# Model details

## 1 Model overview

The model structure is shown in Figure 1. Individuals have a set of environmental behavioural traits that spread through a fixed Watts–Strogatz graph via social interactions with their neighbours. These exchanges are mediated by transmission biases informing from whom an individual learns and how much attention is paid. The influence of individuals on each other is a function of their similarity in environmental identity, where we represent environmental identity computationally by aggregating past agent attitudes towards multiple environmentally related behaviours. To perform a behaviour, agents must both have a sufficiently positive attitude toward the behaviour and overcome a corresponding threshold. This threshold structure, where the desire to perform a behaviour does not equal its enactment, allows for a lack of coherence between attitudes and actual emissions. This leads to a disconnect between what people believe and what they do, such that the social network as whole desires greener behaviours but only a minority performs them.

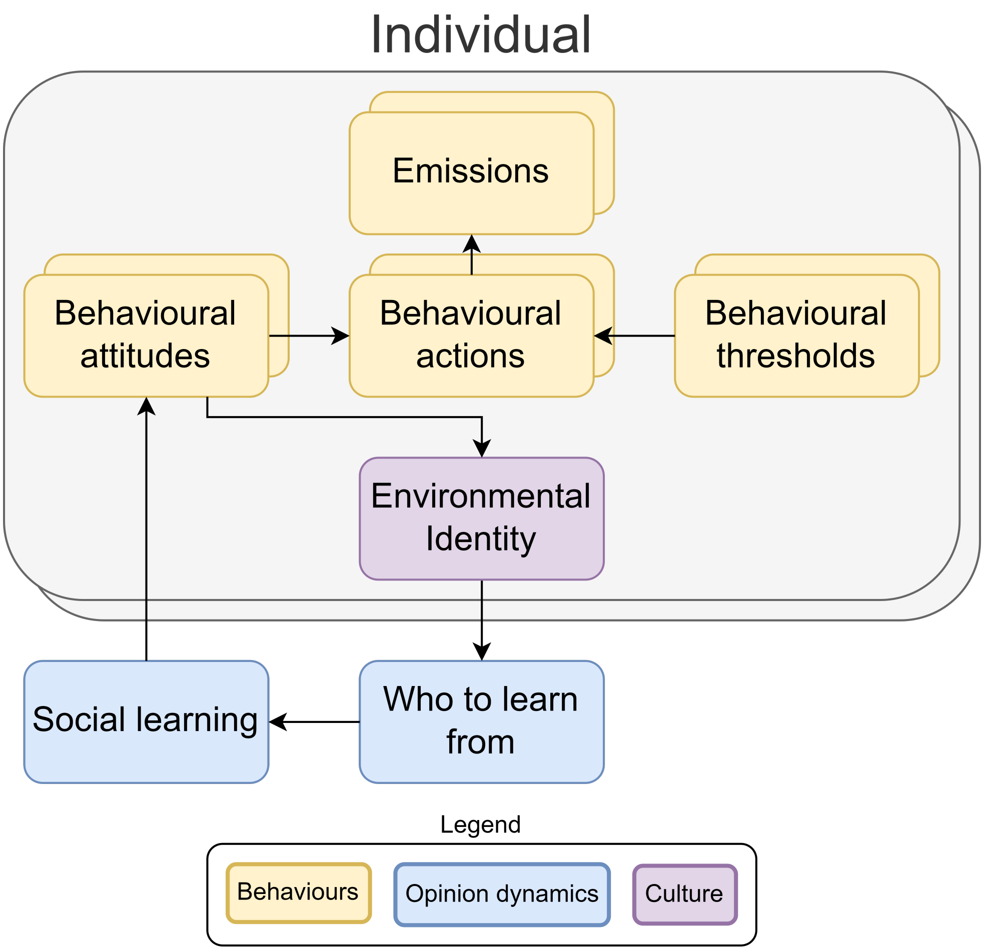


Figure 1: Model structure. Arrows indicate the direction of influence between components and stacked boxes represent multiple individuals (grey) and a behavioural vector (yellow). The past attitude of an individual towards multiple behaviours determines a single environmental identity. This in turn influences who an individual pays attention to when learning socially.

# 2 Details of model procedure

The model initialisation is shown in Figure 2 and the procedure for a single time step in Figure 3. Below the equations that compose the model are summarised.

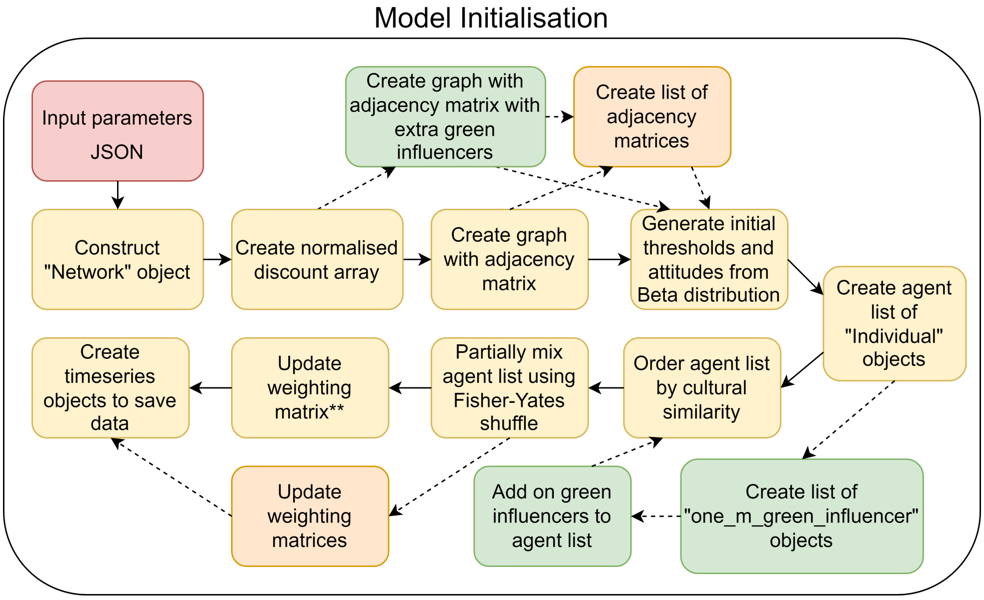


Figure 2: Model initialisation procedure, solid arrows indicate the default run of behavioural interdependence and no green influencers (yellow), whilst dashed represent an alternate scenario such as behavioural independence (Orange) or addition of green influencers (Green). Note that is possible to have both behavioural independence and green influencers. In the case of static uniform neighbour weighting (**), this step is skipped.

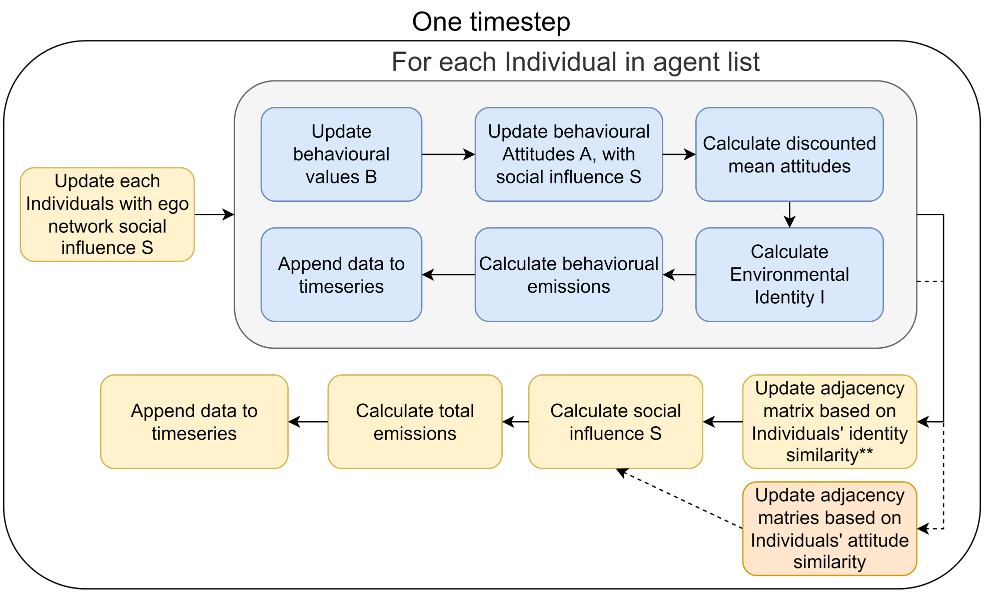


Figure 3: One time step, blue components indicated those occurring within the Individual object. In the case of static culturally determined neighbour weighting (**), this step is skipped.

## 2.1 Behavioural values, attitudes and thresholds

The behavioural value $B_{t,n,m}$ is determined by two continuous variables: the individual's attitude $A_{t,n,m} \in [0,1]$ towards the behaviour and the threshold or barrier of entry for performing a behaviour $T_{n,m} \in [0,1]$. A value of $A_{t,n,m} = 1$ is the "greenest" attitude and $A_{t,n,m} = 0$ the "brownest" or most indifferent to environmental impacts, and similarly $T_{n,m} = 1$ is the highest barrier of entry and $T_{n,m} = 0$ the lowest. Therefore the behavioural value $B_{t,n,m}$ is given by

$$B_{t,n,m} = A_{t,n,m} - T_{n,m}, \tag{1}$$

where initial values $T_{n,m}$ and $A_{0,n,m}$ are generated separately using a Beta distribution

## 2.2 Behavioural emissions

The total emissions $E_t$ produced by the population of size $N$ is given by the summation over each of the multiple behaviours performed by each individual,

$$E_t = \sum_{n=1}^{N} \sum_{m=1}^{M} \frac{1 - B_{t,n,m}}{2}, \tag{2}$$

## 2.3 Environmental identity

We define the environmental identity $I_n$ of agent $n$ at time $t$ as

$$I_{t,n} = \left[ \frac{1}{\sum\limits_{s=0}^{\rho} \delta^s} \right] \sum_{s=0}^{\rho} \delta^s \bar{A}_{t-s,n}, \tag{3}$$

where the mean behavioural attitude $\bar{A}$ over $M$ behaviours is given by

$$\bar{A}_{t,n} = \frac{1}{M} \sum_{m=1}^{M} A_{t,n,m}, \tag{4}$$

where a low attitude value toward pro-environmental behaviour contributes to an identity of indifference towards the environment. Additionally, $\rho$ is a cultural inertia parameter representing the duration of the past considered, $s$ is a dummy variable for the discrete-time step in the present or past evaluated, $\bar{A}$, the average attitude over one time step for $M$ behaviours, $\delta \in [0, 1]$ is a discount parameter that produces a hyperbolic discounting effect.

## 2.4 Dynamic behavioural attitudes

The evolution of individuals' attitudes is given by

$$A_{t+1,n,m} = [1 - \phi_m] A_{t,n,m} + [\phi_m] S_{t,n,m}, \tag{5}$$

where the attitude to behaviour $m$ is modulated by $\phi_m$, a measure of conspicuous consumption, and $S_{t,n,m}$ is the social influence component due to an agent´s $K_n$ neighbours.

## 2.5 Imperfect social learning

We implement a multi-dimensional continuous model inspired by **?** to represent social learning in our model, with a weighted mean to aggregate the impact of each agent's neighbourhood. To model the imperfection of social transmission we add a Gaussian error $\varepsilon = G(0, \sigma_\varepsilon^2)$. A separate transmission error is applied to each of the $M$ behaviours after an individual has aggregated the attitudes of their ego network. Thus, the social learning component is given by

$$S_{t,n,m} = \left[ \sum_{k=1}^{K_n} \alpha_{n,k} A_{t,k,m} \right] + \varepsilon_{t,n,m}, \tag{6}$$

where $K_n$ is the total number of neighbours in the $n^{th}$ agent's social network. This is weighted by $\alpha_{n,k}$, which represents how much focal agent $n$ values the opinion of neighbouring agent $k$ within their ego network. Finally, $A_{t,m,k}$ is the positive attitude of agent $k$ towards behaviour $m$ in agent $n$'s neighbourhood.

## 2.6 Heterogeneous agent weighting in imitation

Our social network weighting matrix $\alpha_{t,n,k}$ is given by the softmax function

$$\alpha_{t,n,k} = \frac{e^{-\theta|I_n - I_k|}}{\sum_{k \neq n}^{K_n} e^{-\theta|I_n - I_k|}}, \tag{7}$$

where we consider the identity distance of $K_n$ neighbours and $\theta$ is a measure of confirmation bias. Thus, if $I_n$ and $I_k$ grow further apart over time, the influence they exert on each other also decreases. To ensure that the total weighting in an agent's ego neighbourhood is equal to one we normalise the values against the total weighting calculated.

# 3 Parameter description

In Table 1 we summarise the parameters required to initialise the model. We create a list of behaviour social susceptibility by linearly dividing the range phi_lower to phi_upper into $M$ values such that each behaviour has its own $\phi_m$. Additionally, there is the alpha_change parameter that determines how often identity is updated and whether there is behavioural independence. The four possible values are "static_uniform_weights", "static_culturally_determined_weights", "dynamic_culturally_determined_weights", "behavioural_independence". The default value is "dynamic_culturally_determined_weights". The first case, static uniform neighbour weighting, represents a society in which individuals value the opinion of all their neighbours in the Watts–Strogatz network equally such that $\alpha_{t,n,k} = 1/K_n$ and are unable to change their weighting over time. This is equivalent to having no confirmation bias, $\theta = 0$. In the second case, static culturally determined neighbour weighting, individuals calculate their social network weighting once based on their initial identities, according to Equation 7. Similarly to the first case, this is fixed for subsequent time periods of the experiment. In the third case, dynamic cultural neighbour weighting, we update $\alpha_{t,n,k}$ every time step, representing frequent social interactions.

Table 1: List of model input parameters including default values and ranges explored.

| Code Name | Variable Name | Representation | Definition | Range explored | Default value |
|---|---|---|---|---|---|
| save_timeseries_data | Time series Boolean | - | Whether or not to save time series data | - | 1 |
| compression_factor | Compression factor | - | How many time steps to skip between saves of timeseries data | - | 10 |
| set_seed | Stochastic seed | - | Sets stochastic seed used for Networkx and Numpy libraries | - | 1 |
| seed_list | Seed list | - | List of stochastic seed used when running the same experimental set-up over multiple stochastic starts | - | - |
| N | Number of individuals | $N$ | Total number of individuals in the social network | [10,500] | 200 |
| M | Number of behaviours | $M$ | Behaviours modelled per agent | [100,500] | 3 |
| K | Number of agent neighbours | $K$ | Average number of social connections per agent, varies slightly due to re-wiring probability of small-world network | [2,30] | 20 |
| prob_rewire | Probability of re-wiring | $p_r$ | Likelihood that connection between neighbouring agents is swapped to form long-distance or weak tie | - | 0.1 |
| cultural_inertia | Cultural inertia | $\rho$ | Duration over which past states influence current identity state | [1,1000] | 1000 |
| learning_error_scale | Social learning error standard deviation | $\sigma_\varepsilon$ | Standard deviation of Gaussian learning error representing the degree of perfect imitation | [0,0.5] | 0.02 |
| discount_factor | Discount factor | $\delta$ | Decrease in relative importance between two adjacent moments in time | [0,1] | 0.95 |
| homophily | Attribute homophily | $h$ | Degree of identity homogeneity in the initial social network | [0,1] | 0.95 |
| confirmation_bias | Confirmation bias | $\theta$ | How much agents only listen to those neighbours with similar identities | [0,100] | 20 |
| a_attitude | Initial attitude Beta | $a_A$ | One component of input into Beta distribution to generate initial distribution of agent attitudes | [0.05, 8] | 1 |
| b_attitude | Initial attitude Beta | $b_A$ | One component of input into Beta distribution to generate initial distribution of agent attitudes | [0.05, 8] | 1 |
| a_threshold | Initial threshold Beta | $a_T$ | One component of input into Beta distribution to generate initial distribution of agent thresholds | [0.05, 8] | 1 |
| b_threshold | Initial threshold Beta | $b_T$ | One component of input into Beta distribution to generate initial distribution of agent thresholds | [0.05, 8] | 1 |
| green_N | Number of green influencers | - | Number of individuals that broadcast one perfectly green behavioural attitude | - | 20 |
| time_steps_max | Max time steps | $\tau$ | Number of steps in simulation | - | 3000 |
| phi_lower | Lower bound of conspicuous consumption factor | $\phi_m$ | Lower bound of behaviour specific social susceptibility | - | 0.01 |
| phi_upper | Lower bound of conspicuous consumption factor | $\phi_m$ | Upper bound of behaviour specific social susceptibility | - | 0.05 |