Evolution and Kantian morality: A correction and addendum

Ingela Alger, Jörgen Weibull

Evolution and Kantian morality: a correction and addendum

Ingela Alger* and Jörgen Weibull†

June 26, 2022. Revised January 13, 2023

Abstract. Theorem 1 in Alger and Weibull (Games and Economic Behavior, 2016) consists of two statements. The first establishes that *Homo moralis* with the right degree of morality is evolutionarily stable. The second statement is a claim about sufficient conditions for other goal functions to be evolutionarily unstable. However, the proof given for that claim presumes that all relevant sets are non-empty, while the hypothesis of the theorem does not guarantee that. We here prove instability under a stronger hypothesis that guarantees existence, and we also establish a new and closely related result. As a by-product, we also obtain an extension of Theorem 1 in Alger and Weibull (Econometrica, 2013).

Keywords: C73, D01, D03.

JEL codes: Preference evolution, evolutionary stability, morality, Homo moralis.

Theorem 1 in Alger and Weibull (2016), henceforth AW, consists of two statements. The first establishes that *Homo moralis* with the right morality profile is evolutionarily stable. The second statement is a claim about sufficient conditions for other goal functions to be evolutionarily unstable: "Any $f \in F$ with $X(f) \cap X(f^*) = \varnothing$ is evolutionarily unstable". Here f^* is the goal function of *Homo moralis* with the right morality profile (that is, identical with the assortativity profile of the matching process), $X(f)$ (resp. $X(f^*)$) is the set of strategies $x \in X$ that are best replies to themselves with respect to goal function f (resp. f^*), and a goal function $f \in F$ is *evolutionarily unstable* if there exists another goal function $g \in F$ such that there for every $\bar{\varepsilon} > 0$ exists a smaller but positive mutant population share ε and at least one associated Nash equilibrium in which the mutants earn a higher material payoff than the residents. However, the proof given for the second claim presumes that all relevant sets are non-empty and that $X(f)$ is a singleton set, while the hypothesis of

*I.A. acknowledges IAST funding from the French National Research Agency (ANR) under grant ANR-17-EURE-0010 (Investissements de l'avenir program). I.A. also acknowledges funding from the European Research Council (ERC) under the European Union Horizon 2020 research and innovation programme (grant agreement No 789111 - ERC Evolving Economics).

†J.W. acknowledges financial support from the Jan Wallander and Tom Hedelius Foundation.

the theorem does not guarantee this. Here, we provide new sufficient conditions for a goal function f to be evolutionarily unstable. As a by-product, we also obtain an extension of Theorem 1 in Alger and Weibull (2013).

Since the definition of instability in AW requires equilibrium existence, we first ensure this in order to prove instability. Throughout, we therefore make the following assumption:[1]

> *Assumption:* The material-payoff function π is such that $f^*(x, y)$ is concave in $x \in X$, the own (potentially multi-dimensional) strategy, for any strategy $y \in X$ used by an opponent.

Let $F^c \subset F$ denote the subset of goal functions that are concave with respect to their first argument, $x \in X$.

Lemma 1. *If $f \in F^c$, then*

1. *$X(f)$ is non-empty,*
2. *$B^{NE}(f, f^*, \varepsilon) \neq \varnothing$ for all $\varepsilon \in (0, 1)$,*
3. *the correspondence $B^{NE}(f, f^*, \cdot) : (0, 1) \rightrightarrows X^n$ is u.h.c. and compact-valued.*

Proof: The first two claims follow from the Kakutani-Glicksberg-Fan fixed-point theorem, since f^* and f are continuous and concave in their first argument, and X is a nonempty, convex and compact set in a normed vector space (see Corollary 17.55 in Aliprantis and Border, 2006). The third statement follows from Berge's maximum theorem (see Theorem 17.31, op. cit.). **Q.E.D.**

We are now in a position to provide the new sufficient conditions for evolutionary instability of goal functions:

Proposition 1. *Any goal function $f \in F^c$ for which $X(f) \cap X(f^*) = \varnothing$ is evolutionarily unstable.*

Proof: Consider any $f \in F^c$. The non-emptiness of $X(f)$ implies that $B^{NE}(f, f^*, 0)$ is non-empty too, since $(x^*, y^*) \in B^{NE}(f, f^*, 0)$ if and only if $x^* \in X(f)$ and

$$y^* \in \arg\max_{y \in X} f^*(y, (x^*, ..., x^*)),$$

[1] Proposition 4 in Bomze et al. (2021) provides necessary and sufficient conditions for the required concavity property of the *Homo moralis* goal function f^* when applied to the mixed-strategy extension of finite two-player games in material payoffs.

where the latter set is non-empty by Weierstrass' maximum theorem (f^* is continuous and X is non-empty and compact). Hence, the domain of the u.h.c. correspondence $B^{NE}(f, f^*, \cdot)$ can be extended to include $\varepsilon = 0$.

Let $(x^*, y^*) \in B^{NE}(f, f^*, 0)$. Then $x^* \notin \arg\max_{y \in X} f^*(y, (x^*, ..., x^*))$, since otherwise $x^* \in X(f^*)$, contradicting the hypothesis $X(f) \cap X(f^*) = \varnothing$. Thus

$$\Pi_R(x^*, y^*, 0) = f^*(x^*, (x^*, ..., x^*)) < f^*(y^*, (x^*, ..., x^*)) = \Pi_M(x^*, y^*, 0).$$

Let the function $D : X^2 \times [0,1] \to \mathbb{R}$ be defined by $D(x, y, \varepsilon) = \Pi_M(x, y, \varepsilon) - \Pi_R(x, y, \varepsilon)$. Then $D(x, y, \varepsilon) > 0$ for all $(x, y) \in B^{NE}(f, f^*, 0)$. Since $\varnothing \neq B^{NE}(f, f^*, 0) \subseteq X^2$ is compact and the function D is continuous, there exists, by Weierstrass' maximum theorem, a $\delta > 0$ such that $D(x, y, 0) \geq \delta$ for all $(x, y) \in B^{NE}(f, f^*, 0)$. Again by continuity of D, there exists an $\bar{\varepsilon} > 0$ such that $D(x, y, \varepsilon) \geq \delta/2$ for all $(x, y, \varepsilon) \in U \times [0, \bar{\varepsilon}]$ where $U \subset X^2$ is the $\bar{\varepsilon}$-neighborhood of the compact set $B^{NE}(f, f^*, 0) \subset X^2$. Since $B^{NE}(f, f^*, \cdot) : [0,1] \to X^n$ is u.h.c., $\varnothing \neq B^{NE}(f, f^*, \varepsilon) \subseteq U$ for all $\varepsilon \in [0, \bar{\varepsilon},]$ sufficiently small. In sum: for all small $\varepsilon > 0$ there exist equilibria $(x, y) \in B^{NE}(f, f^*, \varepsilon)$, and in all those equilibria $\Pi_R(x, y, \varepsilon) < \Pi_M(x, y, \varepsilon)$. **Q.E.D.**

This proof in fact establishes a "strong" form of evolutionary instability of goal functions $f \in F^c$ for which $X(f) \cap X(f^*) = \varnothing$, in the sense that residents with such a goal function earn a strictly lower material payoff in *all* Nash equilibria for $\varepsilon > 0$ small. (We did not impose such a stringent condition in the definition of instability in AW; it only required that there exist at least one equilibrium for $\varepsilon > 0$ small enough in which residents earn a strictly lower material payoff than mutants.)

An interesting novelty compared to our previous analyses is that in the new proof the mutant is *Homo moralis*, and not a mutant always using the same strategy, that can invade a population where the resident type is some $f \in F^c$ for which $X(f) \cap X(f^*) = \varnothing$.

Remark 1. *In Alger and Weibull (2013) we required for a goal function to be unstable that residents with this goal function earn a lower material payoff against some mutant goal function in all Nash equilibria for $\varepsilon > 0$ small, without requiring existence of such Nash equilibria. Proposition 1 also establishes an extension of the second claim in Theorem 1 in that paper, by way of (a) dispensing with the hypothesis that the set $X(f)$ (there denoted X_θ) is a singleton, (b) replacing the hypothesis that the type set (there denoted Θ) is "rich" by the hypothesis that this set contains Homo moralis with degree of morality equal to the index of assortativity (these are defined for two-player games), (c) requiring a concavity property of the material payoff function and the goal function under examination, and (d) establishing existence of Nash equilibria between residents and the mutant.*

The proof of Proposition 1 can be adapted to obtain a result that does not require concavity of the resident type. For this result, recall the definition in AW of a behavioral alike to *Homo moralis*. This is a preference type which for at least one strategy $\hat{x}$ belonging to the set $X(f^*)$ of symmetric equilibrium strategies for the game between *Homo moralis*, has a best response $\hat{y}$ to $\hat{\boldsymbol{x}} = (\hat{x}, \hat{x}, ..., \hat{x}) \in X^{n-1}$ that is also a best response for *Homo moralis*.

Proposition 2. *Consider a goal function $f \in F$ that is not a behavioral alike to Homo moralis, for which $X(f) \neq \varnothing$ and for which there exists some $\bar{\varepsilon} > 0$ such that $B^{NE}(f, f^*, \varepsilon) \neq \varnothing$ for all $\varepsilon \in (0, \bar{\varepsilon})$. Then f is evolutionarily unstable.*

Proof: Consider a goal function f with the assumed properties. The non-emptiness of $X(f)$ implies that $B^{NE}(f, f^*, 0)$ is non-empty too, since $(x^*, y^*) \in B^{NE}(f, f^*, 0)$ if and only if $x^* \in X(f)$ and

$$y^* \in \arg\max_{y \in X} f^*(y, (x^*, ..., x^*)),$$

where the latter set is non-empty by Weierstrass' maximum theorem (f^* is continuous and X is non-empty and compact). Hence, the domain of the u.h.c. correspondence $B^{NE}(f, f^*, \cdot)$ can be extended to include $\varepsilon = 0$.

Consider any $(x^*, y^*) \in B^{NE}(f, f^*, 0)$. Then $x^* \notin \arg\max_{y \in X} f^*(y, (x^*, ..., x^*))$, since otherwise x^* would also belong to $X(f^*)$, and f would then be a behavioral alike to f^*. Thus, for all $(x^*, y^*) \in B^{NE}(f, f^*, 0)$,

$$\Pi_R(x^*, y^*, 0) = f^*(x^*, (x^*, ..., x^*)) < f^*(y^*, (x^*, ..., x^*)) = \Pi_M(x^*, y^*, 0).$$

Since there exists some $\bar{\varepsilon}$ such that $B^{NE}(f, f^*, \varepsilon) \neq \varnothing$ for all $\varepsilon \in (0, \bar{\varepsilon})$ (by assumption), and noting that the correspondence $B^{NE}(f, f^*, \varepsilon) : (0, 1) \rightrightarrows X^n$ is u.h.c. and compact-valued (by Berge's maximum theorem), the arguments given in the proof of Proposition 1 apply here as well. **Q.E.D.**

We end by briefly considering a counter-example to the instability claim in Theorem 1 of AW. This example builds upon Example 3 in Bomze et al. (2020).

Example 1. *Let π be the mixed-strategy payoff function for the generalized Rock-Paper-Scissors game with material-payoff matrix (for the row player)*

$$P(a) = \begin{pmatrix} 1 & 2-a & 0 \\ 0 & 1 & 2-a \\ 2-a & 0 & 1 \end{pmatrix}$$

for some $a < 1$. With mixed strategies represented as column vectors, the goal function f^*_κ for Homo moralis with degree of morality $\kappa \in [0, 1]$ is defined by

$$f^*_\kappa(x, y) = (1 - \kappa) x^T P(a) y + \kappa x^T P(a) x \quad \forall x, y \in \Delta$$

where Δ is the unit simplex in $\mathbb{R}^3$. As shown in Bomze et al. (2020), f^*_κ is strictly convex in x for all $a \in (0,1)$ and $\kappa \in (0,1)$, and then $X(f^*_\kappa) = \varnothing$. Hence, if $\sigma \in (0,1)$ is the index of assortativity in the matching process, then f^*_κ, for $\kappa = \sigma$, is evolutionarily stable according to the first claim in Theorem 1 in AW, and yet $f = f^*_\sigma$ meets the hypothesis for instability in the second claim in the same theorem, "$f \in F$ with $X(f) \cap X(f^*) = \varnothing$". By definition, an evolutionarily stable goal function cannot be evolutionarily unstable.

REFERENCES

[1] Alger, I., and J. Weibull (2013): "Homo moralis—preference evolution under incomplete information and assortative matching", *Econometrica 81*, 2269-2302.

[2] Alger, I., and J. Weibull (2016): "Evolution and Kantian morality", *Games and Economic Behavior* 98, 56-67.

[3] Aliprantis C., and K. Border (2006): *Infinite-Dimensional Analysis: a Hitchhiker's Guide*. Third edition. Berlin: Springer Verlag.

[4] Bomze, I., W. Schachinger, and J. Weibull (2021): "Does moral play equilibrate?", *Economic Theory* 71, 305-315.

Declarations of interest: none.