Journal of Evolutionary Economics
https://doi.org/10.1007/s00191-023-00826-6

REGULAR ARTICLE

Five shades of green: Heterogeneous environmental attitudes in an evolutionary game model

Angelo Antoci[1] · **Simone Borghesi**[2,3] · **Giulio Galdi**[4]

Accepted: 24 May 2023
© The Author(s) 2023

Abstract
An environmental policy to foster virtuous behaviour does not automatically establish a social norm in a population; that is, the policy might not be socially acceptable or enforceable. Some agents feel compelled to abide by environmental social norms and embrace them, but others do not. Some might want to imitate their peers, while others might prefer not to conform and play the role of a maverick. In this model, we describe the heterogeneity of preferences by proposing a taxonomy of five possible agent types that enrich the traditional triplet presented in the literature. We then employ a random matching model to study how a social norm spreads within a population when its composition changes. Considering three relevant population compositions (scenarios), we show that what is most important for the successful diffusion of social norms is not whether, but why agents abide by it.

Keywords Environmental externalities · Evolutionary games · Replicator dynamics · Social norms · Green behaviour

JEL Classification C73 · D62 · D91 · Q5

1 Introduction

The existence of different preferences across agents is one major challenge in the preservation of environmental goods, as it makes it harder to broker an agreement (de Zeeuw 2015) or to uphold it afterwards (Klaser et al. 2021). Indeed, even when

✉ Simone Borghesi
simone.borghesi@eui.eu

1 Department of Economic and Business Sciences, University of Sassari, Sassari, Italy

2 Department of Political and International Sciences, University of Siena, Siena, Italy

3 Florence School of Regulation, European University Institute, Florence, Italy

4 Department of Economics and Management, University of Trento, Trento, Italy

Published online: 07 August 2023

environmental norms and agreements are put in place, they can be disregarded by agents if they find them too hard to comply with or not in line with their preferences. In this paper, we propose a model showing how the uptake and diffusion of a pro-environmental behaviour can indeed be influenced by the evolution of agents' preferences.

We study a population of agents that interact through random matching and, at each encounter, must choose between two strategies: Pollute (P) and Non-pollute (NP). Agents are heterogeneous with respect to their preferences, i.e. the ordering of payoffs attributed to the four possible outcomes of the game: (P, P), (NP, NP), (NP, P), and (P, NP). We consider five different types of agents that are characterised by their preferences over the possible outcomes of the game. Following replicator dynamics, the shares of agents in the population change according to the payoff they obtain. Agents can recognise the type of other agents as in Ferguson and Flynn (2016), react to the expected behaviour of those agents, and then possibly adjust their preferences based on their interaction through a cultural transmission mechanism like that studied by Bisin and Verdier (2005).

We contribute to the extant literature by shedding light on an aspect that has been mainly ignored so far, namely, that agents' preferences and their underlying motives are more important than their behaviour to assess whether a pro-environmental behaviour will diffuse in a population of agents. Moreover, we also show that social dynamics are often path-dependent; that is, history, as expressed by the initial distribution of agent types in the population, determines the final outcome (e.g. environmental quality) the society tends to achieve.

Our work aligns with studies stressing the importance of endogenous dynamics in the evolution of preferences in a population (Antoci et al. 2000; Akçay and Cleve 2012; Lehmann et al. 2015). Indeed, agents' preferences can be influenced by others through several channels, e.g. expected cooperation (Bruhin et al. 2016; Drouvelis and Georgantzis 2019), retaliation due to failed cooperation attempts (Andreoni 1995; Richter and Grasman 2013), imitation of the most successful agents (Bar-Gill and Fershtman 2005), or preference for non-conformity (Antoci et al. 2018).

The remainder of this paper is structured as follows. Section 2 characterises the five types of agents considered in the study. Section 3 describes the evolutionary dynamics of the population. Section 4, 5, 6 consider three distinct types of triads and present the possible scenarios emerging in those cases. Section 7 presents our final considerations.

2 Player characterisation

2.1 Five types of agents

A very common conceptual framework in the literature on heterogeneity of agents distinguishes them into two types, cooperators and defectors. The actual distinction between these types can be described in a variety of ways. In experimental settings, the two types can be directly elicited from their decisions: defectors contribute either less than the players they are matched with or nothing at all, whereas cooperators contribute more than what their match contributes or what they should rationally

contribute in general (Fischbacher et al. 2001). In evolutionary approaches similar to ours, cooperators and defectors are typically identified by their contribution to a common pool or adoption of altruistic behaviour (Doebeli et al. 2004). However, we introduce an additional layer to agents. We sort them into one of five different agent types according to their behaviour and their preferences. As illustrated in more detail later, classic (unconditional) cooperators are represented by our environmentalist (E) agents and defectors are represented by our non-environmentalist (NE) agents.

Besides cooperators and defectors, the literature highlights the strong presence of the so-called conditional cooperators in the population, that is the agents who reciprocate cooperation and defect otherwise (Fischbacher et al. 2001; Drouvelis and Georgantzis 2019). These agents use the tit-for-tat strategy, initially choosing cooperation and then following the strategy of the agent with whom they are matched. This strategy has been found to yield higher long-term payoffs than other strategies (Axelrod and Hamilton 1981). In our study, conditional cooperators always imitate the action of the other agent, but would still prefer cooperation. Analogously to the environmentalist and non-environmentalist agents, we call this agent a conditional environmentalist (CE). The three agent types presented so far (cooperators, defectors, and conditional cooperators) represent the main pillars of the agent taxonomy commonly used in the scientific literature.[1] However, some phenomena have compelled scholars to expand this basic view by considering additional types of agents.

For instance, a well-known finding in public good game experiments is that the average contributions decrease in successive rounds. A common explanation in the literature is the presence of conditional cooperators. When these agents choose to contribute slightly less than the others, it ignites a vicious cycle resulting in contributions falling dramatically by the end of the experiment (Fischbacher and Gächter 2010; Chaudhuri 2011; Richter and Grasman 2013). Another reason is the presence of agents who would rather not contribute, but are required to abide by what is perceived as a desirable contribution level. When this social norm weakens, that is the average contribution level decreases, they feel less peer pressure and revert to no contribution. This line of behaviour is also in line with the Goal-Framing Theory (GFT), which has been proposed and used by the environmental psychology literature to explain pro-environmental behaviour (Lindenberg and Steg 2007; Gkargkavouzi et al. 2019a). Under the GFT framework, people act with the following three main motives: gain (selfish), hedonic (experiential), and normative (moral). If an individual considers an environmental action morally right but contrasts it with the gain motive (e.g. by considering it burdensome), it could induce the individual to defect/not cooperate as soon as the moral norm is perceived to be weak or weakening (Steg et al. 2014). In other words, the fear of social stigma might push some individuals into cooperative behaviours (Tavoni et al. 2012). We integrate this phenomenon by introducing the ashamed non-environmentalist (AnE) as an agent type behaving similarly to the conditional environmentalist (who reciprocates both cooperation and defection) but reversing preferences over payoffs. In terms of outcome, the AnE type prefers to be matched with a defector and defect, rather than be matched with a cooperator and

[1] They are sometimes considered to be generous, spiteful, and conditionally generous agents (Gul and Pesendorfer 2016).

thus be compelled to cooperate. In our interpretation, this agent type is particularly susceptible to social influences and thus decides to cooperate due to a sense of shame when matched with cooperators.

Another component of pro-environmental behaviour typically analysed is identity, which can, for instance, influence the propensity to buy electric cars (Barbarossa et al. 2017) and bio-plastic products (Confente et al. 2020), offset their own carbon emissions (Whitmarsh and O'Neill 2010), or support renewable energy sources (Bauwens and Devine-Wright 2018). In addition, while most people respond negatively to the anti-social behaviour of other agents (Drouvelis and Georgantzis 2019), some agents might obtain some gratification by belonging to a virtuous minority (Antoci et al. 2018). The reason could be that belonging to a minority group reinforces their identity (Akerlof and Kranton 2010) or induces a relatively warm glow effect (Ferguson and Flynn 2016; Andreoni 1990). We capture this effect by introducing the snob environmentalist (SnE), who, like environmentalists, always cooperates. However, the SnE actually prefers to be the only one adopting pro-environmental behaviours. In line with previous studies (see, e.g. Mancha and Yoder 2015; Gkargkavouzi et al. 2019b), we find that identity plays a crucial role in upholding pro-environmental behaviours in a population where it is slightly diffused.

2.2 The general context

The game we analyse has a random matching structure; two agents are chosen at random from an infinite population to share the enjoyment of a given environmental good. In general, agents interact in such a manner that their type or attitude towards pro-environmental behaviours becomes visible. That is, they can choose their best response to the action of the other agent depending on their own preferences, i.e. their payoff structure. Agents may either adhere to the environmental social norm or not; that is, they have two choices:

(a) Agents may $pollute$ (P) (broadly speaking) an environmental good.
(b) Agents may $not\ pollute$ (NP) an environmental good.

Our agents are matched in pairs; thus, we have four possible outcomes from the interaction between agents:

$$(P, P),\ (NP, NP),\ (P, NP),\ (NP, P),$$

where the entries of each pair represent the choices of any two matched agents. We construct our taxonomy on the preference ordering of these outcomes and describe the following agent types mentioned in the previous subsection: *environmentalist*, *non-environmentalist*, *conditional environmentalist*, *ashamed non-environmentalist*, and *snob environmentalist*. We do not claim this to be an exhaustive taxonomy of all possible attitudes of people towards the environment. We assume that adhering to the environmental social norm requires the agent to play NP. We also assume that agents observe the type of the other agents with whom they are matched. We now characterise the five agent types by their preferences based on the above outcomes and summarise the resulting taxonomy in Table 1.

2.2.1 The environmentalist

Environmentalists (E) always play strategy NP irrespective of the strategy of the agent with whom they are matched. Indeed, the value they attribute to the environment is greater than their effort related to NP:

$$(NP, NP) \succ (NP, P) \succ (P, NP) \succ (P, P),$$

where $A \succ B$ must be read as A is preferred to B. The ordering of the payoffs reflects the preferences of the first agent in the couplet. Given this preference ordering, the favourite outcome is the one where both agents preserve the environment.

2.2.2 The non-environmentalist

Non-environmentalists (NE) are almost at the other end of the behavioural spectrum; that is, they always choose not to adhere to the environmental social norm, considering it too burdensome:

$$(P, NP) \succ (P, P) \succ (NP, NP) \succ (NP, P)$$

Although non-environmentalists NE are not willing to strive to preserve the environment, they would still desire that the other agents do so and play NP. Therefore, NE achieve their highest payoff when matched with someone playing NP.

2.2.3 The conditional environmentalist

Conditional environmentalists (CE) would desire that both agents adhere to the environmental social norms. However, contrary to the environmentalists, they hate being the only ones doing so and would choose to punish the non-abiding agents by increasing pollution, lowering the payoff of both agents. Therefore, the ranking of preferences will be as follows:

$$(NP, NP) \succ (P, NP) \succ (P, P) \succ (NP, P)$$

Table 1 Preference orderings for all agent types. Outcomes that can be actually realised are highlighted in bold

	Best outcome → Worst outcome						
Environmentalist	$\boldsymbol{(NP, NP)}$	$\succ$	$\boldsymbol{(NP, P)}$	$\succ$	(P, NP)	$\succ$	(P, P)
Non-environmentalist	$\boldsymbol{(P, NP)}$	$\succ$	$\boldsymbol{(P, P)}$	$\succ$	(NP, NP)	$\succ$	(NP, P)
Conditional Environmentalist	$\boldsymbol{(NP, NP)}$	$\succ$	(P, NP)	$\succ$	$\boldsymbol{(P, P)}$	$\succ$	(NP, P)
Ashamed Non-environmentalist	$\boldsymbol{(P, P)}$	$\succ$	$\boldsymbol{(NP, NP)}$	$\succ$	(NP, P)	$\succ$	(P, NP)
Snob Environmentalist	$\boldsymbol{(NP, P)}$	$\succ$	$\boldsymbol{(NP, NP)}$	$\succ$	(P, P)	$\succ$	(P, NP)

However, their attitude leads CE to always do what the other agent does (tit-for-tat strategy), such that only the first and third outcomes would ever occur. Note that this preference ordering is similar to that of the environmentalist, the only difference being that (NP, P) is now the least-preferred outcome. We extend the variety of types by including the following two agents.

2.2.4 The ashamed non-environmentalist

Ashamed non-environmentalists (ANE) represent an NE who feels peer pressure to adhere to the environmental social norm (or feels shame in not doing so) and thus prefers to play NP rather than being the only one to play P:

$$(P, P) \succ (NP, NP) \succ (NP, P) \succ (P, NP).$$

Like CE, this agent type always prefers to do what the other agent does, but for a different reason and with different ranking, where (P, P) is the favourable outcome. This agent type is compelled to cooperate on account of guilt, also referred to as social stigma (Tavoni et al. 2012) or cold shiver (Bruvoll and Nyborg 2004).

2.2.5 The snob environmentalist

Finally, snob environmentalists (SnE) identify great value in sustaining the environment, just as do E and CE agents. Following Akerlof and Kranton (2010), SnE agents derive an additional payoff by being the only agents to play NP, or a sort of identity bonus for playing an outsider role:

$$(NP, P) \succ (NP, NP) \succ (P, P) \succ (P, NP)$$

SnE always choose to play NP but would prefer that the other agent does not do so. Thus, their payoff is highest when matched with NE, but they are disappointed (get a lower payoff) when matched with an agent who also plays NP. These agents are the ones for whom identity is the strongest driver of pro-environmental behaviour, as in Whitmarsh and O'Neill (2010) and Ferguson and Flynn (2016).

We have seen how ashamed non-environmentalists (ANE) and snob environmentalists (SnE) behave similarly to the conditional environmentalists (CE) and environmentalists (E), respectively. Indeed, both ANE and CE behave as reciprocators (or conditional cooperators), whereas SnE and E always choose to abide by the environmental social norm. The behavioural equivalence between CE and ANE and between E and SnE is the reason why we include ANE and SnE in our taxonomy. As shown in the following sections, we exploit this equivalence to show that preferences could be more relevant than actual behaviour to determine the diffusion of pro-environmental behaviours. We study the dynamics of three populations, each consisting of three agent types. As mapping all possible populations is beyond the scope of this work, we select the triplets with most relevant comparisons. In the first scenario (benchmark), we study the population composed of E, NE, and CE agents, to find that only undesirable steady states can be asymptotically stable. In the second

scenario (identity), we substitute E with SnE, to find the emergence of more desirable asymptotically stable steady states. Finally, in the third scenario (social norm), we replace CE with ANE in the benchmark, to find that social norms (shame or pressure) may be more effective than punishment in pushing the system towards a desirable final state. We summarise the interactions and payoffs obtained for each agent type at the beginning of Sections 4–6.

3 Evolutionary dynamics

This section analyses the social dynamics following an evolutionary game approach[2]. We consider the population composed of three agent types at a time, selecting the triplets we deem more relevant to the discussion. The state of a population is represented by vector $\mathbf{x}$, whose components x_i represent the shares of type i, where $i = E, NE, CE, SnE, ANE$. The shares need to be non-negative:

$$x_i \geq 0 \; \forall \; i \text{ and } \sum_i x_i = 1. \tag{1}$$

As in (1), the complementarity of shares allows us to describe the share x_i of agent type i with respect to the other types, so that $\mathbf{x}$ belongs to the two-dimensional simplex S. The latter can be represented as a triangle whose vertexes describe the population where $x_i = 1$ for type i, whereas $x_{-i} = 0$.

When two individuals are matched, they behave according to their preference orderings, as summarised in Table 1. We assign a payoff equal to 1 to the agents reaching their optimal outcome (e.g. (P, NP) for NE and (NP, NP) for E). For non-optimal outcomes, the payoff is given by parameters $\alpha, \beta, \gamma < 1$, as reported later in the payoff matrices (5)–(7). The ranking of these parameters depends on the relative disutility of the agents in sub-optimal outcomes. Let $\boldsymbol{A}$ be the payoffs matrix whose elements a_{ij} identify the payoff of agent i when matched with agent j.[3] Given the random matching structure of the game, the (expected) payoff Π_i for an i-individual is

$$\Pi_i = \sum_j a_{ij} x_j. \tag{2}$$

The average population payoff $\overline{\Pi}$ is given by

$$\overline{\Pi} = \sum_i \Pi_i x_i. \tag{3}$$

[2] Sacco and Zamagni (1994) and Doebeli et al. (2004) followed an analogous approach to analyse altruistic behaviour.

[3] An analogous symmetric matrix can be defined for the column agent.

We assume that the shares of the agent types are not fixed, but endogenously determined by their corresponding payoffs. As argued in the introduction to this work, agents may change their type, that is their preferences, out of frustration with the defection of other players (Andreoni 1995; Richter and Grasman 2013) or simply because they want to achieve a higher payoff through imitation (Bar-Gill and Fershtman 2005; Bisin and Verdier 2005). We provide additional insights in Sections 4–6 for the possible interpretations of the dynamics. Following Taylor and Jonker (1978), we assume that the variation in share of each agent type depends on its payoff relative to that of other types present in the population. Formally, the growth rate of the share of type i ($\dot{x}_i/x_i$) individuals is equal to the difference between the expected payoff Π_i and average payoff of the entire population $\overline{\Pi}$ (as defined in (2) and (3), respectively):

$$\dot{x}_i = x_i(\Pi_i - \overline{\Pi}). \tag{4}$$

Equation (4), defined on the invariant two-dimensional simplex S, is the so-called replicator equation describing the dynamics by which the best-performing strategies diffuse in the population (see Samuelson 1997; Weibull 1995). Schlag (1998) and Björnerstedt et al. (1996) offer micro-founded justification of replicator dynamics. Note that under replicator dynamics, the behaviour of individuals has a certain amount of inertia. Individuals do not revise their preferences simultaneously, thus preventing discontinuous jumps from one strategy to another.

4 Benchmark scenario

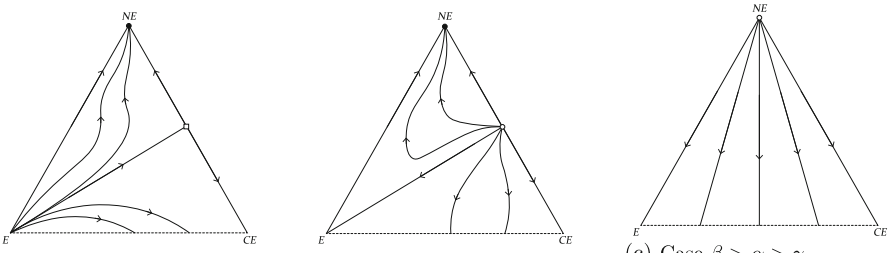
In this section, we analyse the social dynamics in a population composed of environmentalists (E), conditional environmentalists (CE), and non-environmentalists (NE). Figures 1 and 2 present only robust cases; that is, we exclude all cases characterised by equality conditions between parameters for the sake of brevity.[4] Note that in all figures in this paper, repulsors are represented by open dots ∘, attractors by full dots •, and saddle points by squares □. Finally, also note that in all scenarios, the pure population states, that is the states where $x_i = 1$ for type i and $x_{-i} = 0$, are always steady states under replicator dynamics.

In our benchmark population, composed of E, CE, and NE, the possible outcomes and related payoffs are summarised by the following matrices:

$$\begin{array}{c|ccc} & E & CE & NE \\ E & (NP,NP) & (NP,NP) & (NP,P) \\ CE & (NP,NP) & (NP,NP) & (P,P) \\ NE & (P,NP) & (P,P) & (P,P) \end{array} \qquad \begin{array}{c|ccc} & E & CE & NE \\ E & 1 & 1 & \alpha \\ CE & 1 & 1 & \beta \\ NE & 1 & \gamma & \gamma \end{array} \tag{5}$$

where the values denote the payoffs of the row individuals when they meet the column individuals, and $\alpha, \beta, \gamma < 1$. Note that each outcome represents a Nash Equilibrium as both agents adopt their best-reply strategy. In this scenario, E and CE

[4] Further information on non-robust cases can be obtained from the authors upon request.

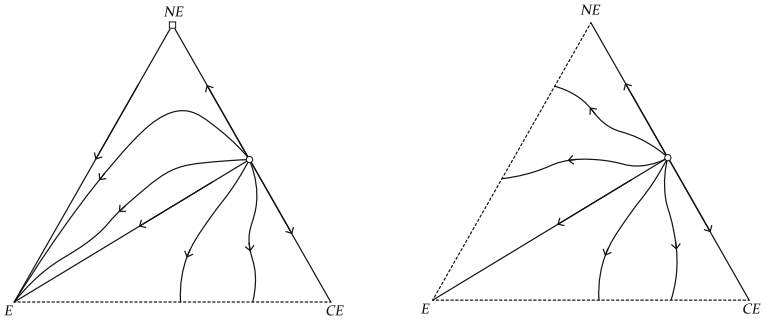


(a) Case: $\gamma > \beta > \alpha$. (b) Case: $\gamma > \alpha > \beta$ (c) Case $\beta > \alpha > \gamma$ or $\alpha > \beta > \gamma$.

Fig. 1 Dynamics in a (E, CE, NE)-population

agents obtain a sub-optimal payoff only when matched with an NE agent, whereas NE agents reach their optimal outcome only when they encounter an E agent. Indeed, when NE is matched with CE, the latter punishes the former by playing P as well, leading both to sub-optimal payoffs. As concerns the ranking of payoffs, note that matrix (5) is full row rank, although it is not full column rank. For instance, in the third column, we cannot assess whether E, CE, or NE perform better when matched with NE; that is, we do not assume the rankings of α, β, and γ. Such ordering depends on the relative disutility agents receive when they reach sub-optimal outcomes.

By applying replicator dynamics (4) to the payoffs of the population composed by the triad (E, CE, NE), we obtain three main results. First, if NE individuals incur only minor disutility when other agents play P, that is $\alpha, \beta < \gamma$, then the NE vertex of S, where all agents are of the NE type, is locally attractive (see Figs 1a-b). Intuitively, matching with NE leads to a sub-optimal outcome for both CE and E. Over time, they could be frustrated by the frequency of encounters with defectors, that is NE, and they decide to let go of their efforts to cooperate. Therefore, if the population already has a sufficiently high number of NE, then E and CE will disappear over time and all agents will be NE. Formally, the only asymptotically stable steady state is the vertex of S, where $x_{NE} = 1$ and $x_E = x_{CE} = 0$. Notably, we find that this is



(a) Case: $\alpha > \gamma > \beta$. (b) Case: $\beta > \gamma > \alpha$

Fig. 2 Dynamics in a (E, CE, NE)-population

the only scenario where the NE vertex, which has all agents choosing to play P, can be locally attractive.

Second, when NEs are relatively bothered by being matched with other agents playing P, that is $\gamma < \alpha$, β the share of NE goes to 0 (see Fig. 1a). This might indicate a situation particularly adverse to NE being matched with other agents playing P, that is CE and other NE. For instance, the environmental consequences of both agents playing P could be so impactful that it is too inconvenient to be an NE, and the E and CE types gain traction. In formal terms, we find the points on the bottom edge $E - CE$, where $x_{NE} = 0$, always stable, even if not asymptotically.

Finally, we find no stable steady state where all behaviour types can coexist under any parametric condition. Even in the cases in which γ takes an intermediate value between α and β, we observe that it is either NE or CE that becomes extinct (see Figs. 2a, b).

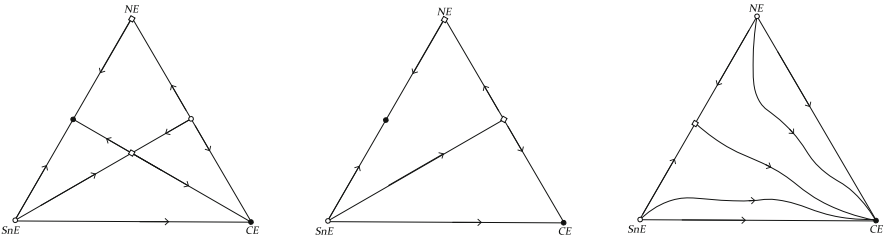
5 Identity scenario

In this section, we analyse the social dynamics in a population where the environmentalists in the benchmark scenario are substituted with the snob environmentalist such that the three agent types are NE, SnE, and CE. In contrast to the previous section, we classify all the cases here, including the non-robust ones, because of their limited number. We summarise the payoff and outcomes in the following matrices:

$$\begin{array}{c|ccc} & SnE & CE & NE \\ SnE & (NP,NP) & (NP,NP) & (NP,P) \\ CE & (NP,NP) & (NP,NP) & (NP,P) \\ NE & (P,NP) & (P,P) & (P,P) \end{array} \qquad \begin{array}{c|ccc} & SNE & CE & NE \\ SnE & \beta & \beta & 1 \\ CE & 1 & 1 & \gamma \\ NE & 1 & \alpha & \alpha \end{array} \tag{6}$$

where the usual interpretation of payoffs applies and $\alpha, \beta, \gamma < 1$. Although SnE always abides by the environmental social norm, like E, they obtain their maximum payoff only when matched with someone who does not abide. Thus, SnE can add an identity bonus to their utility function (Akerlof and Kranton 2010), which is why we denote this scenario as identity scenario. For this reason, SnE receive a higher payoff when they meet an NE and obtain suboptimal payoffs otherwise. Similarly, NE receive the maximum payoff only when matched with SnE. This symmetrical (almost symbiotic) relationship is found to be very relevant to our results.

Three main results emerge when we study the (NE, SnE, CE)-population dynamics (see Fig. 3a-c). First, we find that substituting E in the benchmark scenario with SnE makes the population of NE agents no longer attractive, but makes the population of CE agents attractive. When the share of CE is sufficiently high, it eventually approaches one. As in the previous scenario, consider a case with a large majority of NE agents in the population. SnE agents will achieve their highest payoff when matched with NE agents. Adopting the pro-environmental behaviour under study will become useful for SnE agents to express their identity, thus increasing their performance and attracting imitators in the population. However, as with any fashion



(a) Case: $\alpha > \gamma$ and $(1-\alpha)^2 > (\beta-\alpha)(\gamma-\alpha)$.

(b) Case: $\alpha > \gamma$ and $(1-\alpha)^2 \leq (\beta-\alpha)(\gamma-\alpha)$.

(c) Case: $\alpha \leq \gamma$.

Fig. 3 Dynamics in an (SnE, CE, NE)-population

or fad, when there are too many SnE, they tend to be paired with one another, to their detriment. Here, CE benefit because they would now have many agents in the population adopting a pro-environmental behaviour, increasing the likelihood of their matching leading to their preferred outcome, that is (NP, NP). In short, the presence of SnE can help CE become more prevalent in the population. Formally, vertex CE, where $x_{CE} = 1$ and $x_{NE} = x_{SnE} = 0$, is always an attractor. Notably, everyone abides by the environmental social norm in this steady state, that is all agents decide to play NP. Thus, this population scenario leads to a more desirable outcome for the environment with respect to the benchmark scenario. Indeed, in this scenario the state of the population comprising only NE, all of whom would play P, is not attractive. Snob environmentalists leverage the benefit of belonging to a virtuous minority, render the pro-environmental behaviour NP resistant to the presence of NE, and thus increase the payoff of CE types. In other words, negative assortativity (i.e. heterophilic preferences in agent matching) between SnE and NE enables a favourable condition for CE to diffuse in the population. The attempt of SnE to be the only one to play NP leads to a population where everyone plays it (a similar paradoxical result is shown in Smaldino and Epstein 2015). Note also that in the attractive vertex where all agents play NP, no SnE is left in the population. Even when identity-driven individuals are present only temporarily, it still leads to the diffusion of the sustainable behaviour they promote (and possibly to the overall increase in well-being of the population).

Second, another attractive steady state may exist when the suboptimal payoff of CE is sufficiently low; that is, $\alpha > \gamma$. This attractive state lies in the left-hand edge, where $x_{CE} = 0$, and the population is composed of only NE and SnE. Indeed, we have shown that both types receive the maximum payoff of 1 when they meet such that their payoff differential is null, and none has an incentive to imitate the other. Moreover, as CE receives an excessively low payoff when paired with NE, they are unable to diffuse in the population.

Finally, we find that no attractive steady state exists in the interior of S, where all three agent types are present.

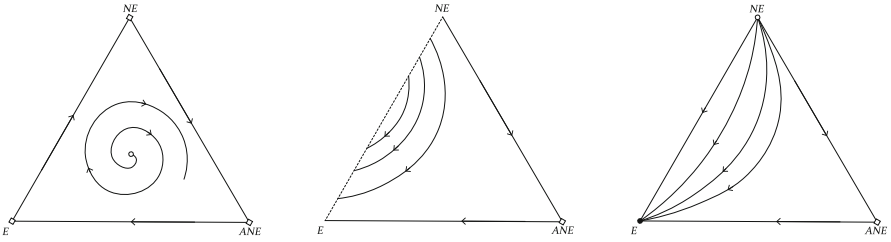
6 Social norm scenario

In this section, we analyse the dynamics of a population composed of environmentalists (E), non-environmentalists (NE), and ashamed non-environmentalists (ANE). Furthermore, we present a full classification of cases, including non-robust ones. We can obtain this population by substituting CE in the benchmark scenario with ANE. Here, the payoff matrix becomes:

$$\begin{array}{c|ccc} & E & ANE & NE \\ \hline E & (NP,NP) & (NP,NP) & (NP,P) \\ ANE & (NP,NP) & (P,P) & (P,P) \\ NE & (P,NP) & (P,P) & (P,P) \end{array} \qquad \begin{array}{c|ccc} & E & ANE & NE \\ \hline E & 1 & 1 & \alpha \\ ANE & \gamma & 1 & 1 \\ NE & 1 & \beta & \beta \end{array} \tag{7}$$

where the usual interpretation of payoff applies and $\alpha, \beta, \gamma < 1$. In this scenario, which we call social norm, E and NE have the same payoff structure as in the benchmark scenario, because ANE behaves similarly to CE. As for ANE, they achieve their highest payoff when matched with other ANE or with NE, because both play P, relieving ANE of the stigma of polluting or not contributing.

When we substitute CE in the benchmark scenario with ANE, we obtain two results (see Figs. 4a–c). First, the only steady state in an (E, NE, ANE)-population that can be attractive is vertex E. This becomes attractive when E dominates NE, that is when $\alpha > \beta$. Furthermore, when E and NE are matched with ANE, E receives maximum payoff whereas NE receives $\beta < 1$. This is so because both would want the other agent to care for the environment in their stead. This implies that E is in a better position with respect to NE when interacting with ANE. Here, the population converges towards a state where all agents abide by the environmental social norm NP. For instance, it could be that NE and ANE are so negatively affected by the level of environmental degradation and peer pressure/shame, respectively, that they gradually internalise the pro-environmental preferences of E agents. Since the difference with respect to the benchmark scenario lies in the presence of ANE instead of CE, we interpret this result in terms of greater efficacy of peer pressure effect (ANE) with regard to punishment (CE). Indeed, shame in the social norm scenario (just like stigma in Tavoni et al. 2012) makes a more desirable steady state attractive, although only a non-desirable steady state can be attractive in the benchmark scenario.



(a) Case: $\alpha < \beta$. (b) Case: $\alpha = \beta$. (c) Case: $\alpha > \beta$.

Fig. 4 Dynamics in (E, ANE, NE)-population

Finally, no stable internal steady state can be found in this scenario. However, an unstable steady state can exist under certain conditions based on the relative value of payoffs (see Fig. 4a). In this case, the trajectories oscillate indefinitely, approaching but not reaching the boundary of S. None of the three agent types leaves the population in these trajectories however close their share may get to zero.

7 Discussion and conclusions

In this study, we investigated the effect of heterogeneity in preference on the diffusion of pro-environmental (or more generally pro-social) behaviours. Our model stresses the importance of agents' preferences and their evolution with respect to their behaviour. In other words, we believe that it is not whether people abide by a social norm that matters most in its diffusion, but why they do so. We considered pro-environmental behaviour in our analysis, rather than the more general pro-social behaviour, because we focused on the pressing environmental challenges. However, we underline that the model can be easily generalised to pro-social behaviours whose diffusion can be affected by heterogeneity in preferences.

We started considering the behavioural triplet commonly found in the literature, defined here as environmentalists (E), non-environmentalists (NE), and conditional environmentalists (CE); we used this as our benchmark scenario. We introduced the snob environmentalist (SnE) in the identity scenario instead of the E type of the benchmark. We found that SnE agents make the pro-environmental behaviour more resistant to the presence of NE type than in the benchmark scenario. Analogously, in the social norm scenario, we introduced the ashamed non-environmentalist (ANE) type instead of the CE of the benchmark. We found that having the ANE type renders a state with only NE no longer attractive. Given that this state represents a situation where the entire population would choose not to adopt the pro-environmental behaviour, this is an improvement with respect to the benchmark scenario. Note that the SnE and ANE type agents follow the same behavioural pattern as the E and CE types that they substitute; that is, they choose the same action under the same circumstances and differ only in their preferences over outcomes. This behavioural equivalence indicates that under random matching with replicator dynamics, preferences are more determinant than agents' actions in the diffusion of a desirable behaviour. Furthermore, the SnE and ANE types relate to identity and peer pressure concepts, which have already attracted the attention of scholars for their role as behavioural drivers.

The role that preferences play in driving the diffusion of pro-environmental behaviours is of great relevance to policymakers. When designing a norm, it could be more effective to characterise such norm in terms of peer pressure and social identity. We may consider the case of the local authority of fictitious GreenVille wanting to prevent the degradation of an environmental common, that is a park or beach. Rather than setting an optimal fee to discourage littering, it would be more effective to promote a social campaign presenting the citizens of GreenVille as proud defenders of their

precious commons or exalting the importance of the latter in the culture of GreenVille. This seems to be especially true when the cost of switching to an environmental social norm is low, although persistent (Moore and Boldero 2017). More generally, enhancing the identity benefit of snob environmentalists or the social pressure felt by ashamed non-environmentalists could prove to be a convenient policy tool. In addition, in line with the findings of Nyborg et al. (2006), the mayor of GreenVille can promote the belief that the distribution of types is different from the actual one, in an attempt to push the population to respond by abiding by the environmental social norm. This might prove especially useful in case of multistability.

Beyond this fictitious example, we believe that a compelling follow-up to this work would be to study the different policies enacted and evaluate their success in the light of the framework proposed here. Similarly, it would be interesting to elicit the distribution of types (or preferences) in a population through an experiment or by exploiting the huge mass of data available in social networks. Furthermore, following Caravaggio and Sodini (2022), an agent-based model can highlight the local properties of our analytical framework to obtain our non-linear dynamics. Finally, even though this study focused on pro-environmental behaviours, our framework can be applied to other contexts where the heterogeneity of agent types is relevant to the diffusion of pro-social behaviours.

Appendix A: Mathematical Appendix

This appendix presents the mathematical results underlying the analyses presented in the study.

Main results for each population

Vertices E, NE, CE, SnE, and ANE of simplex S (see Figs. 1, 2, 3, and 4) represent states where, respectively, only agents of type E, NE, CE, SnE, or ANE are present in the population. An 'edge' of simplex S includes all the population states that do not adopt a given strategy; we denote $i - j$, where $i, j = E, NE, CE, SnE, ANE$, $i \neq j$, the edge where only types i and j are present in the population.

We use the terminology given in Bomze (1983); by 'eigenvalue (EV) of a steady state', we mean an eigenvalue of the linearisation matrix around the steady state. The term 'EV in the direction of vector V' means that V is an eigenvector corresponding to that EV. For simplicity, the propositions in Bomze (1983) are indicated as B♯ (e.g. B4 is Proposition 4 in Bomze's paper). Furthermore, we refer to the numbers relative to Bomze's classification of phase portraits (PP) on the two-dimensional simplex S using notation PP(♯) to indicate the PP number ♯. We denote the time reversal of the trajectories corresponding to PP(♯) by PP($-$♯).

We obtain the following basic mathematical results.

Results for the benchmark population

For a population composed of environmentalists, conditional environmentalists, and non-environmentalists, we obtain the following results.

Proposition 1 *An infinite number of steady states exist under this dynamic; in particular, we have*

(1) Edge E–CE is always pointwise fixed (i.e. each point belonging to it is a steady state).
(2) Edge E–NE is pointwise fixed if and only if (iff) $\alpha = \gamma$; in the other cases, no steady state exists in the interior of edge E-NE.
(3) A unique steady state exists in the interior of edge NE–CE iff $\beta < \gamma$; in this case, the eigenvalue structure of the steady state is $(1-\gamma)(\gamma-\beta)/(1-\beta) > 0$ in the direction of edge NE–CE and $(1-\gamma)(\alpha-\beta)/(1-\beta) > 0$ for $\alpha > \beta$.
(4) An infinite number of steady states exist in the interior of simplex S iff $\alpha = \beta < \gamma$; in the other cases, there are no steady states in the interior of S.

Proof This is derived from the application of B2 and B5.

Proposition 2 *The eigenvalue structure of the pure population steady states E, CE, and NE is as follows:*

(1) E has both eigenvalues equal to zero.
(2) CE an eigenvalue equal to 0 in the direction of edge E–CE and an eigenvalue with the sign of $\gamma - 1 < 0$ in the direction of edge NE-CE.
(3) NE has an eigenvalue with the sign of $\alpha - \gamma$ in the direction of edge E–NE and an eigenvalue with the sign of $\beta - \gamma$ in the direction of edge CE–NE.

Proof This is an application of B1♯.

From the above results and using Bomze's classification, we find that the (robust) PP that can be observed in this regime are those illustrated in Figs. 1–2 (which correspond to PP(25)–(27), PP(-29), and PP(32)).

Results for a (*NE*, *SnE*, *CE*) population

For the population composed of non-environmentalists, snob environmentalists, and conditional environmentalists, the following results hold:

Proposition 3 *The eigenvalue structure of the pure population steady-states NE, SnE, and CE is as follows:*

(1) NE has an eigenvalue with the sign of $1 - \alpha > 0$ in the direction of edge SnE–NE and an eigenvalue with the sign of $\gamma - \alpha$ in the direction of CE–NE.
(2) SnE has an eigenvalue with the sign of $1 - \beta > 0$ in the direction of both edge NE–SnE and edge CE–SnE. Thus, this is always a repulsive steady state.

(3) CE has an eigenvalue with the sign of $\alpha - 1 < 0$ *in the direction of edge NE–CE and an eigenvalue with the sign of* $\beta - 1 < 0$ *in the direction of edge SnE–CE. Thus, this is always an attractive steady state.*

Proof This is an application of B1♯.

Proposition 4 *The steady states structure in the interior of the edges is as follows:*

(1) A unique steady state always exists in the interior of edge NE-SnE; the sign of its eigenvalues is equal to the sign of $\alpha - 1 < 0$ *in the direction of edge NE–SnE and to the sign of* $\gamma - \alpha$ *in the direction of the interior of the simplex S.*
(2) A unique steady state exists in the interior of edge NE–CE iff $\alpha > \gamma$*; in this case, the sign of its eigenvalues is equal to that of* $\alpha - \gamma > 0$ *in the direction of edge NE–CE and to the sign of* $(1-\alpha)^2 - (\beta-\alpha)(\gamma-\alpha)$ *in the direction of the interior of the simplex S. Edge NE–CE cannot be pointwise fixed.*
(3) No steady state exists in the interior of edge SnE–CE.

Proof This depends on the application of B2 and B5♯.

Proposition 5 *A unique steady state exists in the interior of S iff* $\alpha > \gamma$ *and* $(1-\alpha)^2 > (\beta - \alpha)(\gamma - \alpha)$. *No steady state exists in the other cases.*

Proof This is an application of B6♯.

From the above results and using Bomze's classification, we find that the PP observed in this regime are those illustrated in Fig. 3 (corresponding to PP(11), PP(36), and PP(40)).

Results for an (*E*, *NE*, *ANE*) population

Finally, we present the results for a population composed of environmentalists, non-environmentalists, and ashamed non-environmentalists.

Proposition 6 *The eigenvalue structure of the pure population steady states E, NE, and ANE is as follows:*

(1) E has an eigenvalue equal to 0 in the direction of edge NE–E and an eigenvalue with the sign of $\gamma - 1 < 0$ *in the direction of edge ANE–E.*
(2) NE has an eigenvalue with the sign of $\alpha - \beta$ *in the direction of edge E–NE and an eigenvalue with the sign of* $1 - \beta > 0$ *in the direction of edge ANE–NE.*
(3) ANE has an eigenvalue equal to 0 *in the direction of edge E–ANE and an eigenvalue with the sign of* $\beta - 1 < 0$ *in the direction of edge NE–ANE*

Proof This is an application of B1♯.

Proposition 7 *The steady-state structure in the interior of the edges is as follows:*

(1) Edge E–NE is pointwise fixed iff $\alpha = \beta$*; no steady state exists in the other cases.*
(2) No steady states exist in edges E–ANE and NE–ANE.

Proof This is an application of B2 and B5♯.

Proposition 8 *A unique steady state exists in the interior of S iff* $\alpha < \beta$. *No steady state exists in the other cases.*

Proof This is an application of B6♯.

From the above results and using Bomze's classification, we find that the PP observed in this regime are those illustrated in Fig. 4 (corresponding to PP(-17), PP(-33), and PP(43)).

Acknowledgements While taking full responsibility for this paper, we thank Pier Luigi Sacco for his many insightful conversations on the subject. In addition, Giulio Galdi gratefully acknowledges the research grant received from the Department of International and Political Sciences of the University of Siena (ref. DDG 58/2019, n. 3289).

Declarations

Conflicts of interest The authors declare that they have no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Akçay E, Cleve JV (2012) Behavioral responses in structured populations pave the way to group optimality. The American Naturalist 179(2):257–269

Akerlof GA, Kranton RE et al (2010) Identity economics: How our identities shape our work, wages, and well-being. Princeton University Press, NJ, USA

Andreoni J (1990) Impure altruism and donations to public goods: A theory of warm-glow giving. The Economic Journal 100(401):464–477

Andreoni J (1995) Cooperation in public-goods experiments: kindness or confusion? The American Economic Review 85:891–904

Antoci, A., Sacco, P. L., and Zamagni, S. The ecology of altruistic motivations in triadic social environments. In *IEA Conference Volume Series*, volume 130, pages 335–351. Basingstoke; Macmillan Press; New York; St Martin's Press; 1998, 2000

Antoci A, Bellanca N, Galdi G (2018) At the relational crossroads: Narrative selection, contamination, biodiversity in trans-local contexts. Journal of Economic Behavior & Organization 150:98–113

Axelrod R, Hamilton WD (1981) The evolution of cooperation. Science 211(4489):1390–1396

Bar-Gill O, Fershtman C (2005) Public policy with endogenous preferences. Journal of Public Economic Theory 7(5):841–857

Barbarossa C, De Pelsmacker P, Moons I (2017) Personal values, green self-identity and electric car adoption. Ecological Economics 140:190–200

Bauwens T, Devine-Wright P (2018) Positive energies? an empirical study of community energy participation and attitudes to renewable energy. Energy Policy 118:612–625

Bisin A, Verdier T (2005) Cultural transmission. In: Durlauf SN, Blume LE (eds) The New Palgrave Dictionary of Economics. Palgrave Macmillan, London, UK

Björnerstedt, J. and Weibull, J. W. Nash equilibrium and evolution by imitation. In *The Rational Foundations of Economic Behaviour*, pages 155–171. Macmillan, London, UK, 1996
Bruhin A, Fehr E, Schunk D (2016) The many faces of human sociality: Uncovering the distribution and stability of social preferences. Journal of the European Economic Association 17:1025–1069
Bruvoll A, Nyborg K (2004) The cold shiver of not giving enough: on the social cost of recycling campaigns. Land Economics 80(4):539–549
Caravaggio A, Sodini M (2022) Local environmental quality and heterogeneity in an olg agent-based model with spatial externalities. Journal of Economic Interaction and Coordination 17(1):287–317
Chaudhuri A (2011) Sustaining cooperation in laboratory public goods experiments: a selective survey of the literature. Experimental economics 14(1):47–83
Confente I, Scarpi D, Russo I (2020) Marketing a new generation of bio-plastics products for a circular economy: The role of green self-identity, self-congruity, and perceived value. Journal of Business Research 112:431–439
de Zeeuw A (2015) International environmental agreements. Annu. Rev. Resour. Econ. 7(1):151–168
Doebeli M, Hauert C, Killingback T (2004) The evolutionary origin of cooperators and defectors. Science 306(5697):859–862
Drouvelis M, Georgantzis N (2019) Does revealing personality data affect prosocial behaviour? Journal of Economic Behavior & Organization 159:409–420
Ferguson E, Flynn N (2016) Moral relativism as a disconnect between behavioural and experienced warm glow. Journal of Economic Psychology 56:163–175
Fischbacher U, Gächter S (2010) Social preferences, beliefs, and the dynamics of free riding in public goods experiments. American Economic Review 100(1):541–56
Fischbacher U, Gächter S, Fehr E (2001) Are people conditionally cooperative? evidence from a public goods experiment. Economics letters 71(3):397–404
Gkargkavouzi A, Halkos G, Matsiori S (2019) Assessing values, attitudes and threats towards marine biodiversity in a greek coastal port city and their interrelationships. Ocean & Coastal Management 167:115–126
Gkargkavouzi A, Halkos G, Matsiori S (2019) How do motives and knowledge relate to intention to perform environmental behavior? assessing the mediating role of constraints. Ecological Economics 165:106394
Gul F, Pesendorfer W (2016) Interdependent preference models as a theory of intentions. Journal of Economic Theory 165:179–208
Klaser K, Sacconi L, Faillo M (2021) John rawls and compliance to climate change agreements: insights from a laboratory experiment. International Environmental Agreements: Politics, Law and Economics 21(3):531–551
Lehmann L, Alger I, Weibull J (2015) Does evolution lead to maximizing behavior? Evolution 69(7):1858–1873
Lindenberg S, Steg L (2007) Normative, gain and hedonic goal frames guiding environmental behavior. Journal of Social Issues 63(1):117–137
Mancha RM, Yoder CY (2015) Cultural antecedents of green behavioral intent: An environmental theory of planned behavior. Journal of Environmental Psychology 43:145–154
Moore HE, Boldero J (2017) Designing interventions that last: a classification of environmental behaviors in relation to the activities, costs, and effort involved for adoption and maintenance. Frontiers in Psychology 8:1874
Nyborg K, Howarth RB, Brekke KA (2006) Green consumers and public policy: On socially contingent moral motivation. Resource and Energy Economics 28(4):351–366
Richter A, Grasman J (2013) The transmission of sustainable harvesting norms when agents are conditionally cooperative. Ecological Economics 93:202–209
Sacco PL, Zamagni S (1994) Un approccio dinamico evolutivo all'altruismo. Rivista Internazionale di Scienze Sociali 2:223–261
Samuelson L (1997) Evolutionary Games and Equilibrium Selection. The MIT Press, Cambridge, MA, USA
Schlag KH (1998) Why imitate, and if so, how?: A boundedly rational approach to multi-armed bandits. Journal of Economic Theory 78(1):130–156
Smaldino PE, Epstein JM (2015) Social conformity despite individual preferences for distinctiveness. Royal Society Open Science 2(3):140437

Steg L, Bolderdijk JW, Keizer K, Perlaviciute G (2014) An integrated framework for encouraging pro-environmental behaviour: The role of values, situational factors and goals. Journal of Environmental Psychology 38:104–115

Tavoni A, Schlüter M, Levin S (2012) The survival of the conformist: social pressure and renewable resource management. Journal of Theoretical Biology 299:152–161

Taylor PD, Jonker LB (1978) Evolutionary stable strategies and game dynamics. Mathematical Biosciences 40(1–2):145–156

Weibull J (1995) Evolutionary Game Theory. The MIT Press, Cambridge, MA, USA

Whitmarsh L, O'Neill S (2010) Green identity, green living? the role of pro-environmental self-identity in determining consistency across diverse pro-environmental behaviours. Journal of Environmental Psychology 30(3):305–314

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.