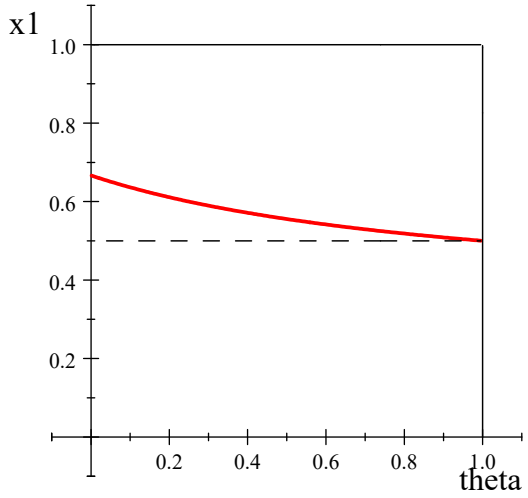
(drawn for $a = 2$ and $b = 1$). Both pure equilibria are robust to a small degree of morality, but the mixed equilibrium is not.

Example 2.2 *Consider the hawk-dove game*

$$A = \begin{pmatrix} 0 & a \\ b & 0 \end{pmatrix}$$

for $a > b > 0$. *Then* $A + A^T$ *is indefinite, but* W *is strictly concave:* $W(x) = 2(a+b)\,x_1(1-x_1)$. *Hence* $u_\theta(x,y)$ *is concave in* x *so there exist at least one fixed point. It is easily verified that the unique fixed point is*

$$x^*_\theta = \begin{pmatrix} (a+\theta b)/(a+b) \\ (b+\theta a)/(a+b) \end{pmatrix}$$



(drawn for $a = 2$ and $b = 1$). The aggressive hawk strategy is used less the more moral the players are.

Proposition 2.2 *Let A be the payoff matrix of a symmetric constant-sum game. For any $\theta < 1$, the set of fixed points is identical with the non-empty set of fixed points when $\theta = 0$, while every $x \in \Delta$ is a fixed point when $\theta = 1$.*

- In other words, all *Homines morales*, except *Homo kantiensis*, behave like *Homo oeconomicus* in all (finite and symmetric two-player) constant-sum games

- The remaining situation to investigate is thus when $\theta > 0$ and W is not concave (as in Example 1)

- We begin by considering an example

Example 2.3 *Consider the generalized Rock-Scissors-Paper (RSP) game matrix*

$$A = \begin{pmatrix} 1 & 2+a & 0 \\ 0 & 1 & 2+a \\ 2+a & 0 & 1 \end{pmatrix}$$

for any $a > -1$. *A constant-sum game iff* $a = 0$. *For* $\theta = 0$, *the unique symmetric Nash equilibrium strategy is the barycenter* $x^o = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$. *As is well-known, this unique equilibrium is unstable in the replicator dynamic if* $a < 0$ *and asymptotically stable if* $a > 0$. *The function* W *is strictly convex if* $a < 0$ *and concave if* $a > 0$. *For any* $x \in \Delta$:

$$W(x) = 2 + a \cdot \left(1 - \|x\|^2\right) .$$

Assume $a < 0$*, fix* $0 < \theta < 1$ *and note that* $\varnothing \neq \beta_\theta(y) \subseteq \{e_1, e_2, e_3\}$ *for all* $y \in \Delta$*. Moreover,* $u_\theta(e_i, e_i) = 1$ *for all* $i \in S$*, and*

$$u_\theta(e_1, e_2) = u_\theta(e_2, e_3) = u_\theta(e_3, e_1) = (1 - \theta)(2 + a) + \theta\,.$$

Hence, no vertex e_i *is a fixed point. Consequently:*

There exists no fixed point if $-1 < a < 0$ *and* $0 < \theta < 1$*.*

More generally:

Proposition 2.3 *If* W *is strictly convex, then* $\beta_\theta(y) \subseteq \{e_1, ..., e_m\}$ *for all* $y \in \Delta$ *and* $\theta > 0$*, and* $e_i \in \Delta$ *is a fixed point under* β_θ *iff*

$$a_{ii} \geq (1 - \theta)\, a_{ki} + \theta a_{kk} \quad \forall k \in S$$

- The usefulness of Propositions 2.1 and 2.3 depends on how easy or hard it is to verify that the welfare function is either concave or strictly convex on the unit simplex.

Proposition 2.4 *Let C be the expansion of the $(m-1) \times (m-1)$ identity matrix to an $(m-1) \times m$ matrix obtained by appending the column $(-1, -1, ..., -1)^T \in \mathbb{R}^{m-1}$. Then W is concave (strictly convex) iff the symmetric $(m-1) \times (m-1)$ matrix*

$$D = C\left(A + A^T\right)C^T$$

is negative semidefinite (positive definite).

- In some applications the payoff matrix A is symmetric; $A^T = A$. In such *potential* games, average payoff increases along all solution trajectories to the replicator dynamic.

- For such games and any positive degree of morality, any global welfare maximizer is a fixed point, and every fixed point is a local maximizer:

Proposition 2.5 *Suppose* $A^T = A$ *and* $\theta > 0$*. Then*

(a) $\arg\max_{z \in \Delta} W(z) \quad \subseteq \quad X_\theta$

(b) $x \in X_\theta \quad \Longrightarrow \quad x \in \arg\max_{z \in \Delta \cap U} W(z)$ *for some neighborhood* U *of* x.

2.3 Incomplete information about others' morality

- We now consider strategic interactions between two *Homines morales* who only know their own degree of morality, not that of the opponent

- We endow the type space Θ with its Euclidean topology and let μ be a Borel probability measure on Θ, representing the type distribution in the population

Definition 2.1 *A* **strategy** *is a Borel-measurable function* $\xi : \Theta \to \Delta$*, assigning to each type* $\theta \in \Theta$ *a strategy* $\xi(\theta) \in \Delta$*.*

- A strategy ξ is *optimal* against a mixed strategy $y \in \Delta$ if

$$\xi(\theta) \in \arg\max_{x \in \Delta} u_\theta(x, y) \qquad \forall \theta \in \Theta.$$

- It follows from measurable selection theory à la Kuratowski-Ryll-Nardzewski that such an optimal strategy $\xi : \Theta \to \Delta$ exists for each $y \in \Delta$.

Definition 2.2 *A strategy $\xi : \Theta \to \Delta$ is a best reply to itself, or, equivalently, (ξ, ξ) is a symmetric* **Nash equilibrium** *under incomplete information, if*

$$\xi(\theta) \in \arg\max_{x \in \Delta} \int_{\Theta} u_\theta(x, \xi(\tau))\, d\mu(\tau) \qquad \forall \theta \in \Theta \tag{2}$$

- By linearity of $u_\theta(x, y)$ with respect to y:

$$\int_{\Theta} u_\theta(x, \xi(\tau))\, d\mu(\tau) = u_\theta\left(x, \overline{\xi}\right)$$

where

$$\overline{\xi} = \mathbb{E}_\mu[\xi(\theta)] = \int_{\Theta} \xi(\theta) d\mu(\theta)$$

is the *representative agent*'s mixed strategy.

- Hence: A strategy $\xi : \Theta \longrightarrow \Delta$ is a best reply to itself iff it is optimal against its own representative agent's mixed strategy

- Existence of symmetric Nash equilibria (ξ, ξ) is non-trivial

- However, one may characterize NE by way of first- and second-order optimality conditions:
  - a first-order (Lagrangian) condition
  - a complementary slackness condition
  - a second-order (curvature) condition (significantly different from standard)

- In order to state the result:

- Let $H(\theta) = \theta\left(A + A^T\right)$, the Hessian of $u_\theta(\cdot, y)$
- Let $g(\theta) = H(\theta)\xi(\theta) + (1-\theta)A\bar{\xi}$, the gradient of $u_\theta(\cdot, y)$ evaluated at $x = \xi(\theta)$
- Let $H_i(\theta) = e_i g^T(\theta) + g(\theta) e_i^T - \xi_i(\theta) H(\theta)$, a rank-two update of $H(\theta)$
- Let $\Gamma_i(\theta) = \left\{v \in e^\perp : \xi_i(\theta) v_j \geq \xi_j(\theta) v_i \ \forall j \in S\right\}$ a polyhedral cone

Theorem 2.6 *For any Borel probability measure μ on Θ, a strategy $\xi : \Theta \to \Delta$ is a best reply to itself iff there are Borel-measurable functions $\lambda_i : \Theta \to \mathbb{R}$ for $i \in \{0, 1, .., m\}$ such that, for all $i \in S$ and $\theta \in \Theta$:*

$$[H(\theta)\xi(\theta)]_i + (1-\theta)[A\bar{\xi}]_i + \lambda_0(\theta) + \lambda_i(\theta) = 0 \tag{3}$$

$$\lambda_i(\theta)\xi_i(\theta) = 0 \tag{4}$$

$$v^T H_i(\theta) v \geq 0 \quad \forall v \in \Gamma_i(\theta) \text{ if } \xi_i(\theta) > 0 \tag{5}$$

- In Example 2.3 we noted that no symmetric Nash equilibrium exists under complete information between equally moral players when $-1 < a < 0$ and $0 < \theta < 1$.

- Formally, such a situation can be represented as incomplete information with a unit Dirac measure placed on a particular type $\theta \in (0, 1)$.

- Consider instead any continuous type distribution μ on $\Theta = [0, 1]$. We may then divide the type space into three disjoint intervals I_k with $\mu(I_k) = 1/3$, for $k = 1, 2, 3$. If all types in I_k play pure strategy k, then all types $\tau \in \Theta$ best respond to $\bar{\xi} = x^o$, the barycenter of the strategy simplex.

- Hence, the non-existence of equilibrium under complete information and equally moral players is, in this example, non-robust to arbitrarily small degrees of incomplete information about morality (as measured in the L^1-norm)

- More generally:

Proposition 2.7 *Suppose that W is convex and that μ is absolutely continuous with PDF $f : \Theta \to \mathbb{R}_+$. If there exists a strategy $x \in \Delta$ whose support coincides with the sets of pure best replies to x for all types θ and τ in the support of f, then there exists a NE (ξ, ξ), and $\bar{\xi} = x$.*

- We note that the required identity between the three subsets of pure strategies holds generically for all θ in some open set $U \subset \Theta$

- The hypothesis of Proposition 2.7 is met in Example 2.3 for $x = x^o$

3 Conclusion

I hope to have answered:

Question 1: *What is meant by "Kantian moral motivation" and "partly morally motivated players"*?

I believe I have answered *positively*:

Question 2: *Is there any empirical evidence for moral motivation?*

And I have answered *negatively*, with necessary and sufficient conditions for existence under complete and incomplete information:

Question 3: *If players are partly morally motivated, do Nash equilibria always exist?*

Thanks for your attention!