RECent
Center for Economic Research

WORKING PAPER SERIES

Andersen's "Emperor's New Clothes" as an Evolutionary Coordination Game with Stochastic Stability

Tampieri Alessandro

Working Paper 164

May 2026

www.recent.unimore.i

RECent: c/o Dipartimento di Economia Politica, Viale Berengario 51, I-41100 Modena, ITALY
Phone +39 059 2056856, Fax +39 059 2056947, E-mail: RECent@unimore.it

Andersen's "Emperor's New Clothes" as an Evolutionary Coordination Game with Stochastic Stability

Alessandro Tampieri*

May 25, 2026

Abstract

We model Andersen's "Emperor's New Clothes" as a coordination game under evolutionary dynamics. Citizens choose between praising non-existent clothes (C) and telling the truth (T), facing conformity gains and stigma for dissent. Payoff differences are monotone in the share of conformists, yielding a unique threshold x^*. Under replicator dynamics, all-C and all-T are asymptotically stable with an unstable threshold. With rare idiosyncratic plays (Young, 1993), we characterise stochastic stability via reduced resistances, and we outline the conditions under which a single truthful outburst causes the conformist convention to collapse, while restoring it requires many coordinated mistakes. The result is robust to payoff specifications that preserve monotone thresholds and clarifies when a small public deviation triggers a norm cascade.

JEL: C73. D83. D91

Keywords: coordination; evolutionary dynamics; stochastic stability.

*Corresponding author. Department of Economics, University of Modena and Reggio Emilia, and CREA, University of Luxembourg. Email: alessandro.tampieri@unimore.it

1 Introduction

Evolutionary game theory (EGT) emerged at the intersection of biology and economics in the 1970s. It has its roots in the work of John Maynard Smith and George Price, who applied game-theoretic reasoning to evolutionary biology and introduced the concept of an evolutionarily stable strategy (ESS). An ESS is a strategy that is resistant to invasion by mutants (Maynard Smith and Price, 1973). This biological foundation was soon complemented by contributions from economists such as Taylor and Jonker (1978), who formulated replicator dynamics by linking population strategy frequencies to differential payoffs.

This note reads Hans Christian Andersen's (1837) fable *The Emperor's New Clothes* (Andersen, 1837) as an evolutionary coordination problem: a fragile norm of public praise—sustained by fear of social stigma—coexists with a norm of truth-telling. We provide a compact selection result under stochastic perturbations. Considering a large population and letting x^* denote the critical share of conformists above which conformity remains self-sustaining, the truth-telling convention is stochastically stable precisely when the conformist threshold x^* exceeds one half. Moreover, when $x^* > 1 - \frac{1}{N}$, where N denotes the population size, a single idiosyncratic truth-telling act suffices to tip the population towards truth-telling, echoing Andersen's child. Interestingly, the tale was written more than twenty years before Darwin (1859)'s *On The Origin of Species*.

Andersen (1837)'s story has been considered relevant for understanding how societies work. For instance, the persistence of the Emperor's invisible clothes is a classic case study in "pluralistic ignorance", a phenomenon where a majority of group members privately reject a norm but assume, incorrectly, that most others accept it (Prentice and Miller, 1993 and Centola et al., 2005, among others). It is often described as the situation where "no one believes, but everyone believes that everyone else believes" (Krech and Crutchfield, 1948). This collective delusion is not a product of ignorance regarding the truth, but rather of ignorance regarding the private beliefs of others. The closest related contribution is Centola

et al. (2005), who provide an agent-based, computational-network account of how unpopular norms emerge and propagate. The paper contributes by translating the Emperor's dilemma from an agent-based account of norm cascades into a canonical evolutionary coordination game with stochastic stability: unlike Centola et al. (2005), this paper derives a closed-form selection criterion for the long-run survival of truth-telling versus conformity under rare perturbations. The result clarifies the exact sense in which Andersen's child is pivotal: a single truthful deviation triggers a cascade only when the conformist convention has a sufficiently small basin of attraction. Thus, the paper supplies an analytical benchmark for a literature dominated by computational and sociological models of pluralistic ignorance, public compliance and norm enforcement.

Consistent with Andersen (1837)'s tale, Noelle-Neumann (1974) and Noelle-Neumann (1984) study the spiral of silence: individuals monitor the perceived climate of opinion and tend to remain silent when they believe their view is unpopular, because they fear social isolation. This silence makes the apparently dominant opinion look even more dominant, which further discourages dissent. Noelle-Neumann's focus is less on strategic payoff modelling and more on public opinion, communication, social psychology, and fear of isolation.

Kuran (1995) studies preference falsification: as in *The Emperor's New Clothes*, people may publicly support a belief, norm, regime, or policy that they privately reject, because truthful expression is socially or politically costly. The key point is the gap between private beliefs/preferences and public declarations. This can generate false public consensus: everyone thinks everyone else supports the norm, while many privately oppose it. That is why social change can look sudden: once some people reveal dissent, others discover that opposition was more widespread than expected.

The present paper should not be read as a complete formalisation of Noelle-Neumann (1974)'s spiral of silence, nor of preference falsification in the sense of Kuran (1995). Those approaches address the broader process through which private beliefs, perceived public opinion,

fear of isolation, and public expression interact. Here, we take as a reduced-form premise the central asymmetry highlighted by that literature: private disbelief may be widespread but latent, while public behaviour is governed by reputational sanctions and conformity incentives. The contribution is instead to provide an evolutionary-selection benchmark for the public convention once this private-public gap is already present. In this benchmark, the question is not why citizens privately disbelieve yet publicly conform, but whether the convention of public praise is dynamically robust, how large its basin of attraction is, and when a visible truthful deviation can trigger a cascade. In this sense, the model complements the preference-falsification and spiral-of-silence approaches: they explain the emergence and persistence of concealed dissent, whereas the present stochastic-stability analysis characterises the evolutionary fragility of the resulting public convention.

The remainder is structured as follows. Section 2 summarises Andersen's tale, while Section 3 introduces the framework. Section 4 develops the equilibrium and stability analysis, as well as some comparative statics. Section 5 briefly concludes.

2 *The Emperor's New Clothes*

The story is about an Emperor who loves fine clothing. Two swindlers arrive in the capital and claim the clothing they will make is the most magnificent. They also say that this clothing is magical: anyone who is incompetent or stupid will not be able to see it.

The Emperor pays the men a huge sum to make these magnificent clothes. Then, the two swindlers pretend to sew and weave the clothing on empty looms. The Emperor sends the members of his court to check on their work. Each of them sees nothing, but all pretend they do, lest they be accused of being incompetent and stupid. Thus, each member of the court lies to the Emperor, saying that the clothing was splendid and of incomparable beauty.

On the day of a great procession, the clothing is brought to the Emperor, who also cannot see it, but he, too, does not want to admit to being stupid or incompetent. Hence he admits

that the clothing is superb. Dressed in the invisible garments, the Emperor marches in the procession in front of the crowd. Everyone sees the Emperor without clothes, but in order not to be accused of being stupid or incompetent, they all praise their Emperor's clothing.

Eventually, a child says, "But he is naked!" Everyone realises that if an innocent child is saying this, then it must be true. Everyone starts exclaiming, "The Emperor doesn't have anything on!" The Emperor then finishes the procession, knowing that the people are right.

3 The model

There is a population of N citizens who attend the Emperor's procession. A social norm emerges within the crowd: "Those unable to see the Emperor's clothes are stupid". In fact, the clothes do not exist, so nobody can see them. Time is continuous and, at each instant, each citizen makes a public statement, choosing either C for "Clothed" (lying by praising the beauty of the Emperor's clothes) or T for "Truth" (admitting that they cannot see any clothes).

In standard evolutionary models, payoffs often derive from random pairwise matching. However, many social phenomena, particularly those involving public norms, reputation, and collective signalling, are better described as population games or games with global externalities. In Hans Christian Andersen's (1837) *The Emperor's New Clothes*, the "stigma" of being perceived as stupid and the "conformity gain" of praising the invisible clothes are not products of isolated encounters between two citizens. Rather, they are functions of the *aggregate social atmosphere* (Horst, 2010).

Let $x \in [0, 1]$ denote the current share of conformists. For a representative citizen, let $u_C(x)$ and $u_T(x)$ be expected payoffs from choosing C or T. We impose a single-crossing *payoff-monotone* structure: the payoff difference

$$D(x) \equiv u_C(x) - u_T(x), \tag{1}$$

with $D'(x) > 0$, $D(0) < 0 < D(1)$, and C chosen if and only if $D(x) \geq 0$. Hence there exists a unique threshold $x^* \in (0,1)$ such that $D(x^*) = 0$. For $x > x^*$ conformity strictly dominates truth-telling; for $x < x^*$ the converse holds. We assume ties are broken in favour of C (a conservative convention).

Assume *threshold best replies*:

$$\text{play } C \iff x \geq x^*, \qquad \text{play } T \iff x < x^*, \tag{2}$$

for a given critical mass $x^* \in (0,1)$. For large x, reputational and social–sanction payoffs make C myopically optimal; once x falls below x^*, T strictly dominates and best replies snowball to T.

3.1 An example

Here we provide an example of the interaction. At each point in time, stage game payoffs reflect four forces:

- *Conformity gain* from matching the crowd, $c > 0$: the psychological or social utility of belonging to the majority.
- *Stigma* from dissenting when most people praise, $s > 0$: the social cost of being a truth-teller, which intensifies as the conformist share x grows.
- *Embarrassment/backfire* if one praises while many call it out, $e > 0$: the embarrassment of praising the clothes when a fraction $(1 - x)$ of the population is calling out the truth.
- *Intrinsic honesty/accuracy* from saying T: $h \geq 0$, plus a small *norm reinforcement* $b \geq 0$ when others also say T.

A parsimonious linear specification is:

$$u_C(x) = c\,x \;-\; e(1-x) \;-\; d, \tag{3}$$

$$u_T(x) = h \;+\; b(1-x) \;-\; s\,x, \tag{4}$$

where $d \geq 0$ is the (possibly tiny) disutility of knowingly lying.

Hence, the per-period payoff difference can be represented as

$$D(x) \equiv u_C(x) - u_T(x) = Bx - A, \tag{5}$$

where $A \equiv h + b + e + d$ and $B \equiv b + s + c + e > 0$. Best responses are threshold-type: in each time period, an agent plays T if and only if $D(x) < 0$, i.e., if and only if $x < x^*$ with

$$x^* = \frac{A}{B}. \tag{6}$$

An agent revises her strategy towards T if $x < x^*$ and towards C if $x \geq x^*$. Notice that the interior condition $x^* \in (0, 1)$ requires $A > 0$ and $A < B$, i.e.

$$h + d < s + c.$$

3.2 Deterministic adjustment

Since payoffs depend on x, the replicator dynamic is:

$$\dot{x} = x(1 - x)[u_C(x) - u_T(x)] = x(1 - x)[Bx - A]. \tag{7}$$

The fixed points are $x = 0$, $x = 1$, and $x = x^*$. At $x = 1$ (all-C), $\frac{d\dot{x}}{dx} = A - B$. Since $x^* < 1 \implies A < B$, the derivative is negative. However, for the share of C, the stability depends on the sign of $u_C - u_T$. If $x > x^*$, $u_C > u_T$, and $\dot{x} > 0$, making $x = 1$ locally stable. By contrast, at $x = 0$ (all-T), $u_T > u_C$, so $\dot{x} < 0$, making $x = 0$ locally stable. Finally, x^* is the unstable threshold partitioning the basins of attraction.

Hence, the absorbing states of the unperturbed process are the two homogeneous conventions, $x = 1$ and $x = 0$. With rare mutations, the resistance from one convention to the basin of attraction of the other is the minimum number of mistaken revisions required along a least-resistance path to cross the threshold x^*.

4 Evolutionary dynamics and stability

Consider a large finite population of size N with best-response revision and rare mutations. The absorbing states of the unperturbed process are the homogeneous conventions: all-C and all-T. With a small mutation rate, the long-run distribution concentrates on states with minimal *stochastic potential*, which can be computed from *reduced resistances* (Kandori et al., 1993, Young, 1993). Intuitively, the resistance to moving from one convention to the basin of the other equals the minimal number of (simultaneous) mistaken moves needed to cross the threshold that flips best replies.[1]

The resistance from the all-C convention to the basin of attraction of T is

$$r_{C\to T} = \min\left\{k \in \{0,\dots,N\} : 1 - \frac{k}{N} < x^*\right\} = \left\lfloor N(1-x^*)\right\rfloor + 1.$$

The resistance from the all-T convention to the basin of attraction of C is

$$r_{T\to C} = \min\left\{k \in \{0,\dots,N\} : \frac{k}{N} \geq x^*\right\} = \left\lceil Nx^*\right\rceil.$$

The next result follows.

THEOREM **1** (Selection under monotone thresholds: finite population)**.** *The truth-telling convention T is stochastically stable if and only if*

$$r_{C\to T} \leq r_{T\to C},$$

that is,

$$\left\lfloor N(1-x^*)\right\rfloor + 1 \leq \left\lceil Nx^*\right\rceil.$$

It is uniquely stochastically stable if and only if the inequality is strict. For large N, this condition converges to the simple threshold condition

$$x^* > \frac{1}{2}.$$

[1] Formal definitions use resistance trees; here we exploit the payoff-monotone structure, which yields closed-form counts.

Proof. Starting from the all-C convention, the share of conformists is $x = 1$. Each mistaken move from C to T reduces the share of conformists by $1/N$. After k such mistakes, the share of conformists is

$$x = 1 - \frac{k}{N}.$$

The process enters the basin of attraction of T when

$$1 - \frac{k}{N} < x^*.$$

Thus the minimum number of mistakes required is

$$r_{C\to T} = \min\left\{k : 1 - \frac{k}{N} < x^*\right\} = \left\lfloor N(1 - x^*)\right\rfloor + 1.$$

Conversely, starting from the all-T convention, the share of conformists is $x = 0$. Each mistaken move from T to C increases the share of conformists by $1/N$. After k such mistakes, the share of conformists is

$$x = \frac{k}{N}.$$

Because ties are broken in favour of C, the process enters the basin of attraction of C when

$$\frac{k}{N} \geq x^*.$$

Hence the minimum number of mistakes required is

$$r_{T\to C} = \min\left\{k : \frac{k}{N} \geq x^*\right\} = \left\lceil Nx^*\right\rceil.$$

The convention with the lower stochastic potential is uniquely stochastically stable; if the two stochastic potentials coincide, both conventions are stochastically stable. Thus T is stochastically stable if and only if

$$r_{C\to T} \leq r_{T\to C},$$

with uniqueness if and only if

$$r_{C\to T} < r_{T\to C}.$$

Substituting the two resistance expressions gives

$$\lfloor N(1-x^*)\rfloor + 1 \leq \lceil Nx^*\rceil.$$

As N becomes large, the integer terms are asymptotically equivalent to

$$N(1-x^*) \quad \text{and} \quad Nx^*,$$

so the condition becomes

$$N(1-x^*) < Nx^*,$$

or equivalently,

$$x^* > \frac{1}{2}.$$

□

The outcome of the tale is summarised in the following corollary. A single truthful deviation is sufficient to move the population from the all-C convention into the basin of attraction of T if and only if

$$r_{C\to T} = 1.$$

Equivalently,

$$\lfloor N(1-x^*)\rfloor + 1 = 1,$$

which holds if and only if

$$N(1-x^*) < 1,$$

or

$$x^* > 1 - \frac{1}{N}.$$

Thus,

COROLLARY **1** ("The Emperor is naked!" result)**.** *If*

$$x^* > 1 - \frac{1}{N},$$

one idiosyncratic truth-telling act is enough to push the population below the conformist threshold and trigger convergence to the truth-telling convention.

It is important to emphasise that the "child" result is not the generic prediction of the model. In general, a single truthful deviation need not overturn a conformist convention: the number of deviations required to trigger a cascade depends on the size of the basin of attraction of the all-C equilibrium. The Andersen case corresponds to the limiting situation in which this basin is very small. Formally, when $x^* > 1 - \frac{1}{N}$, one truthful statement is sufficient to move the population below the conformist threshold; substantively, this means that almost all citizens were already close to abandoning the public lie. The child does not create disbelief from nothing. Rather, the child's statement reveals and coordinates a latent willingness to switch, making public what was already privately shared. Thus, the model should not be read as claiming that any public falsehood collapses after one dissenting voice, but as identifying the precise condition under which Andersen's dramatic ending can occur.

We conclude the section by showing some intuitive comparative statics using the linear specification.

COROLLARY **2.** *Greater stigma s and stronger conformity pull c* lower *the tipping point, strengthening the all-C norm; greater honesty h and lying disutility d* raise *it, weakening the all-C norm. Stronger truth-side reinforcement b and stronger embarrassment from praising when others tell the truth e both raise the tipping point.*

Proof. From Equation (6), with $A = h + b + e + d$ and $B = b + s + c + e$, differentiation of x^* with respect to s, c, h, d, b and e, respectively, yields:

$$\frac{\partial x^*}{\partial s} = -\frac{A}{B^2} < 0, \quad \frac{\partial x^*}{\partial c} = -\frac{A}{B^2} < 0,$$
$$\frac{\partial x^*}{\partial h} = \frac{1}{B} > 0, \frac{\partial x^*}{\partial d} = \frac{1}{B} > 0,$$

while

$$\frac{\partial x^*}{\partial b} = \frac{B - A}{B^2} > 0, \quad \frac{\partial x^*}{\partial e} = \frac{B - A}{B^2} > 0,$$

given that the maintained interior-threshold condition requires $B > A$. □

By Corollary 2, parameters that increase A (e.g. intrinsic honesty, benefits from shared truth, reinforcement of candid norms) raise x^* and thereby favour the selection of T; parameters that increase B (e.g. stigma for dissent, costs of appearing foolish when truthful) reduce x^* and strengthen C. Larger x^* shrinks $r_{C\to T}$ and expands $r_{T\to C}$, making truth-telling both easier to reach and harder to overturn. This formalises the intuition that small public deviations can trigger norm cascades when a latent majority already favours truth (e.g. Centola et al., 2018).

5 Conclusion

The tale captures a coordination problem with two conventions: public deception sustained by stigma versus candid truth. A simple monotone-threshold specification yields clean evolutionary and stochastic selection results: with a large population, truth is selected whenever the conformist threshold exceeds one half, and—in finite populations—one outspoken deviation may suffice to tip society. The analysis connects literary intuition to modern equilibrium-selection theory.

References

Andersen, H., 1837. Fairy Tales Told for Children. First Collection. C. A. Reitzel.

Centola, D., Becker, J., Brackbill, D., Baronchelli, A., 2018. Experimental evidence for tipping points in social convention. Science 360 (6393), 1116–1119.

Centola, D., Willer, R., Macy, M., 2005. The emperor's dilemma: A computational model of self-enforcing norms. American Journal of Sociology 110 (4), 1009–1040.

Darwin, C., 1859. On the origin of species by means of natural selection. London: J. Murray.

Horst, U., 2010. Dynamic systems of social interactions. Journal of Economic Behavior & Organization 73 (2), 158–170.

Kandori, M., Mailath, G. J., Rob, R., 1993. Learning, mutation, and long run equilibria in games. Econometrica: Journal of the Econometric Society , 29–56.

Krech, D., Crutchfield, R. S., 1948. Theory and problems of social psychology. Vol. 36. McGraw-Hill New York.

Kuran, T., 1995. Private Truths, Public Lies: The Social Consequences of Preference Falsification. Harvard University Press, Cambridge, MA.

Maynard Smith, J., Price, G. R., 1973. The logic of animal conflict. Nature 246 (5427), 15–18.

Noelle-Neumann, E., 1974. The spiral of silence: A theory of public opinion. Journal of Communication 24 (2), 43–51.

Noelle-Neumann, E., 1984. The Spiral of Silence: Public Opinion–Our Social Skin. University of Chicago Press, Chicago.

Prentice, D. A., Miller, D. T., 1993. Pluralistic ignorance and alcohol use on campus: some consequences of misperceiving the social norm. Journal of personality and social psychology 64 (2), 243.

Taylor, P. D., Jonker, L. B., 1978. Evolutionarily stable strategies and game dynamics. Mathematical Biosciences 40 (1-2), 145–156.

Young, H. P., 1993. The evolution of conventions. Econometrica: Journal of the Econometric Society 61, 57–84.

RECent Working Papers Series

The 10 most RECent releases are:

No. 163 ANCESTRAL INEQUALITY AND PREFERENCES FOR REDISTRIBUTION (2026)
G. Bertocchi, A. Dimico, Tedeschi G.

No. 162 NONLINEAR BUSINESS-CYCLE ANATOMY
M. Brianti, M. Forni, L. Gambetti, A. Granese

No. 161 FREQUENCY-BAND ESTIMATION OF THE NUMBER OF FACTORS (2024)
M. Avaruccia, M. Cavicchioli M. Forni, P. Zaffaroni

No. 160 INFORMING DSGE MODELS THROUGH DYNAMIC FACTOR MODELS (2024)
M. Forni, L. Gambetti, M. Lippi, L. Sala

No. 159 COMMON COMPONENTS STRUCTURAL VARS (2024)
M. Forni, L. Gambetti, M. Lippi, L. Sala

No. 158 MATH EXPOSURE AND UNIVERSITY PERFORMANCE: CAUSAL EVIDENCE FROM TWINS (2024)
G. Bertocchi, L. Bonacini, M. Joxhe, G. Pignataro

No. 157 FAMILY PLANNING AND ETHNIC HERITAGE: EVIDENCE FROM SUB-SAHARAN AFRICA (2024)
G. Bertocchi, A. Dimico, C. Falco

No. 156 TWO MAIN BUSINESS CYCLE SHOCKS ARE BETTER THAN ONE (2024)
A. Granese

No. 155 TEMPERATURE AND GROWTH: A PANEL MIXED FREQUENCY VAR ANALYSIS USING NUTS2 DATA (2023)
A. Cipollini, F. Parla

No. 154 NATURAL DISASTERS AND PREFERENCES FOR THE ENVIRONMENT: EVIDENCE FROM THE IMPRESSIONABLE YEARS (2022)
C. Falco, R. Corbi

The full list of available working papers, together with their electronic versions, can be found on the RECent website: https://www.recent.unimore.it/recent-working-papers/

RECent: c/o Dipartimento di Economia Politica, Viale Berengario 51, I-41100 Modena, ITALY
Phone +39 059 2056856, Fax +39 059 2056947, E-mail: RECent@unimore.it