John Nash meets Immanuel Kant: moral motivation in strategic interactions

Jörgen Weibull
Stockholm School of Economics

July 23, 2020

1 Introduction

John Nash (1950):

"One may define a concept of *an* n*-person game* in which each player has a finite set of pure strategies and in which a definite set of *payments* to the n players corresponds to each n-tuple of pure strategies, one strategy being taken for each player... Any n-tuple of strategies ... *counters* another if the strategy of each player in the countering *n*-tuple yields the highest obtainable expectation for its player against the $n-1$ strategies of the other players in the countered n-tuple. A self-countering n-tuple is called an *equilibrium* point."

- A world where each individual strives to maximize his or her "payment".

Immanuel Kant (1785):

"Act only according to that maxim whereby you can, at the same time will that it become a universal law."

- A world where each individual strives to "do the right thing" in the given situation

- Do these world-views agree, disagree, can they be reconciled, or are they incompatible?
  - Nash's view is *individualistic* and *consequentialistic*: "What strategy or strategies lead to the best consequences in terms of my "payment"?
  - Kant's view is *universalistic* and *deontological*: "What strategy or strategies is it my, or anybody's, duty to use in this situation?"

- Kant's (1785) lead example:

  "May I, when in distress, make a promise with the intention not to keep it?"

  - His answer:

  "The shortest way... to discover ... whether a lying promise is consistent with duty, is to ask myself if I would be able to say to myself: "Every one may make a deceitful promise when he finds himself in a difficulty from which he cannot otherwise extricate himself"? ... While I can will the lie, I can by no means will that lying should be a universal law. For with such a law there would be no promises at all... Hence my maxim, as soon as it should be made a universal law, would necessarily destroy itself."

- How does this deontological reasoning relate to Nash's consequentialistic approach? That depends, to a large extent what meaning we attach to Nash's term "payment"

- A possible reconciliation between Nash and Kant:
  - Consider symmetric games and use Nash equilibrium (or a refinement thereof) as the solution concept, with a "pure Kantian" player's "payment" defined in terms of the "material consequences" for the players if the other player would use the same strategy as oneself

1.1 Kant's "promise game"

1. Nature flips a fair coin. If "heads", then player 1 falls in distress, if "tail" then player 2 falls in distress

2. The distressed player may or may not ask for a loan - with a payback promise - from the other player

3. The other player may or may not give the loan

4. If the loan is given, then it may or may not be paid back by the loan-taker

- Let the associated "material payoffs" be -1 for the player in distress if no loan is given, 0 for the lender if the loan given but not paid back, 1 for the distressed player if the loan is given and paid back, and let 2 be the "material payoff" to player 2 if not giving the loan or if the loan is given and paid back

- Purely self-interested players: Define "payment" to strategy x, when used against strategy y, as own material payoff. The unique NE outcome is that no loan is given. The expected material payoff to each player is $1/2$.

- Pure Kantian players: Define "payment" to strategy x, when used against strategy y, as own material payoff *if also the other player would use strategy* x. The unique Nash equilibrium is to lend, trust, and honor. The expected material payoff to each player is $3/2$.

- Experimental data from trust games suggest something inbetween

1.2 Morality and economics

These two topics were more intertwined in classical economics than today. Some references:

- Smith (1759), Edgeworth (1881), Arrow (1973), Laffont (1975), Sen (1977), Bacharach (1999), Tabellini (2008), Sugden (2011), Alger and Weibull (2013, 2016, 2017), Romer (2015, 2020), Miettinen, Kosfeld, Fehr and Weibull (2020), Alger, Weibull & Lehmann (2020), Bomze, Schachinger & Weibull (2020)

2 Model

[Alger & Weibull, 2013]

- Consider finite and symmetric two-player games
  - The set of pure strategies: $S = \{1, ..., m\}$
  - The set of mixed strategies: Δ (the unit simplex)
  - "Material" payoff function $\pi : \Delta^2 \to \mathbb{R}$, where

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

    is the "material" payoff to a player using mixed strategy $x \in \Delta$ against an opponent using mixed strategy $y \in \Delta$, where A is the $m \times m$ payoff matrix

- Define the payoff or *utility* function $u_\kappa : \Delta^2 \to \mathbb{R}$ by

$$u_\kappa(x, y) = (1 - \kappa)\,\pi(x, y) + \kappa\pi(x, x) \tag{1}$$

  for some $\kappa \in [0, 1]$, the player's *degree of morality*.

  - Here $\kappa = 0$ is *Homo oeconomicus*, $\kappa = 1$ *Homo Kantiensis*. For any $\kappa \in [0, 1]$, this is the utility function of a *Homo moralis* with degree of morality κ

- Let $\beta_\kappa : \Delta \rightrightarrows \Delta$ be the best-reply correspondence of *Homo moralis* of degree of morality κ, and write

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

- Thus $x \in X_\kappa$ iff (x, x) is a *symmetric NE* between two *Homines morales* with the same degree of morality κ

Question 1: *Is there any empirical evidence for Homo moralis?*

Question 2: *Is there any theoretical foundation for Homo moralis?*

Question 3: *Do symmetric Nash equilibria between Homine morales always exist?*

3 Empirical evidence

[Miettinen, Kosfeld, Fehr & Weibull, 2020]

- Laboratory experiment with 98 master students in Zürich
  - A sequential prisoners' dilemma interaction, preceded by fair coin toss
  - Belief elicitation, the strategy method. No equilibrium assumption.
  - Main result:

TABLE 4: Analysis with 4 homogeneous groups.

preference 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

4 Theoretical foundation

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

- Alger & Weibull (2013, 2016) extended this definition to a property of payoff or utility functions in (finite or infinite) symmetric two-player games, when each individual's utility/payoff function is his or her private information

- Anonymous random matching, which may be uniform or assortative, in a large population. Index of assortativity $\sigma \in [0, 1]$

- The main result (for any compact and convex strategy space X):

Theorem 4.1 (Alger & Weibull, 2013) *Homo moralis with degree of morality $\kappa = \sigma$ is evolutionarily stable against all behaviorally distinct continuous utility functions. Every continuous utility function u that is behaviorally distinct from u_σ is evolutionarily unstable.*

- This result is extended to symmetric n-player games in Alger & Weibull (2016)

- The result is corroborated in a stochastic model of preference evolution in spatially structured populations by Alger, Weibull & Lehmann (2020)

5 Equilibrium play

[Bomze, Schachinger & Weibull, 2020]

- Finite and symmetric two-player games in two distinct information settings:
  - Complete information between equally moral players
  - Incomplete information among moral players drawn from a (morally) heterogeneous population

- Let $W(x) = 2\pi(x,x) = 2x^T Ax$, the (expected) "material" welfare if both play $x \in \Delta$

Proposition 5.1 *The set X_κ is non-empty if $\kappa \in \{0,1\}$. The same is true for all $\kappa \in [0,1]$ if $W : \Delta \to \mathbb{R}$ is concave.*

Example 5.1 *Consider the coordination game*

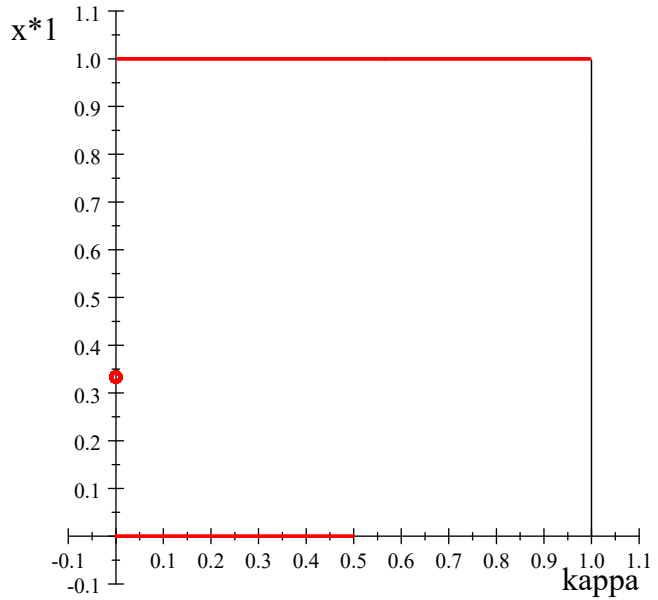
$$A = \begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix}$$

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

$$x^* = \left(\frac{b}{a+b}, \frac{a}{a+b}\right)$$

We note that W is strictly convex. Hence, $u_\kappa(x,y)$ is strictly convex in x for any $\kappa > 0$. Moreover: $\beta_\kappa(y) \subseteq \{e_1, e_2\}$ for all $\kappa > 0$. It is easily verified that $e_2 \in \beta_\kappa(e_2)$ iff $\kappa \leq b/a$.

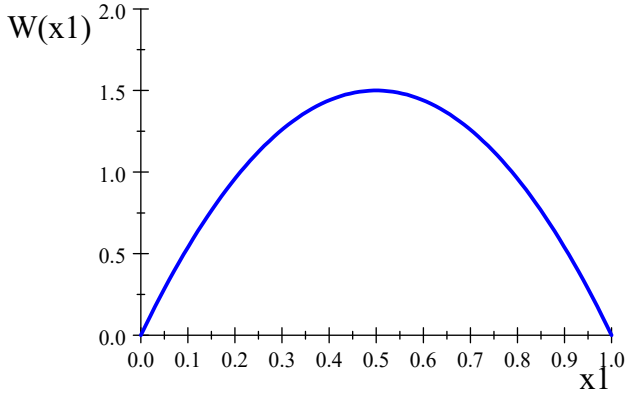
The diagram below shows $x_1^*(\kappa)$ *for* $a = 2$, $b = 1$.



Example 5.2 *Consider the anti-coordination game*

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

for $a > b > 0$. *Then* W *is strictly concave on* Δ



Hence there exist at least one symmetric NE. In fact, it is unique:

$$x^*(\kappa) = \left(\frac{a + \kappa b}{(1+\kappa)(a+b)}, \frac{\kappa a + b}{(1+\kappa)(a+b)} \right)$$

Pure Kantian players ($\kappa = 1$), *randomize 50/50. This maximizes material welfare across symmetric strategy profiles, and the maximum value is*

$\frac{1}{2}(a+b)$. *For players with degree of morality* $k < 1$ *less material welfare is obtained in equilibrium.*

**For* $\kappa \leq a/(2a+b)$ *also* $x_1^* = 1$ *and* $y_1^* = 0$ *is a NE, and for* $\kappa \leq b/(2b+a)$ *also* $x_1^* = 0$ *and* $y_1^* = 1$ *is a NE.*

***By adding a public randomization device that assigns player roles with equal probability, one obtains a symmetric augmented game with a symmetric NE with material payoff* $a+b$.

5.1 Constant-sum games

Proposition 5.2 *Suppose that the game is a constant-sum game. For any* $\kappa < 1$*, the set of fixed points is identical with the non-empty set of fixed points when* $\kappa = 0$*, while every* $x \in \Delta$ *is a fixed point when* $\kappa = 1$*.*

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

5.2 Games with convex welfare functions

Example 5.3 *Consider the generalized Rock-Scissors-Paper game*

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

where $a + 1 > 0$. *A constant-sum game iff* $a = 0$. *For* $\kappa = 0$, *the unique symmetric Nash equilibrium strategy is* $x_0^* = (1/3, 1/3, 1/3)$. *This 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* $a > 0$.

Let $a < 0$ *and* $\kappa \in (0, 1)$. *Then* $\varnothing \neq \beta_\kappa(y) \subseteq \{e_1, e_2, e_3\}$ *for all* $y \in \Delta$. *Moreover,*

$$u_\kappa(e_3, e_1) = (1 - \kappa)(2 + a) + \kappa > 1$$

Hence, there exists no fixed point: $X_\kappa = \varnothing$.

More generally:

Proposition 5.3 *If W is strictly convex, then $\beta_\kappa(y) \subseteq \{e_1, ..., e_m\}$ for all $y \in \Delta$ and $\kappa > 0$. Moreover, $e_i \in X_\kappa$ iff*

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

- The usefulness of the above results depends on how easy or hard it is to verify that the welfare function W is either concave or strictly convex on the unit simplex.

Proposition 5.4 *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) on Δ iff the symmetric $(m-1) \times (m-1)$ matrix*

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

is negative semidefinite (positive definite).

5.3 Incomplete information about others' morality

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

- Let μ be a Borel probability measure on the type space $\Theta = [0, 1]$, the type distribution

Definition 5.1 *A* **strategy** *is a Borel-measurable function* $\xi : \Theta \to \Delta$*, assigning to each type* $\kappa \in \Theta$ *a strategy* $\xi(\kappa) \in \Delta$*.*

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

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

- 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 5.2 *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*

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

- But, by linearity of $u_{\kappa}(x, y)$ with respect to y:

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

where

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

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 NE (ξ, ξ) is still non-trivial. However, since the utility/payoff functions are linear-quadratic, one may obtain necessary and sufficient conditions for a strategy to be a best reply to itself (in terms of first- and second-order optimality conditions)

- See Theorem 1 in Bomze, Schachinger & Weibull (2020).

6 Conclusion

- I hope to have shown that the approaches by Nash and Kant, different as they are, may nevertheless be reconciled and combined

- I also hope to have given answers to:

Question 1: *Is there any empirical evidence for Homo moralis?*

Question 2: *Is there any theoretical foundation for Homo moralis?*

Question 3: *Do symmetric Nash equilibria between Homine morales always exist?*

- Many more interesting, fun and challenging questions remain to be answered. You are all most welcome to join in on this exciting research agenda!

Post-seminar comment:

- In Example 5.3 we noted that no symmetric Nash equilibrium exists under complete information between equally moral players when $-1 < a < 0$ and $0 < \kappa < 1$.

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

- 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 $\overline{\xi} = x_0^*$, the barycenter of the strategy simplex.

- Hence, in that example the non-existence of equilibrium under complete information and equally moral players is non-robust to arbitrarily small degrees of incomplete information about morality