Journal of Economic Interaction and Coordination (2022) 17:613–637
https://doi.org/10.1007/s11403-022-00347-8

REGULAR ARTICLE

Staring at the Abyss: a neurocognitive grounded agent-based model of collective-risk social dilemma under the threat of environmental disaster

Danilo Liuzzi[1] · **Aymeric Vié**[2,3,4,5,6]

Received: 7 October 2019 / Accepted: 18 December 2021 / Published online: 28 February 2022

Abstract

Increasingly visible climate change consequences challenge carbon-based economies worldwide. While expert knowledge on climate change percolates through political initiatives and public awareness, its translation into large-scale policy actions appears limited. Climate change consequences unequally target regions, countries and social classes, a vital issue for social cooperation. When facing an imminent ecological collapse, in which conditions can self-interested agents gain environmental awareness and settle on a sustainable path of actions when their knowledge of the imminent collapse is bounded? This cooperation emerges from the interaction between individuals and the interaction of various cognitive processes within individuals. This article develops an agent-based model for this emergence of cooperation enriched with the Agent Zero neurocognitive grounded cognitive architecture. We investigate when agents endowed with deliberative, affective and social modules can settle on actions that safeguard their environment through numerical simulations. Our results show that cooperation on sustainable actions is the strongest when the system is at the edge of collapse. Policy measures that increase the environment's resilience become internalized by the agents and undermine awareness of the ecological catastrophe. Depending on the cognitive channels activated, agent behaviors and reactions to specific interven-

✉ Aymeric Vié
vie@maths.ox.ac.uk

Danilo Liuzzi
danilo.liuzzi@unimi.it

1 Department of Economics, Management and Quantitative Methods, University of Milan, Milan, Italy

2 Oxford Institute of New Economic Thinking, University of Oxford, Oxford, UK

3 Mathematical Institute, University of Oxford, Oxford, UK

4 New England Complex Systems Institute, Cambridge, USA

5 Paris School of Economics, Paris, France

6 Sciences Po St Germain en Laye, St Germain en Laye, France

tions significantly vary. Our analysis suggests that taking different cognitive channels, deliberative, affective, social, and others into account, significantly impact results. The complexity of agent cognition deserves more attention to assess parameter sensitivity in social simulation models.

Keywords Agent-based model · Dispositional contagion · Social dilemma · Regime shifts · Climate change

1 Introduction

Wildfires, flooding, hurricanes: more extreme weather events attributed to climate change seem to happen each year. Numerous treaties, initiatives, regulations, reports investigate the matter and propose various actions to limit the warming extent. However, environmental awareness as consciousness in the population and political structures of the extend of the threat appears limited, as the undertaken actions fail to match the severity of the danger (Eriksen et al. 2014). One critical obstacle to such awareness may be the unequal distribution of climate change consequences on the planet (King and Harrington 2018). Countries and populations less harmed by climate change may not internalize the extent of the threat for more impacted communities.

Moreover, several difficulties arise when studying the emergence of awareness in a population facing an ecological danger. The complexity of interactions and emergence at the population level have been intensively studied, from the tragedy of the commons (Ostrom 1990) to more recent literature on cooperation reviewed in Sect. 2. We add that even within a single individual, such a decision is not straightforward. One may become aware of climate change and willing to change one's actions for deliberative reasons, such as reading the scientific literature on the topic. Another may adopt this stance facing the emotion of a newsworthy catastrophe somewhere on the planet. Others may be sensitive to the adoption of this opinion in their social circles. In sum, there are many ways environmental awareness could arise.

We propose an agent-based model that explores the dynamics of a collective-risk social dilemma. Uncooperative and myopic exploitation of the environment can trigger a detrimental regime shift, where environmental resources and services are no longer available. In this failure of the social dilemma, resources are only restored once hysteretic dynamics have unfolded. We model the environment as a (shallow) lake (Scheffer 1989), a bistable system where either a clean state or a turbid state can be reached, depending on the history of pollution discharge into the lake. Whenever the turbid state arises, agents must temporarily cease polluting activities, bowing to the restoration they cannot manage. Agent decisions, aggregated over the entire population, may push the lake toward a polluted (turbid) state, but the way back to the clean state is out of their hands.

The agents are endowed with more natural cognitive abilities, building on the Agent Zero framework (Epstein 2014). Deliberative, affective and social components are behind the agent decisions, accounting for both bounded rationality and the plurality of channels by which dispositions to act and opinions are transmitted. This multifaceted model allows us to obtain a more precise understanding of the emergent process

of cooperation induced by environmental awareness, and in so doing, it helps us understand how to prevent the transition to the polluted state possibly.

Simulation results emphasize that environmental awareness is maximum when the socio-environmental system reaches critical pollution levels. Moreover, more resilient natural conditions, such as higher natural pollution decay, do not increase awareness, as the agents internalize these constraints and are encouraged to pollute more. These results explain why, in collective-risk social dilemmas, cooperation is the strongest when the system is at the edge of collapse and that more resistant environments (lower probability of collapse) may shelter a less environmentally concerned population. We analyze how the activation of different modules, or combinations in the Agent Zero framework, impacts how environmental awareness arise. As significant differences arise, as in other uses of this framework (Vié 2019), we suggest that this superposition of cognitive channels for decision-making is helpful for social simulation.

The paper proceeds as follows. Section 2 describes the relevant literature on the topic of social cooperation and environmental dilemmas. Section 3 presents the details of the model. Section 4 presents the results, while Sect. 5 discusses the main findings of the paper. Section 6 concludes and presents avenues for future research.

2 Related work

Humans have come to dominate this planet thanks to their ability to cooperate efficiently and in (extremely) large groups (Wilson 2012; Harari 2014; Turchin 2016). The global economy is an example of the feats of cooperation involving billions of individuals across the globe. However, when it comes to the threat of global warming, the bets are on the dire outcomes, as the ever-increasing global level of atmospheric CO_2[1] and the more and more frequent extreme weather events[2] (e.g., fires, heatwaves, droughts, floods) seem to suggest. A single agent cannot possibly tame global warming, be it a person, a country or a region.

In his 1968 most influential paper, "The Tragedy of the Commons" (Hardin 1968), Garret Hardin warned us that the tragedy is the unavoidable outcome when the management of a common exhaustible resource is left to the consciences of the agents who take advantage from it (Hardin himself recasts this problem as waste disposal in a familiar environment). The paper describes a common pasture to which several herders can bring their animals. Each herder has a strong incentive to increase her herd, given that the personal benefit of adding an animal to the common pasture is more significant than the cost shared among all the herders. In Hardin's view, the dilemma between the personal gain and the collective interest has permanently ruined the common resources as a final result, unless a fundamental extension in morality steps in or a regime of "mutual coercion mutually agreed upon" is enforced. Fortunately, it is not always true that without the intervention of a sanctioning authority (e.g., a coercive state), supposedly selfish agents end up destroying a common pool resource, as convincingly pointed out by the Nobel awarded Elinor Ostrom in her book "Governing The Com-

[1] See https://www.climate.gov for updated statistics.

[2] See the IPCC report at https://www.ipcc.ch/.

mons" (Ostrom 1990). Social dilemmas in which the individually rational decision does not align with the collective optimal outcome are ubiquitous. These dilemmas resemble, theoretically and experimentally, Public Goods Games (PGG). In a typical PGG, participants receive an initial monetary endowment and devote a portion of it to a shared pool. The total amount collected in the shared pool is then multiplied by an enhancement factor s (usually $s = 2$) and equally redistributed to all the participants that walk out of the experiment with the amount accumulated in their account. The rational or selfish choice is not to contribute to the shared pool, possibly ripping the benefit of the others' contributions. Conversely, the collective optimal outcome is reached when each individual transfers her endowment completely to the shared pool. In the absence of specific mechanisms, cooperation almost always declines and vanishes in repeated PGGs (Ostrom et al. 1992; Fehr and Gachter 2000; Rand et al. 2009). Cooperation supporting mechanisms have been studied extensively in the literature (see Nowak 2006, 2011; Rand 2013). Reduced levels of effectiveness of nepotism and reciprocity lead to indirect reciprocity or reputation to take over (Milinski 2002; Nowak and Sigmund 1998, 2005; Jacquet et al. 2011; Milinski 2016; Santos et al. 2018). Network reciprocity or spatially structured populations (Santos et al. 2006; Szabo and Gabor 2007; Righi and Takács 2018), as well as solid reciprocity and altruistic punishment (Fehr and Gachter 2000, 2002; Gächter and Herrmann 2009) and their interaction (Rockenbach and Milinski 2006; Rand et al. 2009), have been found to bolster cooperation.[3] Unfortunately, global warming and other social dilemmas that can lead to catastrophic outcomes cannot be framed as standard PGG. In this respect, some modifications of the standard PGG could provide additional insights. A variant of the PGGs is the Threshold Public Goods Games (TPGG) framework, where the provision of a public good is conditional to the agents' total contribution being equal to or above a certain threshold. An example could be when a community needs individual voluntary contributions to build a dam or erect a defensive wall. Again, the incentives are on the free-riding strategy, but the benefit only materializes if enough contributors live within the community. A version of the TPGG particularly suited to the modeling of the so-called Collective-Risk Social Dilemmas represents the global warming threat. Here the collective failure to reach the threshold exposes the agents to the risk of losing everything—even the initial personal endowments—with a certain probability. In a Collective-Risk Social Dilemma, personal knowledge about the nature and the severity of the prospective consequences of collective failure is crucial to the emergence of cooperation. In particular, the risk of collective failure, i.e., the perceived probability of future losses, plays an essential role. The higher the risk, the easier the cooperative outcome (Milinski et al. 2008; Wang et al. 2009; Santos and Pacheco 2011; Santos et al. 2012; Du et al. 2012, 2014).

The conundrum of the emergence of cooperation among supposedly self-centered individuals has attracted attention from many different fields in recent years, and, as a consequence, many are the research methods employed. In the literature mentioned above, the topic at hand has been tackled both with experimental settings (for example, Milinski et al. 2008) and with evolutionary game-theoretical frameworks (for example, Wang et al. 2009). An approach that can harness the power of both exper-

[3] For an interdisciplinary unifying vocabulary of the above concepts see West et al. (2007).

iments and theory, induction and deduction, is Agent-Based Modeling (ABM). As Robert Axelrod clearly states in the introduction of his 1997 book "The Complexity of Cooperation" (Axelrod 1997), ABM can both be built on a firmly defined set of assumptions, like deduction, and create artificial data analyzed with induction. These characteristics qualify the ABMs as a fundamental tool to understand and mimic emergent phenomena. ABMs represent agents as computer algorithms: agents have a set of decision rules, interact with the environment and other agents, learn from and adapt to the changing environment they contribute to creating. Therefore, the ABMs approach is bottom-up: from the agents' micro-rules of behavior, the idea is to detect patterns of emergent meso- and macro-phenomena (Holland and Miller 1991; Bonabeau 2002; Helbing and Balietti 2011). There is a wide range of applications of ABMs to social sciences, from studies on segregation, cultural evolution and organizational behavior (Schelling 1978, Epstein and Axtell 1996; Lazer and Friedman 2007) to works on opinion dynamics, societies and nation-states formation (Deffuant et al. 2000; Epstein 2006; Cederman 1997). Seminal ABM works on the emergence of cooperation are Axelrod (1984, 1997). "A Cooperative Species" (Bowles and Gintis 2011) discusses proximate and ultimate causes for the existence of cooperation among humans, exploring the subject with a plurality of ABM approaches. In general, provided that agents are replaced by strategies in the stage game and strategies reproduce proportionally to their success, an evolutionary agent-based model can be at the core of analyzing a model concerning social dilemmas (for example, Bowles et al. 2004). Central to the ABM literature and the present work is the possibility of ABMs representing agents with cognitive abilities more in line with actual human beings than the standard economic approach referred to as "Homo Economicus" (Henrich et al. 2001). In Joshua M. Epstein Agent Zero (Epstein 2014), the set of decision rules of the agents are grounded in the state-of-the-art neuroscience literature, allowing for a realistic description of the agents' behavior determinants.

The central question of this paper revolves around whether a community of cognitively human-like agents under the imminent threat of an environmental collapse could cooperatively refrain from polluting and, in so doing, avoid the tragedy of the commons without the intervention of an external/policing authority. Furthermore, our answer is affirmative; there is hope: agents with complex inner life can self-organize to save their environment when a grave danger is looming over them. We take inspiration from the Collective-Risk Social Dilemma devised in Milinski et al. (2008), with two crucial differences. Firstly, agents do not choose whether to pool their resources to reach a target, but they choose whether to restrain from polluting to avoid the destruction of the environment. Secondly and crucially, the economic considerations are absent from the agents' decision process.

The reason for this peculiar modeling decision lies in the focus we put on the proximity of the disaster, on the imminence of the threat, where personal gain can reasonably be put aside. In our model, agents know that the environmental disaster will happen; they do not know when. The agents do not know the whereabouts of the tipping point. The uncertainty lies in personal knowledge (Tavoni et al. 2011; Barrett and Dannenberg 2012; Dannenberg et al. 2015; Hagel et al. 2016, 2017; Kumar and Dutt 2019). We adopt a deterministic threshold-based tipping point, unlike Milinski et al. (2008), where the environmental disaster would have happened with a

certain probability if the agents had not reached the target. We employ the Agent Zero framework (Epstein 2014), whose details will be described in the next section, precisely because we want to take into account the bounded rationality of individuals (Raihani and David 2011), together with the human-like characteristic of dispositional imitation, where a cascade of similar decisions can start without a leader paving the way. Agent Zero embodies an agent whose rules of behavior are in line with modern neuroscience literature. In its most straightforward formulation, the agent "Agent Zero" takes action if its total disposition exceeds a threshold. The total disposition of Agent Zero includes three components: emotional or affective, rational or deliberative and social. The emotional component follows the Rescorla-Wagner equation of conditioning: a series of pairings sums up and strengthen the association between the unconditional and conditional stimulus. The rational component evaluates the evidence available to the agent. In this respect, the evidence is not comprehensive but limited to the range of vision of the agent. This limitation allows for a description of the agent's rationality consistent with the one proposed in the "bounded rationality" literature. The social component considers the dispositions of the other agents that constitute the network of agents surrounding an agent zero, all endowed with the same general characteristic of the prototypical Agent Zero. It is essential to underline that in the social components, it is not the other agents' action that affects Agent Zero' action but the disposition of the other agents. So we do not imitate behavior, but dispositional imitation: the first agent who takes action is not necessarily the leader. In other words, we study how Agent Zero's agents interact with each other and with their environment to harness the imperfect information they have and whether this is enough to avert the tragedy. We find that such agents can develop a distributed environmental awareness and self-organize critically to avoid catastrophe.

3 The model

3.1 The shallow lake environment

The Shallow Lake (henceforth SL) is a potent metaphor for a social-ecological system (see Scheffer 1989). Humans depend on the lake along many different dimensions, from the daily provision of food and clean water to mitigation of temperature gradients, from recreational activities to stock of edible biomass. Human economic activities affect the status of the lake heavily. Urban sewage systems and industrial waste together increase the concentration of nutrients in the lake, e.g., phosphorus. In turn, the changed nutrients distribution increases the presence of planktonic algae that prevents the light from reaching the lake's bottom, making it difficult for the submerged plants to survive. Eventually, the algae take over, and the lake becomes turbid, at the end of a process known as eutrophication. The management of the ecosystem is particularly complicated because the human activities and the lake dynamics unfold at different time scales. The causes of the lake's eutrophication are diffused and occur slower than the typical time scale of human decisions. The consequences are the regime shift from clean to turbid state, which occurs abruptly, and is immediately visible to anyone. Moreover, the eutrophication process is inherently path-dependent; in other words, it

shows hysteris: reducing the level of nutrients load in the lake to a level antecedent to the eutrophication does not restore the lake to its clean state, but a drastic and long-lasting reduction of the load is often the only remedy. Significant references in the literature of the SL are Scheffer (1989, 1998); Scheffer et al. (2001) and Mäler et al. (2003). An inspirational ABM paper about the shallow lake narrative is Martin and Schlüter (2015). Unlike this literature, we do not model the SL dynamics with a system of differential equations, but we use a binary variable representing the two regimes off the lake: clean or turbid. A threshold on the level of waste load separates the clean from the turbid state. Once the lake becomes turbid, a restoring period is needed for the lake to come back to life. We explicit the dynamics of the lake in our model below.

We consider a lake, accounting for the environment state. The lake can have two states summarized by the dummy variable S: clean ($S = 0$) and polluted ($S = 1$). The lake is clean if pollution is below a threshold p^*. The lake has a natural absorption rate δ, a natural amount of pollution it can remove per unit of time without any external intervention. At each period in a discrete-time setting, each agent in a population of n individuals takes action among the choices described by the decision variable X. Three actions are possible. Hence, X can take three mutually exclusive values. $X = 0$ means that the agent renounces to the personal gain from exploiting the natural resource, emitting no pollution externality. $X = 1$ means that the agent exploits the environmental resources in a sustainable way (whose precise interpretation will be clarified below), generating a pollution externality in the form of an amount π of pollution released in the lake. $X = 2$ means that the agent takes a higher amount from the natural resource but releases an equal amount of pollution 2π in the lake. We do not explicitly model the natural resource, but just the environmental degradation that results from its exploitation—we call it pollution, in general. Neither do we model the gain the agents obtain from the resource exploitation. As stated in the introduction, we are only interested in the health status of the lake and in the way agents react when this health status is seriously under threat. The decision rule of the agents proceeds by comparing dispositions in favor of sustainable exploitation, over-exploitation or abstention. We detail the disposition modeling below. The pollution level of the lake denoted p_t at time t follows a dynamic rule given by the aggregation of the pollution externalities emitted by all agents' i actions in n at time t minus the amount of self-cleaned units the lake manages to take rid of at the same time:

$$p_{t+1} = p_t + \sum_{i}^{N} \pi_{it} - \delta \mathbb{1}\{p_t < p^*\} p_t - r\delta \mathbb{1}\{p_t \geq p^*\} p_t, \qquad p_0 \text{ given.} \quad (1)$$

The natural pollution absorption rate δ varies depending on the state of the lake. While a clean lake fully recovers at rate δ, the polluted lake pollution decay is affected. The parameter $r \in [0, 1]$ measures how the environmental collapse affects this natural recovery. As $r \to 0$, this collapse becomes irreversible, as the natural pollution decay in the turbid lake becomes negligible. When $r = 1$, the state change of the Shallow Lake has no impact on its natural self-cleaning. The Shallow Lake starts in a pristine state with zero pollution. If the pollution reaches the threshold $p_t = p^*$, then the lake

flips to its polluted (turbid) state, destroying the ability of the agents to exploit the lake and hindering the self-cleaning process to the value of the parameter r.

3.2 Agent's decision-making mechanism

In a call for greater cognitive realism in generative social science, Epstein (2014) proposed a neurocognitively grounded decision model to generate more "developed and conflicted inner lives" of agents. In this model, the observable behavior of agent zero stems from interaction (conflict) of affective components (based on the Rescorla-Wagner Model of conditioning and extinction), deliberative (heuristics, sample selection bias) and social forces (contagion effects). Binary actions occur by exceeding thresholds that may be heterogeneous and susceptible to dispositional contagion. Epstein's agent i takes action F at time t if her total disposition $\Delta_{F,it}$ toward that action goes beyond the threshold T_F:

$$\Delta_{F,it} = D_{F,it} + A_{F,it} + S_{F,t} \geq T_F \tag{2}$$

In Eq. 2, the total disposition is the sum of the solo deliberative component $D_{F,it}$, the solo affective component $A_{F,it}$, and the contribution to the disposition given by the social component organized in a network $S_{F,t}$. With some appropriate modifications, but in line with Epstein's Agent Zero, we model the ternary decision set of the agent i, $X = 0$, abstain, $X = 1$, exploit sustainably, or $X = 2$, fully exploit, according to which one of the following disposition prevails:

$$X_t = \begin{cases} 0 & \text{if } \max(\Delta_{0,it}, \Delta_{1,it}, \Delta_{2,it}) = \Delta_{0,it} \\ 1 & \text{if } \max(\Delta_{0,it}, \Delta_{1,it}, \Delta_{2,it}) = \Delta_{1,it} \\ 2 & \text{if } \max(\Delta_{0,it}, \Delta_{1,it}, \Delta_{2,it}) = \Delta_{2,it} \end{cases} \tag{3}$$

In the Agent Zero framework context, agents do not cross a threshold but choose the action with the highest disposition. The following subsections show how individuals calculate those dispositions.

3.2.1 Solo disposition: the deliberative component

Now we first describe the two constituents of the solo disposition, then provide details on the social component. In our variation of the solo deliberative component of Agent Zero, agents receive a signal denoted $\hat{p}_{it}$ about the lake pollution level. The signal is imperfect and is identically and independently distributed in a Gaussian distribution. σ_p^2 accounts for the precision of the emitted signal on the lake current pollution level.

$$\hat{p}_{it} = p_t + \varepsilon_{pi} \text{ where } \varepsilon_p \sim N(0, \sigma_p^2) \tag{4}$$

Agents start with two thresholds that materialize beliefs on the lake dynamics. Here each agent compares the signal on lake pollution to confidence and an alarm threshold.

If the signal is below the confidence level, the agent gets strong confidence in the lake state and has incentives to exploit as much as possible, certain that this exploitation will not be detrimental to the lake state. If the inferred pollution lies between the confidence and alarm threshold, concerns and awareness on environmental disasters emerge in the agent's mind, that will be disposed toward sustainable exploitation. If the inferred pollution level is higher than the alarm threshold, the agent internalizes the threat of environmental collapse and has a strong disposition to cease exploitation momentarily. The alarm threshold denoted T_a for each agent i is identically and independently distributed according to a Gaussian distribution centered around the actual lake pollution threshold. σ_t^2 measures the spread of these beliefs in the population. Homogeneous beliefs on lake threshold is obtained with $\sigma_t^2 = 0$.

$$T_{ai} = p^* + \varepsilon_t \text{ where } \varepsilon_{it} \sim N(0, \sigma_t^2) \tag{5}$$

Another belief, the confidence threshold T_c for each agent i, is derived from the alarm threshold as its fraction. Agents' caution is captured by the parameter $c \in [0, 1]$. When this value equals 0, agents are extremely risk-averse: any pollution signal below the alarm threshold will result in environmental threat awareness and disposition toward sustainable exploitation. However, if $c = 1$, the confidence and alarm thresholds do match, preventing concerns to arise gradually. These overconfident agents are unaffected by growing lake pollution until this level overcomes the alarm threshold.

$$T_{ci} = cT_{ai} \tag{6}$$

From Eqs. 4 to 6, it is clear that heterogeneity in agents' actions becomes inherently related to the variance in the lake pollution signal and the diversity in threshold beliefs. In other words, the inherent uncertainty in the environmental problem lies in the imperfect information the agents possess (Barrett and Dannenberg 2012; Dannenberg et al. 2015; Hagel et al. 2017; Kumar and Dutt 2019) and not in the consequences of surpassing the threshold, as in Milinski et al. (2008). Information coming from the environmental shocks (see next subsection) and channelled via networks plays a fundamental role in shaping the behavior of the agents. We denote $D_{0,it}$, $D_{1,it}$ and $D_{2,it}$ the dispositions of agent i at period t to, respectively, abstain, exploit reasonably and fully the lake. The logistic function inspires these disposition equations. The idea is to set that the disposition to adopt a given action is marginally maximal when the inferred pollution level lies in the exact median of that specific action threshold range. Temporary deviations around the median of the interval have a stronger impact on the increase (decrease) of disposition than deviations far from the median.

$$D_{0,it} = \begin{cases} 0 & \text{if } \hat{p}_{it} < T_{ci} \\ \left(1 + \exp(-p_{it} + \frac{T_{ai}+T_{ci}}{2})\right)^{-1} & \text{if } T_{ci} \leq \hat{p}_{it} < T_{ai} \\ 1 & \text{if } \hat{p}_{it} \geq T_{ai} \end{cases} \tag{7}$$

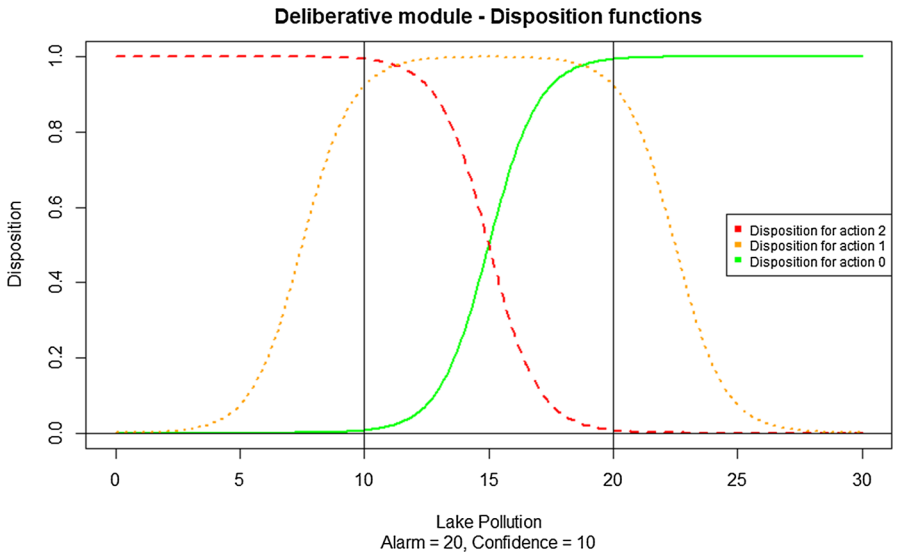

Fig. 1 Disposition functions for a lake threshold of 20, and a confidence ratio of 0.5

$$D_{1,it} = \begin{cases} \left(1 + \exp\left(-\hat{p}_{it} + \frac{T_{ci}+T_{ai}}{4}\right)\right)^{-1} & \text{if } \hat{p}_{it} <= \frac{T_{ci}+T_{ai}}{2} \\ \left(1 + \exp\left(\hat{p}_{it} - 3\frac{T_{ci}+T_{ai}}{4}\right)\right)^{-1} & \text{if } \hat{p}_{it} > \frac{T_{ci}+T_{ai}}{2} \end{cases} \tag{8}$$

$$D_{2,it} = \begin{cases} 1 & \text{if } \hat{p}_{it} < T_{ci} \\ \left(1 + \exp\left(p_{it} - \frac{T_{ai}+T_{ci}}{2}\right)\right)^{-1} & \text{if } T_{ci} \leq \hat{p}_{it} < T_{ai} \\ 0 & \text{if } \hat{p}_{it} \geq T_{ai} \end{cases} \tag{9}$$

As an illustration, Figure 1 displays the disposition for each action and their variation with respect to lake pollution signal. These functions are displayed for a lake threshold equal to 20, and a confidence ratio of 0.5, for illustrative purposes.

3.2.2 Solo disposition: the affective component

In the spirit of Agent Zero, individual disposition to adopt a given action varies with the emotional response associated with environmental signals. Emotional responses can arise by observing pollution-induced weather events or external signs of pollution, such as disappearing species or natural deterioration. Conversely, the emotional response can occur by observing a flourishing world. The probability of occurrence of pollution and good-state shocks at each period, q_{pt} and q_{gt} respectively, nonlinearly scales with the actual current lake pollution: this probability confirms that the environmental state of the Shallow Lake affects the surrounding environment. We further assume

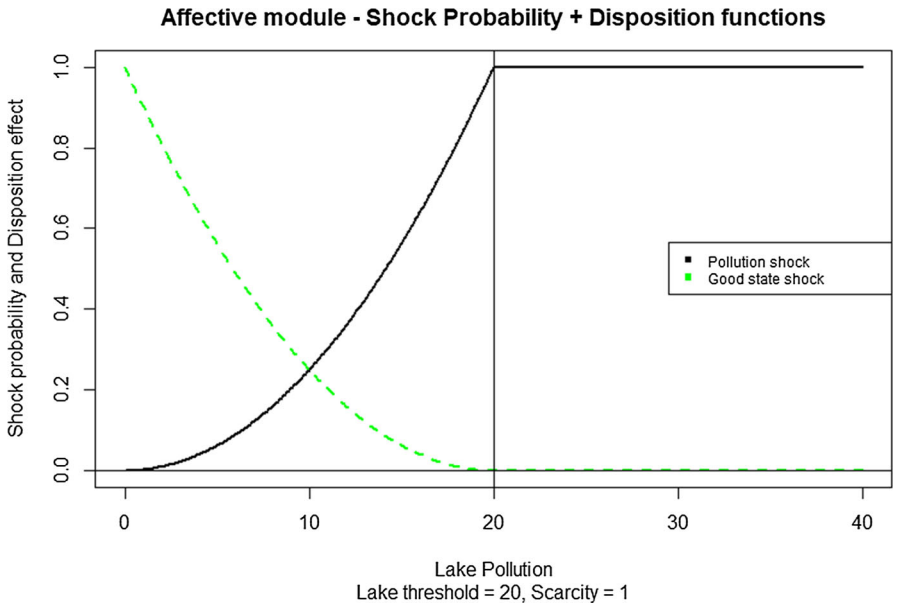

Fig. 2 Affective module shock probabilities and disposition functions

that shocks are rare events, multiplying the obtained probabilities by a scarcity factor $s = 0.2$.

$$q_{pt} = \begin{cases} s\left(\frac{p_t}{p^*}\right)^2 & \text{if } p_t \leq p^* \\ s & \text{if } p_t > p^* \end{cases} \tag{10}$$

$$q_{gt} = \begin{cases} s\left(1 - \frac{p_t}{p^*}\right)^2 & \text{if } p_t \leq p^* \\ 0 & \text{if } p_t > p^* \end{cases} \tag{11}$$

We illustrate the dynamics of the affective component (shock probability and disposition effect) in Figure 2 for a lake threshold of 20 and a scarcity factor of 1 as illustration. In the simulations of the next section, we set $s = 2$.

Shocks occur locally and according to a probability depending on the actual pollution of the lake. Agents are randomly located in a grid torus space. Each squared element j of the grid is subjected to the shock probabilities above. For an agent to observe a shock, that shock must then occur on the squared grid element the agent lies on. Those local shocks are local environmental collapse's early warnings signals, individual-specific, given the random locations of the agents at initialization. An agent observing an early warning for a natural disaster has a strong motive for abstaining from exploiting, and this motive strength varies with the intensity of the signal observed. In this sense, the signal intensity translates the idea of associative strength

introduced by the Rescorla–Wagner equation to the shallow lake. For simplicity, we assume that the signal strength, or intensity of the observed event, is equal to its probability of occurrence multiplied by the parameter s. For example, as pollution is close to the lake threshold, pollution shocks will occur more frequently, but their amplitude will also be higher. So will be the disposition effect generated by this event. If there is no observed signal, there will be no disposition effect. Disposition effects accumulate when two opposite signals pollution and good state are observed simultaneously. The binary variable G captures whether a good state shock occurred for an agent. The binary variable P captures whether a nasty state shock occurred for an agent.

Observing a pollution shock activates disposition to exploit sustainably and to abstain. Observing a good-state shock increases disposition to exploit fully and to exploit sustainably. Denoting $A_{0,it}$, $A_{1,it}$ and $A_{2,it}$ agent i's affective disposition to, respectively, abstaining, exploiting reasonably and exploiting fully, we obtain:

$$A_{0,it} = \begin{cases} 0 & \text{if } G = 0 \text{ and } P = 0 \\ 0 & \text{if } G = 1 \text{ and } P = 0 \\ q_{pt} & \text{if } G = 0 \text{ and } P = 1 \\ q_{pt} & \text{if } G = 1 \text{ and } P = 1 \end{cases} \tag{12}$$

$$A_{1,it} = \begin{cases} 0 & \text{if } G = 0 \text{ and } P = 0 \\ q_{gt} & \text{if } G = 1 \text{ and } P = 0 \\ q_{pt} & \text{if } G = 0 \text{ and } P = 1 \\ q_{gt} & \text{if } G = 1 \text{ and } P = 1 \end{cases} \tag{13}$$

$$A_{2,it} = \begin{cases} 0 & \text{if } G = 0 \text{ and } P = 0 \\ q_{gt} & \text{if } G = 1 \text{ and } P = 0 \\ 0 & \text{if } G = 0 \text{ and } P = 1 \\ q_{gt} & \text{if } G = 1 \text{ and } P = 1 \end{cases} \tag{14}$$

3.2.3 Social disposition to act in multilayer networks

The layouts of the agents' networks play a crucial role both as backbones for cooperation and as channels for information (Santos et al. 2006; Szabo and Gabor 2007). From the deliberative and affective components, player i's own disposition to choose either action at period t, respectively, denoted $\delta_{it,0}$, $\delta_{it,1}$ and $\delta_{it,2}$ is computed as the sum of their dispositions to act.

$$\begin{cases} \delta_{it,0} = D_{0,it} + A_{0,it} \\ \delta_{it,1} = D_{1,it} + A_{1,it} \\ \delta_{it,2} = D_{2,it} + A_{2,it} \end{cases} \tag{15}$$

The social component in the Agent Zero cognitive architecture adds to δ the disposition of other agents within the range of interaction. Concerning Agent Zero, we innovate in that we consider the case of multilayer networks when agents' emotional

and reasoning connections are not the same. We distinguish two different network structures for the propagation of deliberative and emotional dispositions. Transmission of emotional responses through emotional connections, and propagation of reasoning by intellectual connections, provide both essential and different means of the emergence of environmental awareness. Hence, assessing their respective role can be interesting in this study. The former is assumed to be organized around a star network structure, modeling a scientific community or governmental agencies acting as focal points for the agents. The latter, i.e., social networks, friendly or family connections, is here modeled as a random network à la Erdos Rényi with an average degree of β (Barabasi 2016).

Regarding the deliberative propagation network, we study different network structures organized concerning different centralization levels, i.e., the tendency for one or few agents to occupy central positions in the social interconnections. We intend to study how network structure, coupled with the specifications of the cognitive architecture of the agents, influences the emergence of environmental awareness and adoption of either action. A parameter of centralization, denoted α, gives the proportion of agents connected to the central agent. Network density δ is the ratio of realized links to the number of possible links in the population.

Centralized network generation The centralized network model used is derived from the Alpha-centralization model developed by Vié and Morales (2020), which generates graphs with different levels of centralization. It constitutes a generalization of the preferential attachment mechanism (Barabasi and Albert 1999) with an exponent controlling for the emergence and importance of hubs. This specification can generate various webs, from no centralization (randomly and independently distributed edges) to perfectly centralized networks (in one or a few hubs). In between the two extreme cases, an extensive range of scale-free networks is obtained, in which different situations of importance and number of hubs are delivered. The network generation process consists in allocating edges between agents (nodes) as a function of the attachment probability. The probability of node i to create an edge with node j is as follows:

$$p_{ij} \propto k_j^{\alpha} \tag{16}$$

where k_j is the number of connections of node j, and α is the exponent we use to control the influence of the preferential attachment mechanism. If $\alpha = 0$, the attachment probability becomes equal among all nodes, and we obtain a random network with no central hub similar to the Erdos–Ranyi model. If $\alpha = 1$, we obtain the standard Barabási-Albert network with a few hubs. If $\alpha = 2$, we create a network with complete centralization where all nodes are linked to a single central one. This preferential attachment mechanism extension magnifies the degree of heterogeneity among nodes for $\alpha > 1$ and reduces such attractive force for any $\alpha < 1$.

Density network generation To generate a social network with density δ , we implement an Erdos–Renyi random network where each link between two nodes has a probability δ of existence.

Social disposition Having introduced the multilayered social information network, now we account for imitation, communication and interactions between agents: the disposition of or the information coming from other agents is channelled via the emotions disposition network and the deliberative network, respectively. Agents do infer the dispositions of their neighbors in the social network. In particular, they compute both the average deliberative disposition over the N_d agents in their deliberative disposition networks and the average emotional disposition over the N_e agents in their emotional disposition network. The total disposition of each agent i, in accordance with the previous definitions, $\Delta_{0,it}$,$\Delta_{1,it}$ and $\Delta_{2,it}$, are given by the following equations.

$$\begin{cases} \Delta_{0,it} = \delta_{it,0} + \frac{1}{N_d} \sum_k D_{0,kt} + \frac{1}{N_e} \sum_h A_{0,it} \\ \Delta_{1,it} = \delta_{it,1} + \frac{1}{N_d} \sum_k D_{1,kt} + \frac{1}{N_e} \sum_h A_{1,it} \\ \Delta_{2,it} = \delta_{it,2} + \frac{1}{N_d} \sum_k D_{2,kt} + \frac{1}{N_e} \sum_h A_{2,it} \end{cases} \tag{17}$$

Now we are finally in the position to set up all the pieces of our model. The adopted action is the one whose total disposition is the highest among the three alternatives. In the case of equality in dispositions, we assume that the agents conserve the same action as in the last period. The action variable X with possible entries $\{0, 1, 2\}$, we note, with t the period considered and i the agent reads:

$$X_t = \begin{cases} 0 \text{ if } \max(\Delta_{0,it}, \Delta_{1,it}, \Delta_{2,it}) = \Delta_{0,it} \\ 1 \text{ if } \max(\Delta_{0,it}, \Delta_{1,it}, \Delta_{2,it}) = \Delta_{1,it} \\ 2 \text{ if } \max(\Delta_{0,it}, \Delta_{1,it}, \Delta_{2,it}) = \Delta_{2,it} \end{cases} \tag{18}$$

4 Results

4.1 Simulation methods and exploration

Thanks to the richness offered by the Agent Zero framework, we can study the model outcomes with tunable cognition mechanisms. We can study separately the model dynamics where agents are endowed with one particular cognitive module (deliberative or affective population) or with a combination of these modules (deliberative social agents, affective and deliberative agents, affective deliberative and social, any combination). Our results hence present the multifaceted relationship between many of the model parameters and the degree of cooperation's enhancing awareness or environmental concerns in the population, defined as the fraction of agents who decide to abstain from polluting at a given time t and referred to as "degree of environmental awareness" or simply "awareness" hereafter. We can evaluate this relationship by plotting the most relevant parameters against the degree of awareness. The results cover all possible assumptions, or a combination of assumptions, on the population cognition. For example, we can study the emergence of awareness in a fully affective-reasoning

population, or one endowed with deliberative thinking and social cognition, and so on with different combinations. Although empirical identification of the significant parameters involved in this social cognition model is a daunting task, the comparative approach developed in this paper allows us to form general results robust across the cognition assumptions used.

Dealing with several parameters (12), the exploration of the model behavior can be pretty cumbersome. We use an intuitive and efficient way of restoring statistical inference and model understanding in environments with many parameters of interest. The Openmole platform (Reuillon et al. 2013) embeds Sobol Sampling algorithms. Sobol Sampling, introduced by Sobol (1967), generates a sequence of points uniformly distributed in multidimensional space to fully cover the n-dimensional parameter space, where n is the number of parameters. Sobol sequences allow us to obtain a representative sample of observations to implement our statistical analysis. The dimensionality of the exploration task is reduced, with quasi null loss in generality (Fig. 3). We present detailed result tables obtained with both linear and polynomial regressions in the supplementary material.

To implement those samplings, we specified some range of values for the parameters. We adopted a rather large choice of intervals to explore the variety of trajectories of the model. We simulated the model on a 16×16 lattice grid; hence, the maximum value for our network density parameter, which indicates the range up to which affective connections form, is set to 16. In order to deal with the stochasticity of model runs, especially regarding the random physical locations of the agents, we present as results the average of 10 runs for each parameter configuration. Each Sobol sampling implemented cover 1000 different points in the parameter configuration space.

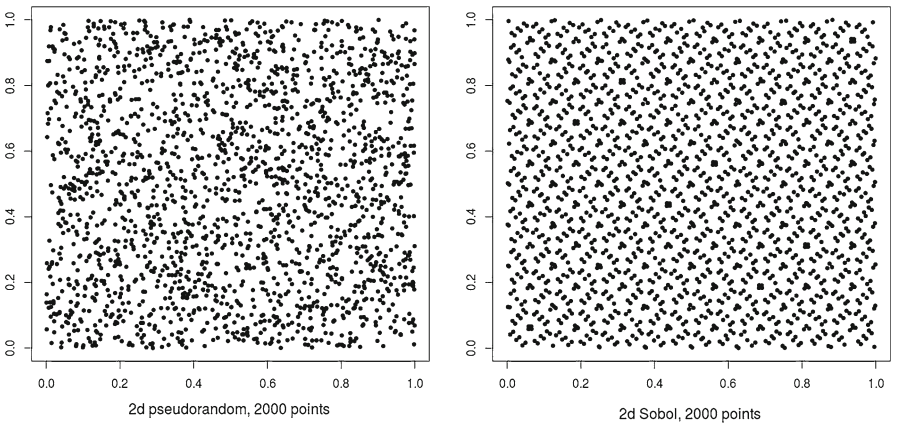

Fig. 3 A visualization of Sobol sequences in a 2-dimensional space (Smith et al. 2017)

4.2 Main results: degree of environmental awareness, internalization of constraints and self-organized criticality

In the following tables, we present the impact of the most relevant parameters on three model outcomes, namely the degree of awareness, the pollution level reached in the lake and the frequency of eutrophications events, that is, the frequency of polluted lake states during the considered period.

In Table 2, we starts with the influence of the confidence ratio c on A, P and E. In all cognition variants, the confidence ratio significantly increases cooperation. It enhances environmental awareness: as the distance between the confidence and the alarm thresholds decreases, the proportion of agents who sustainably exploit the lake increases—the number of environmentally aware agents increases. Since in the implementation of the sustainable action, the deliberative module plays a central role, when the inferred pollution lies between the two decision thresholds, we would first expect that increasing the distance between both, i.e., decreasing the confidence ratio, would increase awareness in the population. Surprisingly, simulations results obtained

Table 1 Parameter ranges used in Sobol samplings

Initial pollution, p_0	[0, 100]
Natural decay, δ	[0, 100]
Collapse recovery, r	[0, 100]
Population, N,	[3, 100]
Lake threshold, p^*	[0, 100]
Signal variance, σ_p	[0, 100]
Threshold variance, σ_t	[0, 100]
Signal scarcity, s	[0, 10]
Confidence ratio, c	[0, 1]
Network centralization, α	[0, 1]
Network range (density), δ	[0, 16]
Correlation parameter, ρ	[0, 1]

Table 2 The impact of the confidence ratio c on awareness and pollution outcomes

	Awareness (%)	Pollution level	Eutrophication frequency (%)
Cognitive modules			
Deliberative	20.675***	34.732***	42.442***
	(1.887)	(2.180)	(2.601)
Delibrative + Social	14.161***	29.338***	39.824***
	(1.926)	(2.212)	(2.548)
Deliberative + Affective	18.058***	30.046***	38.935***
	(1.765)	(2.136)	(2.523)
Deliberative + Affective + Social	18.027***	30.096***	38.905***
	(1.765)	(2.141)	(2.517)

$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$

Table 3 The impact of natural decay δ on awareness and pollution outcomes

	Awareness (%)	Pollution level	Eutrophication frequency (%)
Cognitive modules			
Deliberative	0.345***	0.157***	0.391***
	(0.019)	(0.022)	(0.026)
Affective	0.276***	0.669***	0.360***
	(0.011)	(0.083)	(0.017)
Deliberative + Affective	0.369***	0.118***	0.420***
	(0.018)	(0.021)	(0.025)
Deliberative + Social	0.369***	0.131***	0.381***
	(0.019)	(0.022)	(0.025)
Affective + Social	0.377***	0.443***	0.460***
	(0.011)	(0.048)	(0.017)
Deliberative + Affective + Social	0.371***	0.122***	0.425***
	(0.018)	(0.021)	(0.025)

$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$

by Sobol sampling and processed through regression analysis show the inverse: the closer the two thresholds are, the higher is the awareness in the population.

The intuition behind this result is insightful. When the two bounds are far from each other, the early periods see high initial awareness. The lake's pollution remains low, forcing the system into equilibrium with an exiguous level of awareness, simply because the environment no longer triggers awareness. Polynomial regression shows that this combined positive effect is composed of a first-order intuitive negative effect and a more robust, second-order positive effect in a nonlinear way. On the contrary, when the bounds are close to each other, a high proportion of agents exploit the lake with high-pollution externalities in the early stage, pushing the system into self-organized criticality. In this situation, the emergence of awareness represents the adaptation of agents to the critical environmental state they generated.

It is interesting to note that across all cognitive variants, increasing the confidence ratio, i.e., generating closer and closer confidence and alarm thresholds, has a significant and positive impact on both awareness and pollution, enforcing our explanation of this counterintuitive mechanism. The marginal impact of the confidence ratio over the frequency of the polluted lake state is also high, significant and positive. In this model, having more careless agents increases the spread of the environmental concern in the population, as more careless agents push the system on the verge of collapse. The path to environmental awareness appears more hazardous, as awareness tends to emerge once the system state is critical.

Following the same intuition, in Table 3 we show that the natural decay parameter, namely the quantity of pollution absorbed by the lake, systematically has a positive and significant effect on population awareness, lake pollution and frequency of eutrophication events. However, we would expect a higher natural reduction of pollution to have a positive environmental impact. The explanation for this phenomenon is that agents

Table 4 The impact of collapse recovery r on awareness and pollution outcomes

	Awareness (%)	Pollution level	Eutrophication frequency (%)
Cognitive modules			
Deliberative	− 36.912***	− 17.910***	− 46.573***
	(1.887)	(2.180)	(2.601)
Affective	− 24.801***	− 78.174***	− 37.101***
	(1.064)	(8.310)	(1.736)
Deliberative + Affective	− 36.654***	− 20.588***	− 44.825***
	(1.765)	(2.136)	(2.523)
Deliberative + Social	− 35.292***	− 10.561***	− 42.274***
	(1.926)	(2.213)	(2.549)
Affective + Social	− 24.271***	− 32.177***	− 27.922***
	(1.142)	(4.833)	(1.678)
Deliberative + Affective + Social	− 36.481***	− 18.036***	− 44.775***
	(1.765)	(2.142)	(2.518)

* $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

in the model can internalize the environmental constraints. Similarly to what we could expect in the real world, higher environmental resilience does not discourage people from polluting more: it allows people to exploit and pollute a bit more, knowing that the lake will clean itself. This model shows that more resilient environments are not necessarily better off in this simplified setting, sustainably speaking. This reasoning applies to the recovery abilities of the system under a polluted state, but not to their resistance as characterized by the lake threshold, which exerts a positive linear effect on the lake.

Table 4 deals with the collapse recovery parameter r. This parameter has a strong negative significant marginal impact over awareness (%) and the frequency of eutrophication events, but a positive impact over pollution. The presence of a more robust environment in the situation of polluted state undermines the collective evolution of the population toward a distribution favoring careful, more aware agents. Environments more resistant may shelter a less environmentally concerned population, who pollute more.

Interestingly, the latter draws a contrast with the previous result on the natural decay. In our model, the collapse recovery of the environment is not taken into account by the agents, while the natural decay value is. Environments with higher pollution absorption abilities appear less frequently in the polluted state, but only if the agents do not know or anticipate this feature. The consequences of such a setting are a lake less frequently polluted and paradoxically a population less concerned about the environment. Empirically speaking, public news happy to announce higher than expected absorption of CO2 by the oceans, or the discovery of a new powerful plastic-eating bacteria, may provoke weaker self-control and more pollution externalities from the agents, since people may feel confident that the environment will absorb it.

Table 5 The impact of pollution signal variance σ_p on awareness and pollution outcomes

	Awareness (%)	Pollution level	Eutrophication frequency (%)
Cognitive modules			
Deliberative	0.067***	−0.006	0.057**
	(0.019)	(0.022)	(0.026)
Deliberative + Affective	0.060***	−0.007	0.050**
	(0.018)	(0.021)	(0.025)
Deliberative + Social	0.106***	0.018	0.099***
	(0.019)	(0.022)	(0.025)
Deliberative + Affective + Social	0.061***	−0.005	0.053**
	(0.018)	(0.021)	(0.025)

*$p < 0.1$; **$p < 0.05$; ***$p < 0.01$

Table 6 The impact of threshold variance σ_t on awareness and pollution outcomes

	Awareness (%)	Pollution level	Eutrophication frequency (%)
Cognitive modules			
Deliberative	0.033*	−0.009	−0.007
	(0.019)	(0.022)	(0.026)
Deliberative + Affective	0.026	−0.016	−0.006
	(0.018)	(0.021)	(0.025)
Deliberative + Social	0.024	−0.013	−0.028
	(0.019)	(0.022)	(0.025)
Deliberative + Affective + Social	0.025	−0.025	−0.009
	(0.018)	(0.021)	(0.025)

*$p < 0.1$; **$p < 0.05$; ***$p < 0.01$

Tables 5 and 6 show the role of the Signal Variance σ_p and the Threshold Variance σ_t. As for the information processing, the model results outline that the variance of the pollution signals have a positive impact on awareness and the frequency of the polluted lake state, as they push the system into self-organized criticality from erroneous evaluation of the lake state (and less good decisions for its actual state). However, the impact of the variance of the thresholds in the population never achieves a satisfying and robust significance level. In this model of environmental awareness emergence, differences in information about the environment state seem to matter far more than differences in information processing in the population. In other words, the difference in opinions or concerns about the environments may not be as much an obstacle to its protection and the emergence of collective awareness as noisy environmental information could be.

Tables 7 and 8 study the effects of Network Centralization α and Network Density ? on the outcomes of the model. In the context of our multilayer network analysis, network centralization significantly reduces the proportion of agents sustainably exploiting the lake in the deliberative social cognitive setting but does not seem to

Table 7 The impact of network centralization α on awareness and pollution outcomes

	Awareness (%)	Pollution level	Eutrophication frequency (%)
Cognitive modules			
Deliberative + Social	− 4.296**	3.953*	− 4.378*
	(1.925)	(2.212)	(2.548)
Affective + Social	1.269	− 0.999	1.020
	(1.142)	(4.832)	(1.678)
Deliberative + Affective + Social	− 0.464	− 1.860	− 1.081
	(1.765)	(2.141)	(2.517)

* $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Table 8 The impact of network density δ on awareness and pollution outcomes

	Awareness (%)	Pollution level	Eutrophication frequency (%)
Cognitive modules			
Deliberative + Social	0.054	− 0.070	0.104
	(0.120)	(0.138)	(0.159)
Affective + Social	0.270***	− 0.366	0.560***
	(0.071)	(0.302)	(0.105)
Deliberative + Affective + Social	0.063	0.006	0.155
	(0.110)	(0.134)	(0.157)

* $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

have a significant impact in other settings. Since this centralization network transmits deliberative dispositions, it is interesting to note that this negative effect disappears in the combined model with affective, deliberative and social cognitive modules. The affective module tends to counterbalance the negative effect of network centralization. Agents may be too heavily dependent on a single entity, favoring the homophily of behaviors in the population. The density of the network conveying peer-to-peer emotional dispositions positively impacts awareness percentage in the affective and social setting. This mechanism increases the frequency of the polluted lake state in time. Fear-responses generated by local individual-specific shocks spread in the population, like fear acquisition without personal exposure to danger in Epstein (2014).

So far in this discussion, we have devoted our focus to a multilayer setting, in which we assume that the network layers are independent. In other words, the analysis above implicitly assumes that individual social circles are separated: circles that transmit deliberative disposition differ from social circles that convey affective or emotional responses. We here embrace the possibility that the two layers may be correlated. By setting a correlation parameter $\rho \in [0, 1]$, we denote the probability for any individual link of one layer to exist in the other layer. In other words, we are interested to know how our results on awareness and pollution change when we consider various levels of correlation of network layers. That is the extent to which deliberative and affective

Table 9 The impact of correlation between network layers on awareness and pollution outcomes (full model, Affective + Deliberative + Social modules)

	Awareness (%)	Pollution level	Eutrophication Frequency (%)
Independent variables			
Correlation	−6.955***	−1.166	−4.329*
	(1.909)	(2.146)	(2.369)
Correlation × Network centralization	2.184	2.510	4.949
	(4.973)	(5.604)	(6.185)
Correlation × Network density	−1.005***	−0.377	−0.767**
	(0.310)	(0.350)	(0.386)
Network centralization	−3.544*	1.054	−2.683
	(1.908)	(2.145)	(2.368)
Network density	−0.483***	−0.412***	−0.451***
	(0.119)	(0.134)	(0.148)

*$p < 0.1$; **$p < 0.05$; ***$p < 0.01$

connections can be confounded. We may indeed imagine that individuals do share with their link neighbors both types of dispositions.

The first part of table 9 regresses model outcomes on the same predictors as before, adding the Correlation parameter. The second part regresses on product variables Correlation × Network centralization and Correlation × Network density for additional information. We first observe that, in general terms, the correlation between network layers tends to reduce awareness in the population and decrease the frequency of the pollution state. We interpret it as an indicator of the efficiency of social learning in the population. As the correlation of layers creates a more connected population, messages spread faster, allowing a better collective self-organized regulation, causing paradoxically, despite the faster connection, a lower level of awareness, as the danger level would be lower. When we look at product effects of correlation with network parameters, we observe that most of this variation from correlation seems to stem from the combination with network density. In contrast with previous results, network density seems to have a sizeable, desirable effect on pollution through the efficiency of emotional responses. A higher correlation contributes to enforcing this channel, adding deliberative links to the contagion of emotional dispositions. The quasi null effect of network centralization on awareness and pollution may be explained by the failure of hierarchic systems to favor social learning. In contrast, collective decentralized intelligence, illustrated by network density, and emphasized by correlation, seems to provide a way out of the polluted state.

5 Discussion

The main insights of these results deal with the mechanisms underlying the emergence of cooperation via environmental awareness or consciousness of the proximity of environmental collapse. In our model, across all cognitive modules combinations, self-

organized criticality favors the emergence of environmental awareness in collective-risk social dilemmas. Both intuitively and paradoxically, the agents become more and more conscious of the environmental risk as the lake's pollution becomes closer to the catastrophe threshold. In other words, more cooperative behaviors emerge on the brink of collapse, in line with the previous empirical literature. The risk of collective failure provides an escape to the tragedy of the commons.

Our results on cooperation emerging from awareness and the relaxation of the moral standards as people feel that tragedy is far or less likely are consistent with the existing literature. Indeed, in Milinski et al. (2008), the authors have empirically demonstrated that people tend to cooperate more when the risk of a common failure in preventing an ecological disaster increases. Moreover, the number of cooperators drastically fall if the likelihood of the disaster decreases. A natural starting point for future research would be to calibrate our agent-based model based on the result of this paper.

In this paper, we decided to simplify the dynamics of the lake, choosing to model its hysteresis with a binary variable ruled by an activation threshold and a recovery time. Nevertheless, we are eager to enrich the spatio-temporal dynamics of the lake with more realistic modeling of its unfolding (see Scheffer 1998 and Martin and Schlüter 2015). Other directions for further research include the possibility of the agent punishing non-cooperative behavior, as in the literature of altruistic punishment (see Fehr and Gachter 2002). Finally, we did not explicitly reference economic motivations as drivers of the agents' actions: these motivations remained in the background because our focus was on the emergence of cooperation at the edge of a possible environmental collapse, where economic considerations—we deem—could be put aside. We acknowledge that relaxing economic considerations when facing a sudden case of *force majeure* is valid as a first-order approximation, but that those economic stakes would appear again after some time. In this respect, it would be interesting to apply the agent zero frameworks to the path toward the collapse in its early stages when the threat is remote, but the seeds of the crisis are already there, obfuscated by economic, business, as usual, agents' preferences.

Recent applications of the Agent Zero framework by Vié (2019) emphasized the existence of the emergence of particular phenomena, in the case of opinion dynamics and information selection, that result from the combination of different modules. This work outlined that parameter effects vary in significance, sign and magnitude depending on the cognitive assumptions made on the population, showing the richness of the neurocognitive framework of Epstein (2014) and its ability to generate complex phenomena from few simple individual rules (Epstein 2006).

6 Conclusion

This article explores an application of the Agent Zero neurocognitive decision-making model (Epstein 2014) to the emergence of environmental alertness and cooperation in a population of artificial agents. Affective, deliberative and social channels are modeled and take part in the acquisition and processing ecological signals. The emergence of cooperation enhancing's environmental awareness in the population modeled by an agent-based model where units endowed with the Agent Zero mind architecture choose

to exploit, exploit sustainably or abstain from exploiting natural resources, generating consequent pollution externalities. We examine the impact of network structure on the emergence of cooperation and the resulting lake state. Indeed network structures are at the heart of dispositions channelling both in the affective and in the deliberative module.

In order to avoid environmental collapse, ecological awareness must be widespread in the population. This attention emerges primarily when the system is at the edge of catastrophe. Natural processing of pollution, impacting the natural rate of absorption, or recovery abilities in the polluted equilibrium, does not encourage environmental awareness in the population, as the agents internalize the relaxed constraints and feel the possibility of polluting without consequences.

This framework also provides insights on the determinants of the transition to environmental awareness and avoidance of the polluted lake equilibrium by highlighting the positive role of precise information on the environmental state. While the variance of the information on the world's ecological state pushes the system into criticality through erroneous evaluation of the lake state by the agents, it also favors the spread of sustainable behaviors. Variance in the way agents process information, i.e., how heterogeneous agents react to environmental pollution signals, is nevertheless not an obstacle to this emergence.

Results also show the robustness of emergent alertness to decision threshold variance and social network structure. While network centralization reduces the proportion of alert agents in some cognitive settings, its impact appears negligible in the combined (deliberative, affective and social) model, showing the possibility of reaching a clean lake equilibrium across a wide variety of social networks topologies. Network density appears to contribute to the spreading of affective dispositions between populations observing and populations not observing ecological shocks, allowing the whole system to become more environmentally conscious.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11403-022-00347-8.

Acknowledgements The authors thank Prof. Katheline Schubert of the Paris School of Economics Environment & Regulation research group, the Young Researchers in Complex Systems Society, and Sciences Po Saint-Germain-en-Laye (Chaire Citoyenneté) for their support. The authors would like to thank the participants of the 2019 Workshop on Economics of Heterogeneous Interacting Agents (WEHIA) for their valuable remarks and discussions. The authors are grateful to the OpenMole team of the Paris Complex Systems Institute (ISCPIF, CNRS) and especially Romain Reuillon for their help in the technical aspects of the model analysis.

References

Axelrod R (1984) The evolution of cooperation. Basic Books. ISBN: 0-465-02122-0

Axelrod R (1997) The complexity of cooperation: agent-based models of competition and collaboration. Complexity 3(3):46–48

Barabasi A (2016) Network science. Cambridge University Press, Cambridge

Barabasi AL, Albert R (1999) Emergence of scaling in random networks. Science 286(5439):509–512

Barrett S, Dannenberg A (2012) Climate negotiations under scientific uncertainty. PNAS 109(43):17372–17376

Bonabeau E (2002) Agent-based modeling: methods and techniques for simulating human systems. Proc Natl Acad Sci USA 99(Suppl 3):7280–7287

Bowles S et al (2004) The evolution of strong reciprocity: cooperation in heterogeneous populations. Theor Popul Biol 65(1):17–28. https://doi.org/10.1016/j.tpb.2003.07.001

Cederman L (1997) Emergent actors in world politics: how states and nations develop and dissolve. Princeton University Press, Princeton

Dannenberg A et al (2015) On the provision of public goods with probabilistic and ambiguous thresholds. Environ Resour Econ 61:365–383. https://doi.org/10.1007/s10640-014-9796-6

Deffuant G, Neau D, Amblard F, Weisbuch G (2000) Mixing beliefs among interacting agents. Adv Complex Syst 3(01n04):87–98

Du et al (2012) Evolution of global cooperation driven by risks. Phys Rev E 85:056117

Du et al (2014) Climate collective risk dilemma with feedback of real-time temperatures. EPL 107:60005

Epstein JM (2014) Agent zero: toward neurocognitive generative social science

Epstein JM (2006) Generative social science: studies in agent-based computational modeling. Princeton University Press, Princeton

Epstein JM, Axtell RL (1996) Growing artificial societies: social science from the bottom up. MIT Press, Cambridge

Eriksen S, Inderberg TH, O'Brien K, Sygna L (2014) Introduction: development as usual is not enough. In: Climate change adaptation and development. Routledge, pp 17–34

Fehr E, Gachter S (2000) Cooperation and punishment in public goods experiments. Am Econ Rev 90:980–994

Fehr E, Gachter S (2002) Altruistic punishment in humans. Nature 415:137–40

Gächter S, Herrmann B (2009) Reciprocity, culture and human cooperation: previous insights and a new cross-cultural experiment. Philos Trans R Soc B Biol Sci 364:791–806

Hagel K et al (2016) Which risk scenarios can drive the emergence of costly cooperation? Sci Rep 6:19269. https://doi.org/10.1038/srep19269

Hagel K, Milinski M, Marotzke J (2017) The level of climate change mitigation depends on how humans assess the risk arising from missing the 2°C target. Palgrave Commun 3:17027. https://doi.org/10.1057/palcomms.2017.27

Harari Y-N (2014) Homo sapiens. A brief history of humankind. Vintage, London

Hardin G (1968) The tragedy of the commons. Science 162:1243–1248

Helbing D, Balietti S (2011) How to do agent-based simulations in the future: from modeling social mechanisms to emergent phenomena and interactive systems design. Santa Fe Institute Working Paper No. 2011-06-024. Santa Fe Institute

Henrich J et al (2001) In search of Homo economicus: behavioral experiments in 15 small-scale societies. Am Econ Rev 91:73–78

Holland J, Miller J (1991) Artificial adaptive agents in economic theory. Am Econ Rev 81:365–370

Jacquet J et al (2011) Shame and honour drive cooperation. Biol Lett 7:899–901

King AD, Harrington LJ (2018) The inequality of climate change from 1.5 to 2 C of global warming. Geophys Res Lett 45(10):5030–5033

Kumar M, Dutt V (2019) Collective risk social dilemma: role of information availability in achieving cooperation against climate change. JDDM 5. Article 2. 1

Lazer D, Friedman A (2007) The network structure of exploration and exploitation. Adm Sci Q 52(4):667–694

Mäler KG et al (2003) The economics of shallow lakes. Environ Resour Econ 26(4):603–624

Martin R, Schlüter M (2015) Combining system dynamics and agent-based modeling to analyze social-ecological interactions—an example from modeling restoration of a shallow lake. Front Environ Sci 3:66

Milinski M (2016) Reputation, a universal currency for human social interactions. Philos Trans R Soc 371:20150100
Milinski M et al (2002) Reputation helps solve the 'tragedy of the commons'. Nature 415:424–426
Milinski M et al (2008) The collective-risk social dilemma and the prevention of simulated dangerous climate change. PNAS USA 105(7):2291–2294
Nowak M (2006) Five rules for the evolution of cooperation. Science 314(5805):1560–1563. https://doi.org/10.1126/science.1133755
Nowak M (2011) Supercooperators: the mathematics of evolution, altruism and human behaviour. Canongate Books, New York
Nowak MA, Sigmund K (1998) Evolution of indirect reciprocity by image scoring. Nature 393:573–577
Nowak MA, Sigmund K (2005) Evolution of indirect reciprocity. Nature 437:1291–1298
Ostrom E (1990) Governing the commons: the evolution of institutions for collective action
Ostrom E et al (1992) Covenants with and without a sword: self-governance is possible. Am Polit Sci Rev 86:404–417
Raihani N, David A (2011) Uncertainty, rationality and cooperation in the context of climate change. Clim Change 108:47–55
Rand DG, Nowak MA (2013) Human cooperation. Trends Cogn Sci 17(8):413–425
Rand DG et al (2009) Positive interactions promote public cooperation. Science 325:1272–1275
Reuillon R et al (2013) OpenMOLE, a workflow engine specifically tailored for the distributed exploration of simulation models. Future Gener Comput Syst 29:1981–1990
Righi S, Takács K (2018) Social closure and the evolution of cooperation via indirect reciprocity. Sci Rep 8:11149
Rockenbach B, Milinski M (2006) The efficient interaction of indirect reciprocity and costly punishment. Nature 444:718–723
Santos FC, Pacheco JM (2011) Risk of collective failure provides an escape from the tragedy of the commons. PNAS USA 108:10421–10425
Santos FC, Rodrigues JF, Pacheco JM (2006) Graph topology plays a determinant role in the evolution of cooperation. Proc Biol Sci 273:51–55
Santos et al (2012) Evolutionary dynamics of climate change under collective-risk dilemmas. Math Models Methods Appl Sci 22(Suppl.):1140004. https://doi.org/10.1142/S0218202511400045
Santos FP, Santos FC, Pacheco J (2018) Social norm complexity and past reputations in the evolution of cooperation. Nature 555:242–245
Scheffer M (1989) Alternative stable states in eutrophic, shallow freshwater systems: a minimal model. Hydrobiol Bull 23:73–83
Scheffer M (1998) Ecology of shallow lakes. Chapman and Hall, London
Scheffer M et al (2001) Catastrophic shifts in ecosystems. Nature 413:591–596
Smith A, Lovelace R, Birkin M (2017) Population synthesis with quasirandom integer sampling. J Artif Soc Soc Simul 20(4):1–14
Sobol IM (1967) Distribution of points in a cube and approximate evaluation of integrals. USSR Comput Math Math Phys 7:86–112
Szabo G, Gabor F (2007) Evolutionary games on graphs. Phys Rep 446(4–6):97–216
Tavoni A et al (2011) Inequality, communication and the avoidance of disastrous climate change in a public goods game. PNAS USA 108:11825–11829
Turchin P (2016) Ultrasociety: how 10.000 years of war made humans the greatest cooperators on earth. Beresta Books, Chaplin
Vié A (2019) Information selection efficiency in networks: a neurocognitive-founded agent-based model. In: Network theory and agent-based modelling in economics and finance. Springer, pp 11–34
Vié A, Morales AJ (2020) How connected is too connected? Impact of network topology on systemic risk and collapse of complex economic systems. In: Computational economics, pp 1–25
Wang J, Feng F, Te W, Wang L (2009) Emergence of social cooperation in threshold public goods games with collective risk. Phys Rev E 80:016101
West SA, Griffin AS, Gardner A (2007) Social semantics: altruism, cooperation, mutualism, strong reciprocity and group selection. J Evol Biol 20(2):415–432
Wilson EO (2012) The social conquest of earth. Liveright, New York