Candidate number: 49045

TITLE HERE

Supervisor: Ekaterina (Katya) Oparina

Word count:

Submitted as partial fulfilment for

MSc in Behavioural Science

Department of Psychological and Behavioural Science

The London School of Economics and Political Science

Total word limit: 10000

Abstract (100-150)

Introduction (1000)

Gender gap in labour force participation is a world-wide phenomenon which is particularly pronounced in developing countries. Globally, the rate of labour force participation is about 75% among men while only 50% among women **(International Labour Organization, 2021)**. In regions such as Middle East, North Africa, and South Asia, the gender gap is even greater, with around 75% of men and 20% of women participating in the labour force **(International Labour Organization, 2021)**. In search of potential measures to narrow the gender gap, its causes have been extensively studied with a recent attention to the prominent role of social norms. Researchers speculate that interventions aiming at changing social norms may be the key to achieve greater gender equality in labour force participation **(Bursztyn et al., 2020; Codazzi et al., 2018; Jayachandran, 2021)**.

A recent attempt in developing such a social norm intervention was made by Bursztyn and colleagues **(2020)** in Saudi Arabia. In Saudi Arabia, gender norms exist that expect women to be absent from labour or segregated from men at workplace, and that women need approval from their male "guardian" (usually the husband or father) if they want to work outside the home **(Bursztyn et al., 2018)**. These may lead us to reasonably speculate that the low female labour participation rate in Saudi Arabia **(28% in 2022, International Labour Organization)** is because men don't allow their wives to work outside the home (WWOH) due to the internalisation of the gender norms (i.e., personal beliefs aligning with the norm that women should not work outside the home). Interestingly, however, Bursztyn and colleagues **(2020)** found that the critical factor was rather Saudi men's misperception of the injunctive norms regarding WWOH. Their findings revealed that 80-85% of surveyed Saudi men who were married and aged 18-35 reported to agree with the statement that women should be allowed to work outside the home, while more than three quarters underestimated this percentage. This phenomenon is also referred to as pluralistic ignorance (PI) **(Bursztyn et al., 2020)**.

Based on this finding, Bursztyn and colleagues **(2020)** designed an intervention to correct men's misperception regarding WWOH and found supporting evidence for its effectiveness in changing their labour supply decisions. They found that men who received the information of the true percentage of WWOH supporters (vs. those who did not receive the information) were significantly more likely to sign up for job matching service for their wives immediately after the intervention. Their wives were also more likely to have applied and interviewed for a job three to five months after the intervention. The researchers also tested the effect of a similar intervention on Saudi women in a field setting, finding that women who received the information on the percentage of men supporting WWOH (vs. those who did not received the information) were more likely to take up a part-time job outside the home instead of a position to work at home.

Recognising the potential of correcting norm misperception as an intervention measure to create behavioural change, the present research aims to use computer simulations to further explore its theoretical effectiveness when scaled up. Specifically, an agent-based model (ABM) will be built to simulate PI regarding WWOH in Saudi Arabia and its interventions. ABMs are a type of computational model that simulates in a synthetic environment the actions and interactions of autonomous agents, whose behaviours determine the evolution of the entire system **(Bandini et al., 2009)**. ABMs are particularly suitable to model and study dynamics and emergent properties in complex systems that are irreducible to lower-level descriptions (e.g., mathematical equations) **(Gustafsson & Sternad, 2010)**. One example of such a system is where agents interact dynamically and irregularly over time **(Johnson, 2009; Gustafsson & Sternad, 2010)**. The ability in modelling multi-agent complex systems makes ABMs widely applied in studying interventions to human behaviours in various fields. These include but are not limited to the effect of lock-down policy in epidemiology **(de Mooij et al., 2022)**, fishery policy in sustainable management **(Bailey et al., 2018)**, and interventions to the spread of misinformation **(Pilditch et al., 2022)**.

By building an ABM, the present research seeks to answer two broad questions: under what conditions PI regarding WWOH in Saudi Arabia are sustained and under what conditions social norm interventions can reduce PI and encourage WWOH behaviours? The paper proceeds as follows. Section 1 reviews literature that is relevant to model-building in the present research, including mathematical and computational models of belief and action dynamics, ABMs of pluralistic ignorance, and empirical research of existing social norm intervention strategies. Section 2 introduces the details of the model and the simulation. Section 3 presents the results. Section 4 discusses the implications of the findings and concludes.

1 Literature Review (2000)

1.1 Models of Belief and Action Dynamics

Pluralistic ignorance regarding WWOH as a population-level phenomenon is underlain by complex interpersonal interactions among people's private beliefs and public actions. One's labour supply decision and action can presumably influence their acquaintances' private belief and norm perception, and these influences in turn impact their acquaintances' labour supply decision and action as a function of private belief and norm perception. Such belief and action dynamics among interconnected and mutually influencing individuals have been studied widely through mathematical and computational models. These models assume that agents hold beliefs regarding certain issues. At each time step, these beliefs are modified based on their neighbours' beliefs according to some rules. The dynamics of belief distribution in a population over time is studied. This section will review formal methods to model beliefs and actions dynamics in a population and some applications of these models, which will serve as a basic framework for modelling PI regarding WWOH in the present research.

Models of belief and action dynamics are characterised by different assumptions about how agents' beliefs and actions are updated based on those of others **(Hassani et al., 2022)**. The most simplistic models represent agents' beliefs using binary variables, the classic ones of which include the Ising model **(Li et al., 2019)**, the voter model **(Holley & Liggett, 1975)**, and the Sznajd model **(Sznajd-Weron & Sznajd, 2000)**. These models usually assume that agents update beliefs by replacing them with those of their neighbours. For example, in the voter model, an agent i is randomly chosen at each time step, together with one of its neighbour j. The agent i then abandons its previous belief and takes up the opinion of its neighbour j. Therefore, in these simplistic models, the updating usually results in the agents being memoryless about their previous beliefs.

Other simple models represent agents' beliefs as continuous variables that take the value of a real number. These are exemplified by the classic Degroot model **(Berger, 1981)** and the bounded confidence model **(Rainer & Krause, 2002)**. Models of this type usually update agents' beliefs using a weighted average of an individual's belief and those of their neighbour(s). For example, in the bounded confidence model, an agent i is randomly chosen at each time step t, together with one of its neighbour j, who hold the beliefs $\sigma_i(t)$ and $\sigma_j(t)$, respectively. When the condition $|\sigma_i(t) - \sigma_j(t)| < \epsilon$ is met, the agent i updates its belief according to $\sigma_i(t+1) = (1-\alpha)\sigma_i(t) + \alpha\sigma_j(t)$, and the agent j according to $\sigma_j(t+1) = (1-\alpha)\sigma_j(t) + \alpha\sigma_i(t)$. In other words, when the selected agents hold similar enough beliefs according to a threshold ϵ, they update their beliefs as the weighted average of their previous beliefs based on a convergence parameter α.

More sophisticated models recognise the distinction between agents' private beliefs and public actions, modelling the former as a continuous variable and the latter a discrete one. The Continuous Opinions and Discrete Actions (CODA) model **(Martins, 2013)** and the recent social network opinions and actions evolutions (SNOAEs) model **(Zhan et al., 2022)** are examples of this category. To take the CODA model as an example, it models private beliefs as $P(A) \in [0, 1]$, the probability of the A being the best alternative. Agents have no access to other agents' private beliefs but only their public actions, which is used to update one's own private beliefs. For example, upon observing neighbour j displaying the action A at the time step t, an agent update its private beliefs according to the Bayes' theorem $P_{t+1}(A) = P(A|a_j = +1) \propto P_t(A)P(a_j = +1|A)$, where $P(A|a_j = +1)$ denotes the probability of A being the best action conditioned on the neighbour j displaying A, and $P(a_j = +1|A)$ the probability of the neighbour j displaying A conditioned on A being the best action (which is modelled as a constant).

It is important to notice that there is a recent trend to build psychologically realistic models to represent belief and behaviour dynamics in a population **(Duggins, 2017; Gavrilets, 2021; Tverskoi et al., 2023)**. Apart from agents' beliefs and actions, these models further consider the influence of social norm, the psychological tendency to conform to peers, external authorities, as well as material cost-benefit considerations **(Gavrilets, 2021)**. To incorporate these psychological and social complexities, a recent unifying modelling framework **(Gavrilets, 2021)** models agents' perception of others' private beliefs and public actions as two variables independent from their own beliefs and actions. This allows the representation of various psychological factors in updating beliefs and actions, including cognitive dissonance in taking a private belief (i.e., the aversion towards misaligned belief and action), social projection in perceiving others' private beliefs (i.e., the tendency to project one's own private belief onto others), and compliance with authority.

Models of belief and action dynamics have been applied to studying multiple themes associated with opinion change in a population. These include but are not limited to consensus reaching and polarisation **(Acemoglu & Ozdaglar, 2010; Hassani et al., 2022; Jager & Amblard, 2004; Duggins, 2017)**, the spread of (mis)information **(Watts, 2002; Watts & Dodds, 2007; Pilditch et al., 2022)**, and echo chamber formation **(Madsen et al., 2018; Fränken & Pilditch, 2021)**. Researchers are mostly interested in exploring the assumptions and conditions that can give rise to certain phenomena. For instance, a study reviews a series of mathematical models to inquire into the belief updating mechanisms that facilitate consensus reaching, information aggregation, and the spread of misinformation **(Acemoglu & Ozdaglar, 2010)**. It finds that in both Bayesian and non-Bayesian models, there is a tendency towards reaching consensus. Misinformation spreads with limited extent, which is due to the limited influence of misinformation on Bayesian agents in Bayesian models, and the lack of persistent disagreements in the population in non-Bayesian models. Another example is a series of research that studies how assumptions about network structures can influence information cascade, a phenomenon where a group of individuals make the same decision sequentially **(Watts, 2002; Watts & Dodds, 2007)** It finds that the distribution of the size of cascades depends on the connectivity of the network **(Watts, 2002)**, and that influential individuals (i.e., those who have the highest number of connections to others) are not sufficient for triggering cascades, but the existence of a group of individuals who are easily influenced by them is critical **(Watts & Dodds, 2007)**.

1.2 Agent-Based Models of Pluralistic Ignorance

Although the discussion about pluralistic ignorance is born in psychological research **(e.g., Prentice & Miller, 1993)**, there has been a growing interest in using ABMs to study the conditions under which this phenomenon can arise. This line of research usually defines PI as the situation where the majority of people in a population express opinions different from their private beliefs **(e.g., Seeme et al., 2016; Wang et al., 2014; Ye et al., 2019)**. The seminal simulation study on this topic focuses on the effects of network properties on holding inconsistent private beliefs and expressed opinions **(Centola et al., 2005)**. The simulation inserts to a population a few “true believers” of an unpopular norm (i.e., agents who hold a private belief held by few individuals in the population) who enforce other agents to adopt the belief. It reveals that the phenomenon of holding inconsistent private beliefs and expressed opinions cannot spread widely in the population if the population is fully connected, if the true believers are scattered in the population rather than clustered, or if ties in the network are randomly rewired, breaking those that originally exist between local neighbours.

Recent simulation studies of PI start to model psychological processes that have been theorised by research in social psychology. Two processes that have been mostly attended to are social conformity and cognitive dissonance **(e.g., Wang et al., 2014; Seeme, 2019)**. For instance, a research models the change in agents' expressed opinions as a result of the influence from the opinions groups formed among the neighbours, representing people's psychological tendency to conform to the group **(Wang et al., 2014)**. Moreover, the model follows the cognitive dissonance theory and asks the agents change their private attitudes when the group influence is moderate. When the group influence is strong, the model represents the situation where people are aware of the group influence and doesn't ask agents to change their private beliefs. Under these assumptions, the research finds that there is a widespread inconsistency between agents' private beliefs and expressed opinions, that is, pluralistic ignorance exists by definition.

Another example is a study that considers social conformity and cognitive dissonance while assuming the rationality of agents **(Seeme, 2019)**. The model assumes that the agents have a utility function that is proportional to the rewards from high conformity with its opinion group (i.e., expressing similar opinions with the mean opinion of the group) and the rewards from low cognitive dissonance (i.e., holding similar private beliefs and expressed opinions). Agents update their private beliefs and expressed opinions by maximising the value of the utility function. The study finds that different degrees of pluralistic ignorance arise under different conditions. When all agents make the update simultaneously, the more opinion groups in the population, the more agents demonstrate inconsistency between private beliefs and expressed opinions; when the updating happens sequentially (i.e., agents make the updating one by one), all agents end up demonstrating inconsistency between private beliefs and expressed opinions.

1.3 Social Norm Interventions in Empirical Studies

The strategy adopted by **Bursztyn and colleagues (2020)** to correct Saudi Arabian men's misperception regarding the injunctive norm about WWOH is by providing information about the true percentage of men supporting WWOH. Such social norm interventions that provide people with information about peers have been widely studied in order to change behaviours **(Yamin et al., 2019; Miller & Prentice, 2016)**. They usually provides people with summary information regarding the beliefs or actions in a population in a statistical format such as "82% of married Saudi men aged 18–35 agreed that women should be allowed to work outside the home" **(Bursztyn et al., 2020; Yamin et al., 2019)**. In addition to the intervention to Saudi men's misperception about the WWOH norms **(Bursztyn et al., 2020)**, social norm interventions via summary information have also been applied to changing many other behaviours. For example, a study gave the information that most other students were concerned about climate change to those who were themselves concerned, and found that the student provided with such information were more willing to participate in a discussion on climate change compared to those not provided with the information **(Geiger & Swim, 2016)**.

A variation of this type of interventions provides people with personalised information combining the the summary information with comments about how individuals' beliefs or actions compare to their peers **(Yamin et al., 2019; Miller & Prentice, 2016)**. Those interventions usually follow this structure: "You said you drink 10 drinks per week and that you think the typical student drinks 15. The actual average is 4.6 drinks. You drink more than 80% of other college students" **(Neighbors et al. 2011)**. This type of interventions has been applied to and found generally effective in reducing alcohol consumption, increasing tax compliance, and promoting sustainable behaviours **(Miller et al., 2013; Schultz, 1999; Hallsworth et al., 2017)**.

On closer scrutiny, interventions via summary information essentially intend to change people's beliefs about social norms via a piece of persuasive message. We then can expect factors that have been discussed by psychological research about persuasion will also be important to the effect of the social norm Interventions under discussion. Source credibility is such a factor as it has been both theorised as a heuristic cue to processing persuasive messages **(Petty & Cacioppo, 1986; Chen & Chaiken, 1999)** and been attended to as a factor that needs to be considered in designing policy interventions **(Tankard & Paluck, 2016; Dolan et al., 2012)**. Despite the existence of some nuances, high source credibility has been generally found to have positive effects on people's attitudes in contexts including making arguments and political persuasion **(Housholder & LaMarre, 2014; Chaiken & Maheswaran, 2014; see Pornpitakpan, 2004 for a review)**. Few studies exist that explicitly compare the effects of social norm intervention via summary information with various levels of source credibility. A lab experiment conducted among American college students that explored this topic found that when perceived as having high credibility, the information that intended to reduce misperception of drinking norm was more effective in decreasing perceived weekly drinking compared to when it was perceived as having low credibility **(Hummer & Davison, 2016)**.

1.4 The Present Research

To briefly outline the model built by the present study, agents are generated being interconnected via a social network. Their belief regarding WWOH (i.e., whether women should be allowed to work outside the home) and norm perception (i.e., how many other men believe that women should be allowed to work outside the home) are treated as two independent beliefs that can be revised by external information. The belief is influenced by other agents' beliefs (if accessible) and the norm perception is influenced by either others' beliefs (if accessible) or actions. The belief regarding WWOH and norm perception together determine agents' preference for WWOH, which in turn determines their actual WWOH actions. The summary information given by the intervention affects agents' norm perception, the degree of which is influenced by perceived credibility of the information.

The present study makes contributions to the existing literature in the following aspects. First of all, it treats norm perception as an independent belief which can be revised over time based on observations of other people (See the Methodology section for a discussion on the theoretical basis of this treatment). This is to adapt the novel conceptualisation made by recent work on mathematical models of belief and action dynamics to the context of computer simulations **(Gavrilets, 2021; Tverskoi et al., 2023)**. This approach has not been extensively used yet as most models of belief and action dynamics primarily focus on the change in beliefs about a certain issue itself rather than those about the norm **(e.g., Rainer & Krause, 2002; Martins, 2013)**, while models of PI usually model the influence of norm as a direct result of observations of others **(e.g., Wang et al., 2014; Seeme, 2019)**. Second, rooted in a real-world phenomenon, the model explores both the conditions for PI to be sustained and effectiveness of its interventions. This adds to the existing models' primary focus on the conditions for the emergence of PI **(Seeme, 2019)**. Finally, the present study applies ABMs to social norm interventions. Due to the empirical focus of the existing social norm intervention studies **(Yamin et al., 2019)**, this is an area of policy interventions that have been rarely explored by simulation studies despite its widespread application in other policy contexts.

2 Methodology (2000-3000)

2.1 Model Setup

The present research builds a model populated by $n = 100$ agents who are connected by undirected ties and form a social network on a two-dimensional 61×61 lattice. The social network structure follows the small-world network model, which is characterised by being both highly clustered and having low average path lengths **(Watts & Strogatz, 1998)**. This network model is chosen because many real-world social networks, such as email messages, academic co-authorship, and Facebook following, resemble the small-world network, which makes this model a suitable baseline assumption for modelling an unknown network **(Newman, 2003; Ugander et al., 2011)**. The network is programmed by generating agents on a circle, wiring all agents with their closest two neighbours, and rewiring p_r of all ties **(Watts & Strogatz, 1998; Wilensky, 2015)**.

2.1.1 Initial Setup

Agents are initialised with a belief regarding WWOH B_i ($B_i \in [0, 1]$), a norm perception N_i (a random variable with a normal distribution with the mean μ_{N_i} and standard deviation σ_{N_i}), and a WWOH action A_i ($A_i \in \{0, 1\}$). The variable B_i represents the private belief regarding to what extent women should be allowed to work outside the home. Agents with $B_i \in [0.5, 1]$ are regarded as WWOH supporters and the others are regarded as non-supporters. As the initial condition,

$p_s = 0.8$ of all agents are generated as supporters, and the B_i of these agents are drawn from the upper half of a truncated normal distribution with $\mu_B = 0.5$, $\sigma_B = 0.2$ (truncated between 0 and 1). The other agents have a B_i from the lower half of this distribution.

The initial norm perception of each agent is determined by five samples drawn from a truncated normal distribution (between 0 and 1) with $\mu_{N_{prior}} = 0.5$, $\sigma_{N_{prior}} = 0.2$. The mean of the five sampled values provides the prior norm perception (i.e., μ_{N_i} of the distribution of N_i) and the standard deviation provides the initial uncertainty about the perception (i.e., σ_{N_i}).

A certain proportion $p_A = 0.05$ of WWOH supporters demonstrate WWOH actions. The value of p_s and p_A are drawn from the empirical statistics about Saudi Arabian men's beliefs and actions regarding WWOH **(Bursztyn et al., 2020)**.

2.1.2 Updating Procedures

At each time step, agents reach out to all linked neighbours and can communicate with each of them with a probability p_c. If an agent communicates with a neighbour, the agent updates both its belief and norm perception regarding WWOH based on the neighbour's belief; if the communication doesn't happen, the agent updates its norm perception using the inferred belief from the neighbour's action. Figure 1 gives the flow chart of this procedure. This approach follows the CODA model and its extensions in terms of modelling beliefs as a continuous variable and actions a binary one, as well as updating beliefs depending on whether other agents' beliefs are accessible **(Martins, 2008; Zhan et al., 2022)**. This is aligned with the aim of the current study to model the WWOH action, which is a binary decision, while also capturing the possibly continuous nature of beliefs regarding WWOH.

Agent i, upon communicating with and accessing the belief of neighbour j, updating its own belief with bounded confidence. The bounded confidence model is able to capture the ubiquitous psychological phenomenon of being more likely to be influenced by people like ourselves while keeps the computation simple **(Cialdini & Goldstein, 2004)**. Previous research has successfully nested the bounded confidence model of opinion dynamics in the framework of the CODA model in the context of a social network, which serves as a foundation for the current model **(Zhan et**

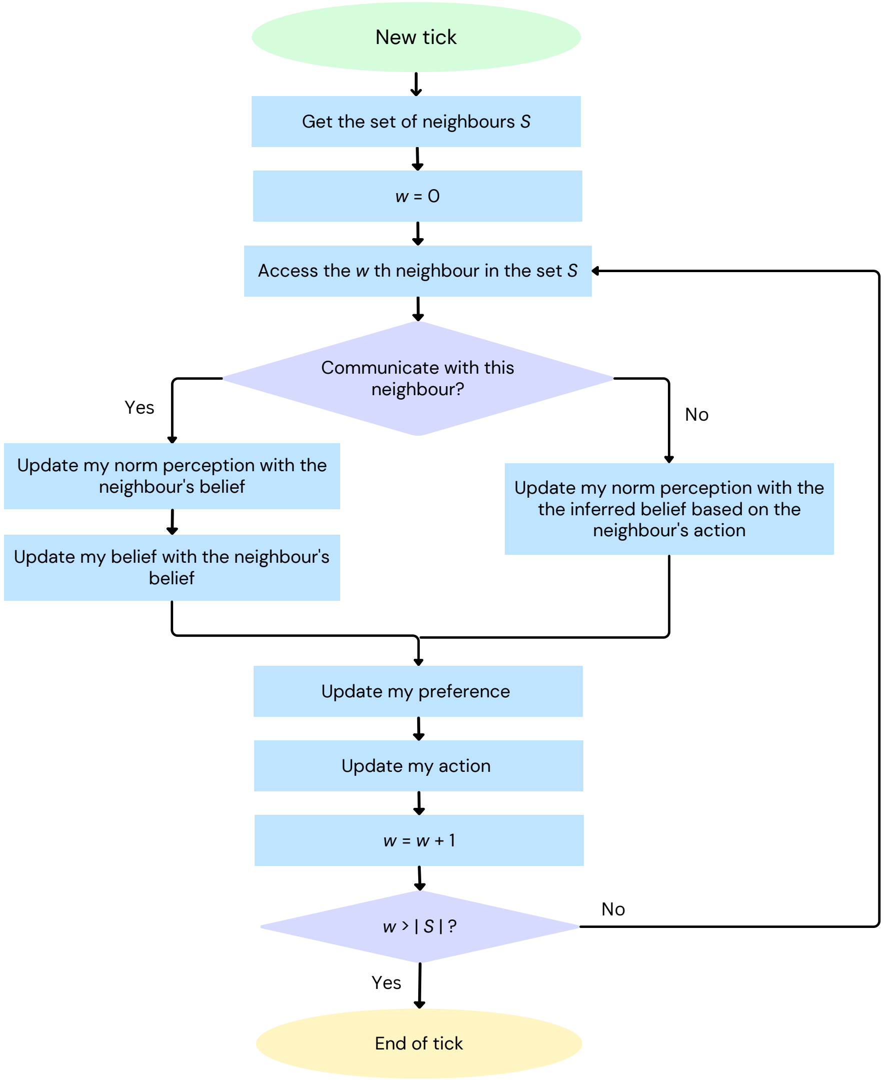
New tick
Get the set of neighbours *S*
w = 0
Access the *w* th neighbour in the set *S*
Communicate with this neighbour?
Yes
No
Update my norm perception with the neighbour's belief
Update my belief with the neighbour's belief
Update my norm perception with the the inferred belief based on the neighbour's action
Update my preference
Update my action
w = *w* + 1
w > | *S* | ?
No
Yes
End of tick

Figure 1: Main updating procedure. The flow chart shows each agent's behaviour at each time step.

al., 2022). Specifically, agent i updates its belief based on neighbour j's belief using the following equation:

$$B_i' = \begin{cases} B_i, & \text{if } |B_i - B_j| > \epsilon_i \\ B_i + \alpha(B_j - B_i), & \text{if } |B_i - B_j| \leq \epsilon_i \end{cases} \tag{1}$$

where $\alpha \in (0, 0.5]$ is the convergence parameter and ϵ_i is the bounded confidence threshold of agent i which is a value drawn from an exponential distribution with the mean ϵ_{mean} and is a fixed value over time. An exponential distribution is chosen so that the value is guaranteed to be greater than zero but does not have an upper limit (The same reason applies to other parameters that are drawn from an exponential distribution).

Agents update their norm perception N_i following the Bayesian inference model **(Krauß et al.,1999).** Perception of social norms is very suitable to be modelled using models of probabilistic inference such as the Bayesian inference model. This is because norm perception can be seen as a piece of abstract knowledge that is inferred from limited observational data, as we can state what we believe to be most others' beliefs while we are only able to observe (at most) the beliefs and actions of our acquaintances. To understand this kind of beliefs, cognitive scientists have been using various probabilistic models for conceptualisation **(Tenenbaum et al., 2011; Chater et al., 2006)**. The present research adopts one specific type of probability models, the Bayesian inference model, to model norm perception. The Bayesian model is chosen since it is widely adopted in research of human perception and memory and has been shown adequate in modelling inference about everyday events **(Griffiths and Tenenbaum, 2006; Tavoni et al., 2022).**

Using Bayes' theorem, we can update the conditional probability of agent i's norm perception given evidence using the following equation:

$$f(N_i \mid E_i) = \frac{f(N_i)f(E_i \mid N_i)}{\int f(N_i)f(E_i \mid N_i)dN_i}, \tag{2}$$

where $f(N_i)$ denotes the prior probability of norm perception of agent i, E_i the evidence (either agent j's belief or inferred belief from its action, see Equation 5), and $f(E_i \mid N_i)$ the conditional probability of the evidence given N_i. Assuming normal distributions for both $f(N_i)$ and $f(E_i \mid N_i)$, the following updating rules for the mean and standard deviation of agents' norm perception can be derived (see the appendices for the derivation):

$$\mu'_{N_i} = \lambda\mu_{N_i} + (1-\lambda)E_i, \tag{3}$$

$$\sigma'^2_{N_i} = \lambda \sigma^2_{N_i}, \tag{4}$$

where $\lambda = \frac{c_i^2}{c_i^2 + \sigma^2_{N_i}}$. The variable c_i is the degree of agent i's lack of confidence in its norm perception (i.e., the standard deviation of the conditional probability of the evidence, $f(E_i \mid N_i)$). c_i is a value drawn from an exponential distribution with the mean c_{mean} and is a fixed value over time.

Since agents may or may not communicate with neighbour j, the value of E is given by the following equation:

$$E_i = \begin{cases} B_j, & \text{if agent } i \text{ communicates with agent } j \\ P(B_j \mid A_j), & \text{if agent } i \text{ doesn't communicate with agent } j \end{cases} \tag{5}$$

$P(B_j \mid A_j = 1) = 0.9$ and $P(B_j \mid A_j = 0) = 0.1$ by stipulation in the simulations reported in the present research.

After updating their belief and norm perception, agents update preference and action regarding WWOH based on these two values. The preference is determined by the equation

$$P_i = r_i B_i + (1 - r_i) N_i, \tag{6}$$

where r_i represents each agent's resistance to social norm, which is drawn from an exponential distribution with the mean r_{mean} and is a fixed value over time. The WWOH action is determined by the equation:

$$A_i = \begin{cases} 0, & \text{if } P_i \in [0, 0.5) \\ 1, & \text{if } P_i \in [0.5, 1] \end{cases} \tag{7}$$

2.1.3 Interventions

A social norm intervention using summary information about the norm is implemented in the model. In the intervention, the summary information about the true norm $T_t = \overline{B_i}$ is distributed to all agents at the selected time step t. Agents take this as a persuasive message, the credibility of which is perceived as s_i, and update their norm perception in a similar manner as they update their beliefs about WWOH upon learning other agents' beliefs. Formally, this is expressed as:

$$\mu'_{N_i} = \begin{cases} \mu_{N_i}, & \text{if } |\mu_{N_i} - T_t| > \epsilon_i \\ \mu_{N_i} + \alpha \cdot s_i \cdot (T_t - \mu_{N_i}), & \text{if } |\mu_{N_i} - T_t| \leq \epsilon_i \end{cases} \tag{8}$$

The variable s_i is drawn from a truncated normal distribution (between 0 and 1) with the mean s_{mean} and s_{sd} and is a fixed value over time.

The time steps when the intervention is implemented are $t_1 = 30$, $t_2 = 40$, and $t_3 = 50$. The intervention is implemented in three ways. First, the summary information is distributed once at time t_1. Second, the information is distributed twice at time t_1 and t_2. Finally, the information is distributed at all three time steps. At the intervention time steps, agents do not update their norm perception in the regular manner via observation (or communication) but only through the summary information. Other procedures remain unchanged.

2.2 Simulations

The model will be run using the parameter values listed in Table 1. For each set of parameter combination, 25 simulations will be run and each simulation will run for 150 time steps. The model is written in NetLogo 6.3 **(Wilensky, 1999)** and the simulations are run in RNetLogo package **xx.xx** in R **xx.xx (Thiele 2014; citation needed)**.

The following main outcomes are measured at each time step of the simulations: 1) $P_u = P(\mu_{N_i} < P_s)$, proportion of agents who underestimate the proportion of WWOH supporters, 2) $P_I = P((B_i \geq 0.5 \wedge A_i = 0) \vee (B_i < 0.5 \wedge A_i = 1))$, the proportion of agents with inconsistent beliefs and actions regarding WWOH, and 3) $P_A = P(A_i = 1)$, proportion of agents who demonstrate WWOH actions. Outcome 1) and 2) are the key measures for PI regarding WWOH, despite assuming different definitions of PI. Outcome 1) follows the empirical study about Saudi Arabia and defines PI as the phenomenon that exists if and only if more than half of all agents underestimate the true proportion of agents who support WWOH **(Bursztyn et al., 2020)**.

Outcome 2) follows previous models of PI and defines PI as the situation where more than half of all agents act differently from their beliefs **(e.g., Ye et al., 2019)**.

For exploratory and potentially explanatory purpose, three extra outcomes are also measured at each time step: 1) $\overline{B_i}$, average beliefs regarding WWOH among all agents, 2) $P_s = P(B_i \geq 0.5)$, proportion of agents who are WWOH supporters, and 3) $\overline{\mu_{N_i}}$, average norm perception regarding WWOH among all agents.

Table 1: A List of Parameters

Parameters	Description	Values
Model setup		
n	Number of agents	100
p_r	Rewiring probability in small-world network setup	0.3, 0.5, 0.7
p_s	Proportion of agents who are WWOH supporters at the beginning of simulations	0.8
μ_B	Mean of the normal distribution from which the initial belief regarding WWOH is drawn	0.5
σ_B	Standard deviation of the above distribution	0.2
$\mu_{N_{prior}}$	Mean of the normal distribution from which the initial norm perception is drawn	0.5
$\sigma_{N_{prior}}$	Standard deviation of the above distribution	0.2
p_A	Proportion of agents who demonstrate WWOH action at the beginning of simulations	0.05
Updating Procedures		
p_c	Probability of communicating with each neighbour	0.1, 0.3, 0.5
ϵ_{mean}	Mean of the exponential distribution from which the bounded confidence threshold is drawn	0.3, 0.5, 0.7
α	Convergence parameter	0.1, 0.3, 0.5
c_{mean}	Mean of the exponential distribution from which the agents' degree of lack of confidence in norm perception is drawn	0.1, 0.3, 0.5
r_{mean}	Mean of the exponential distribution from which the resistance to norm is drawn	0.1, 0.3, 0.5
Interventions		
s_{mean}	Mean of the normal distribution from which the credibility perception of the summary information is drawn	0.3, 0.5, 0.7
s_{sd}	Standard deviation of the above distribution	0.2

3 Results (2000-2500)

4 Discussion and Conclusion (1000-1500)

Reference

Appendices

The Derivation of Equation 3 and 4

Since agents' norm perception N_i is a random variable with a normal distribution with the mean μ_{N_i} and standard deviation σ_{N_i}, we have the probability density function (PDF) of N_i as:

$$f(N_i) \propto \text{exp}(-\frac{(N_i - \mu_{N_i})^2}{2\sigma_{N_i}^2}).$$

The conditional PDF of the evidence given agent i's prior norm perception in Equation 2 is given by

$$f(E_i \mid N_i) \propto \text{exp}(-\frac{(E_i - N_i)^2}{2c_i^2}).$$

Following the Equation 2, we have the posterior PDF of agent i's norm perception as

$$\begin{aligned} f(N_i \mid E_i) &= C \cdot \text{exp}(-\frac{(N_i - \mu_{N_i})^2}{2\sigma_{N_i}^2})\text{exp}(-\frac{(E_i - N_i)^2}{2c_i^2}) \\ &= C \cdot \text{exp}(-\frac{(N_i - \mu_{N_i})^2}{2\sigma_{N_i}^2} - \frac{(E_i - N_i)^2}{2c_i^2}) \\ &= C \cdot \text{exp}(-\frac{(N_i - \mu'_{N_i})^2}{2\sigma'^2_{N_i}} - \frac{(E_i - \mu_{N_i})^2}{2(\sigma_{N_i}^2 + c_i^2)}), \end{aligned}$$

where C is a normalising constant. Since the second term in the exponent does not involve the random variable N_i, we can incorporate it into the constant and rewrite the expression as

$$f(N_i \mid E_i) \propto \text{exp}(-\frac{(N_i - \mu'_{N_i})^2}{2\sigma'^2_{N_i}}),$$

where

$$\mu'_{N_i} = \frac{E_i\sigma_{N_i}^2 + \mu_{N_i}c_i^2}{\sigma_{N_i}^2 + c_i^2} \text{ and } \sigma'^2_{N_i} = \frac{\sigma_{N_i}^2 c_i^2}{\sigma_{N_i}^2 + c_i^2}.$$