

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

# Contents

<b>TITLE HERE</b>	<b>1</b>
<b>Abstract (100-150)</b>	<b>4</b>
<b>Introduction (1000)</b>	<b>5</b>
<b>1 Literature Review (2000)</b>	<b>7</b>
1.1 Models of Belief and Action Dynamics . . . . .	7
1.2 Agent-Based Models of Pluralistic Ignorance . . . . .	9
1.3 Social Norm Interventions in Empirical Studies . . . . .	11
1.4 The Present Research . . . . .	12
<b>2 Methodology (2000-3000)</b>	<b>13</b>
2.1 Model Setup . . . . .	13
2.1.1 Initial Setup . . . . .	14
2.1.2 Updating Procedures . . . . .	14
2.1.3 Interventions . . . . .	18
2.2 Simulations . . . . .	18
<b>3 Results (2000-2500)</b>	<b>21</b>
3.1 Factors Associated With Persistent Pluralistic Ignorance . . . . .	21
3.2 Intervention Effects on Pluralistic Ignorance and WWOH Action . . . . .	23
<b>4 Discussion (1000-1500)</b>	<b>26</b>
<b>5 Conclusion</b>	<b>30</b>
<b>Reference</b>	<b>31</b>
<b>Appendix: The Derivation of Equation 3 and 4</b>	<b>32</b>

## List of Figures

- 1 **Main updating procedure.** The flow chart shows each agent's behaviour at each time step. . . . . 15
- 2 **Pluralistic ignorance and WWOH action over time when no intervention is implemented.**  $P_U$ ,  $P_I$ ,  $P_A$  denote the rate of underestimation, the rate of inconsistency, and the rate of WWOH actions, respectively. Central lines are average values of simulations with the same parameter setting, and shaded areas are the 95% confidence intervals. Probability of communication with neighbours  $p_c$  is shown on facet columns, and the mean resistance to norm  $r_{mean}$  is shown on facet rows. Parameters not shown on the plot are set at the following values:  $p_r = 0.5$ ,  $\epsilon_{mean} = 0.5$ ,  $\alpha = 0.3$ ,  $c_{mean} = 0.3$ . . . . . 33
- 3 **Estimated effects of the intervention on the rate of inconsistency at the final time step.** Point estimates of the effects of the intervention are plotted under different combinations of parameter values. Blue dashed lines indicate 0 effect. Convergence parameter  $\alpha$  is shown on facet columns, and the mean bounded confidence threshold  $\epsilon_{mean}$  is shown on facet rows. Parameters not shown on the plot are set at the following value:  $p_r = 0.5$ . . . . . 34
- 4 **Estimated effects of the intervention on the rate of WWOH actions at the final time step.** Point estimates of the effects of the intervention are plotted under different combinations of parameter values. Blue dashed lines indicate 0 effect. Convergence parameter  $\alpha$  is shown on facet columns, and the mean bounded confidence threshold  $\epsilon_{mean}$  is shown on facet rows. Parameters not shown on the plot are set at the following value:  $p_r = 0.5$ . . . . . 35

5	<b>Effects of the intervention on pluralistic ignorance and WWOH action over time.</b> $P_U$ , $P_I$ , $P_A$ denote the rate of underestimation, the rate of inconsistency, and the rate of WWOH actions, respectively. Central lines are average values of simulations with the same parameter setting, and shaded areas are the 95% confidence intervals. Blue dashed lines indicate the time step when the intervention is implemented. Mean perceived credibility of the information source $s_{mean}$ is shown on facet columns, and the intervention type is shown on facet rows. Parameters not shown on the plot are set at the following values: $p_r = 0.5$ , $p_c = 0.3$ , $\epsilon_{mean} = 0.5$ , $\alpha = 0.3$ , $c_{mean} = 0.1$ , $r_{mean} = 0.1$ . . . . .	36
---	--	----

**List of Tables**

1	<b>A List of Parameters</b> . . . . .	20
2	<b>Effects of parameters on pluralistic ignorance and WWOH action at the final time step when no intervention is implemented</b> . . . . .	22
3	<b>Effects of interventions on pluralistic ignorance and WWOH action at the final time step</b> . . . . .	24

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 per-

centage 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 study investigates the factors associated with the persistence of PI regarding WWOH in Saudi Arabia and the conditions under which 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. Section 5 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  $\epsilon$ , they update their beliefs as the weighted average of their previous beliefs based on a convergence parameter  $\alpha$ .

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. The initial condition is set to align with the empirical findings in Saudi Arabia such that pluralistic ignorance exists in the population and a small proportion of agents demonstrate WWOH actions (**Bursztyn et al., 2020**). Agents' 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 updated by external information. The belief regarding WWOH and norm perception together update agents' preference for WWOH, which in turn determines their actual WWOH actions. A social norm intervention via summary information can be implemented once, twice, or three times. The intervention, if any, updates 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. This is to adapt the cutting-edge conceptualisation made by recent works 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 computational models of PI which primarily focus on the conditions for the emergence of PI (**Seeme, 2019**). Finally, the present study applies ABMs to simulating 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 have been rarely explored by simulation studies despite its widespread application in other policy contexts.

The present study asks two questions: 1) what factors represented by the model are associated with the persistence of PI regarding WWOH in Saudi Arabia, and 2) when PI regarding WWOH persists at the empirical level, whether social norm interventions can reduce PI and encourage WWOH behaviours, and if so, under what conditions. No explicit hypotheses are formed concerning the first question because it is exploratory in nature. Regarding the second question, I expect that social norm interventions can reduce PI and encourage WWOH behaviours if the intervention is implemented multiple times or if the perceived credibility of the information source is high.

## **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  $61 \times 61$  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  $\mu_{N_i}$  and standard deviation  $\sigma_{N_i}$ ), 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.

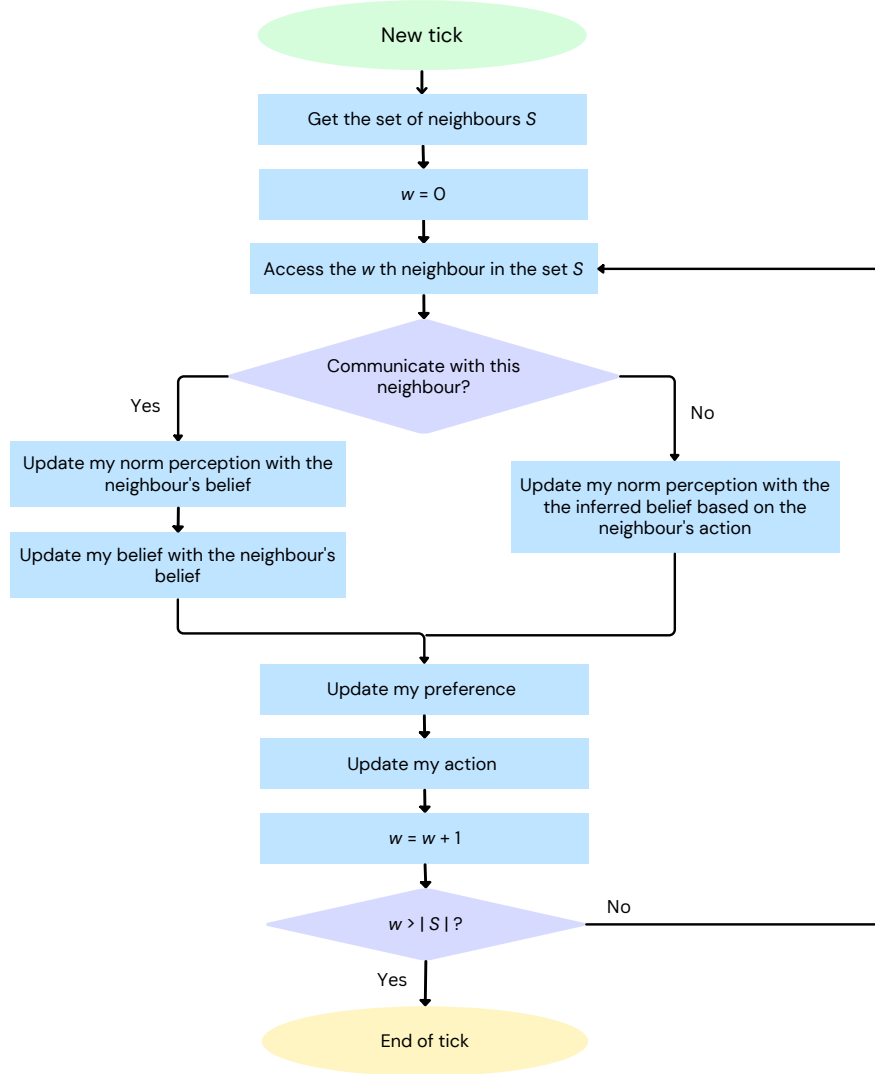
The initial norm perception of each agent is determined by five samples drawn from a truncated normal distribution (between 0 and 1) with  $\mu_{N_{prior}} = 0.5$ ,  $\sigma_{N_{prior}} = 0.2$ . The mean of the five sampled values provides the prior norm perception (i.e.,  $\mu_{N_i}$  of the distribution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

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.



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

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 al., 2022). Specifically, agent  $i$  updates its belief based on neighbour  $j$ ’s belief using the following



equation:

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

where  $\alpha \in (0, 0.5]$  is the convergence parameter and  $\epsilon_i$  is the bounded confidence threshold of agent  $i$  which is a value drawn from an exponential distribution with the mean  $\epsilon_{mean}$  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 (**KrauB 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:

$$f(N_i | E_i) = \frac{f(N_i)f(E_i | N_i)}{\int f(N_i)f(E_i | N_i)dN_i}, \quad (2)$$

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  $f(E_i | N_i)$  the conditional

probability of the evidence given  $N_i$ . Assuming normal distributions for both  $f(N_i)$  and  $f(E_i | N_i)$ , the following updating rules for the mean and standard deviation of agents' norm perception can be derived (see the appendices for the derivation):

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

$$\sigma'^2_{N_i} = \lambda \sigma^2_{N_i}, \quad (4)$$

where  $\lambda = \frac{c_i^2}{c_i^2 + \sigma_{N_i}^2}$ . 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,  $f(E_i | N_i)$ ).  $c_i$  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:

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

$P(B_j | A_j = 1) = 0.9$  and  $P(B_j | A_j = 0) = 0.1$  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

$$P_i = r_i B_i + (1 - r_i) N_i, \quad (6)$$

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:

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

### 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:

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

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 and the simulations are run in NetLogo 6.3 (Wilensky, 1999). The model and data are available in the online repository (OSF: [\(insert link here\)](#)).

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 (also referred to as the rate of underestimation), 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 (also referred to as the rate of inconsistency), and 3)  $P_A = P(A_i = 1)$ , proportion of agents who demonstrate WWOH actions (also referred to the rate of 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 phenomenon that exists if and only if more than half of all agents act differently from their private beliefs (**e.g., Ye et al., 2019**).

For exploratory and potentially explanatory purpose, four 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, 3)  $\overline{\mu_{N_i}}$ , average norm perception regarding WWOH among all agents, and 4)  $\overline{\sigma_{N_i}}$ , average uncertainty of norm perception among all agents.

**Table 1: A List of Parameters**

Parameters	Description	Values
<i>Model setup</i>		
$n$	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
$\mu_B$	Mean of the normal distribution from which the initial belief regarding WWOH is drawn	0.5
$\sigma_B$	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
<i>Updating Procedures</i>		
$p_c$	Probability of communication with each neighbour	0.1, 0.3, 0.5
$\epsilon_{mean}$	Mean of the exponential distribution from which the bounded confidence threshold is drawn	0.3, 0.5, 0.7
$\alpha$	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
<i>Interventions</i>		
$s_{mean}$	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)

Given the large number of varying parameters and resulting parameter combinations, not all results are reported in the main text for the sake of brevity. The results that support the claims in this section but can't be reported here due to space limitation are reported in the Appendices.

#### 3.1 Factors Associated With Persistent Pluralistic Ignorance

To explore the factors associated with the persistent level of pluralistic ignorance regarding WWOH that was found in the empirical study, the present study investigates the three outcome variables when no intervention is implemented. Linear regressions with full interactions are run including all varied parameters as predictors (except for  $s_{mean}$  which is only relevant to the intervention). Table 2 shows the effects of each parameter at the final time step on the rate of underestimation, the rate of inconsistency, and the rate of WWOH actions when controlled for other parameters. All else equal, no parameters or interaction terms have statistically significant effects on the rate of underestimation. The probability of communication with neighbours ( $p_c$ ) and mean resistance to norm ( $r_{mean}$ ) have negative effects on the rate of inconsistency and positive effects on the rate of WWOH actions. No interaction term has a significant effect on either the rate of inconsistency or the rate of WWOH actions.

To revealed the process of the two parameters influencing the outcomes, full processes of the simulations are plotted. As Figure 2 shows, both the probability of communication and mean resistance to norm influence the change of outcome variables over time. Particularly, resistance to norm being equal, higher probability of communication decreases the rate of underestimation and the rate of inconsistency while increasing the rate of WWOH actions. Higher resistance to norm has similar effects on the three outcome variables when the probability of communication is held equal.

When the resistance to norm is low ( $r_{mean} = 0.1$ , facet row one), low ( $p_c = 0.1$ ) and medium probability of communication ( $p_c = 0.3$ ) keep the rate of underestimation and the rate of inconsistency above 0.5, and the rate of WWOH actions low. Pluralistic ignorance exists by either one of the two definitions ( $P_U > 0.5$  or  $P_I > 0.5$ ). When the resistance to norm is low and the probability of com-

**Table 2: Effects of parameters on pluralistic ignorance and WWOH action at the final time step when no intervention is implemented**

	$P_U$			$P_I$			$P_A$		
	$\beta$	$t$	$p$	$\beta$	$t$	$p$	$\beta$	$t$	$p$
$p_r$	0.03 (0.32)	0.10	.919	0.21 (0.27)	0.76	.447	-0.23 (0.27)	-0.84	.402
$p_c$	0.34 (0.49)	0.69	.492	-1.65*** (0.42)	-3.97	<.001	1.93*** (0.42)	4.59	<.001
$\epsilon_{mean}$	0.16 (0.32)	0.50	.616	0.31 (0.27)	1.13	.259	-0.13 (0.27)	-0.47	.638
$\alpha$	0.30 (0.49)	0.60	.546	0.28 (0.42)	0.68	.495	-0.13 (0.42)	-0.32	.749
$c_{mean}$	0.09 (0.49)	0.17	.861	-0.23 (0.42)	-0.54	.588	0.32 (0.42)	0.77	.441
$r_{mean}$	-0.17 (0.49)	-0.34	.736	-1.58*** (0.42)	-3.80	<.001	1.88*** (0.42)	4.48	<.001

*Note.*  $P_U$ ,  $P_I$ ,  $P_A$  denote the rate of underestimation, the rate of inconsistency, and the rate of WWOH actions, respectively. Linear regressions are run with full interaction. Intercepts and interaction terms are omitted for brevity (see Supplementary Table S1 for the full table).

\* $p < 0.1$ , \*\* $p < 0.05$ , \*\*\* $p < 0.01$ .

munication is high ( $p_c = 0.5$ ), the rate of underestimation slightly decreases but still remains above 0.5, while the rate of inconsistency drops under 0.5 and the rate of WWOH increases drastically to a high level.

Under the condition of medium resistance to norm ( $r_{mean} = 0.3$ , facet row 2) and low probability of communication ( $p_c = 0.1$ ), the rate of underestimation remains virtually constant around 1. The rate of inconsistency decreases moderately but still remains above 0.5, and the rate of WWOH actions increases moderately. Pluralistic ignorance also exists by either one of the two definitions. In all other conditions, although the rate of underestimation remains high, the rate of inconsistency drops under 0.5 and the rate of WWOH increases quickly at the very beginning of the simulations.

In sum, whenever the probability of communication with neighbours is low or medium ( $p_c = 0.1$  or  $p_c = 0.3$ ) and the mean resistance to norm is low ( $r_{mean} = 0.1$ ), pluralistic ignorance (PI) exists by both definitions mentioned in the main text ( $P_U > 0.5$  or  $P_I > 0.5$ ). Sensitivity analyses indicate that this conclusion holds true in all combinations of  $p_r$ ,  $\epsilon_{mean}$ ,  $\alpha$ ,  $c_{mean}$  (see Table S1 in Supplementary Materials). Moreover, when the probability of communication with neighbours is

low ( $p_c = 0.1$ ) and the mean resistance to norm is medium ( $r_{mean} = 0.3$ ), whether PI exists as defined through the rate of inconsistency ( $P_I > 0.5$ ) depends on the values of other parameters (Table S1).

### 3.2 Intervention Effects on Pluralistic Ignorance and WWOH Action

To investigate the the effects of interventions on PI regarding WWOH, parameter combinations that sustain PI and WWOH action at the level reported in the empirical study are first identified. The empirical rate of underestimation of WWOH norm among Saudi men is around  $P_U = 0.8$  and the rate of WWOH actions is around  $P_A = 0.05$  (Bursztyn et al., 2020). One-sample  $t$ -tests on the outcome variables reveal that no parameter combinations produce  $P_U$  and  $P_A$  close to these values (i.e., the empirical rate falls within the 95% confidence intervals of  $P_U$  and  $P_A$  produced by the model) simultaneously at the final time step. Some parameter combinations do give rise to  $P_U$  close to 0.8, but the rates of WWOH actions are significantly different from 0.05 (most are over 0.5, see Appendix Table C1). Some parameter combinations give rise to  $P_A$  close to 0.05. Although the rates of underestimation under these conditions are also significantly different from 0.8, most of them lie within a reasonable proximity to the empirical level (see Appendix Table C1). Therefore, the parameter combinations that give rise to  $P_A$  close to 0.05 are used for further analyses.

Linear regressions are run on the rate of underestimation, the rate of inconsistency, and the rate of WWOH actions using the following model <sup>1</sup>:

$$y = a + b_1 I + b_2 p_r + b_3 p_c + b_4 \epsilon_{mean} + b_5 \alpha + b_6 s_{mean} + \\ + b_7 I \cdot p_r + b_8 I \cdot p_c + b_9 I \cdot \epsilon_{mean} + b_{10} I \cdot \alpha + b_{11} I \cdot s_{mean} + \gamma, \quad (9)$$

where  $\gamma$  denotes random errors. Interaction terms are included to detect potential dependency of the intervention effect on other parameter values.

The results reveal that all else equal, the intervention does not have a statistically significant effect on the rate of underestimation, be it implemented once, twice, or three times (Table 3.2). The intervention does have a significant effect on the rate of inconsistency regardless of the number

---

<sup>1</sup>Subscripts are omitted for readability.  $c_{mean}$  and  $r_{mean}$  are not included as predictors since the former is perfectly correlated with  $p_c$  and the latter lacks variation in the data.



of times it is implemented, and its effect significantly depends on  $p_c$ ,  $\epsilon_{mean}$ ,  $\alpha$ , and  $s_{mean}$ . The intervention has a significant effect on the rate of WWOH actions only when it is implemented twice or three times, and this effect also significantly depends on  $p_c$ ,  $\epsilon_{mean}$ ,  $\alpha$ , and  $s_{mean}$ .

**Table 3: Effects of interventions on pluralistic ignorance and WWOH action at the final time step**

	$P_U$			$P_I$			$P_A$		
	$\beta$	$t$	$p$	$\beta$	$t$	$p$	$\beta$	$t$	$p$
(Intercept)	1.08*** (0.02)	52.87	<.001	0.83*** (0.02)	36.40	<.001	0.07*** (0.01)	4.86	<.001
$I_1$	0.04 (0.03)	1.37	.172	0.07** (0.03)	2.24	.025	-0.03 (0.02)	-1.19	.234
$I_2$	0.002 (0.03)	0.05	.957	0.18*** (0.03)	5.48	<.001	-0.16*** (0.02)	-7.75	<.001
$I_3$	-0.01 (0.03)	-0.28	.778	0.27*** (0.03)	8.33	<.001	-0.28*** (0.02)	-13.39	<.001
$p_r$	-0.03 (0.02)	-1.12	0.261	0.05** (0.02)	2.04	.041	-0.04*** (0.02)	-2.63	.009
$p_c$	-0.16*** (0.03)	-5.34	<.001	-0.04 (0.03)	-1.27	.204	-0.05** (0.02)	-2.42	.015
$\epsilon_{mean}$	-0.06*** (0.02)	-3.47	<.001	0.03 (0.02)	1.35	.178	0.003 (0.01)	0.29	.774
$\alpha$	-0.09*** (0.02)	-5.78	<.001	-0.03 (0.02)	-1.58	.115	0.002 (0.01)	0.13	.893
$s_{mean}$	0.01 (0.02)	0.29	.773	<.001 (0.02)	<.001	.999	0.02 (0.01)	1.38	.169
$p_r \times I_1$	-0.03 (0.03)	-0.97	.330	-0.02 (0.04)	-0.57	.570	-0.01 (0.02)	-0.25	.805
$p_r \times I_2$	0.01 (0.03)	0.21	.833	-0.05 (0.04)	-1.55	.122	0.03 (0.02)	1.50	.133
$p_r \times I_3$	0.03 (0.03)	0.84	.400	-0.01 (0.04)	-0.16	.873	-0.01 (0.02)	-0.21	.835
$p_c \times I_1$	-0.05 (0.04)	-1.26	.207	-0.12** (0.05)	-2.54	.011	0.03 (0.03)	1.05	.292
$p_c \times I_2$	0.02 (0.04)	0.47	.636	-0.16*** (0.05)	-3.37	<.001	0.13*** (0.03)	4.34	<.001
$p_c \times I_3$	0.01 (0.04)	0.19	.850	-0.21*** (0.05)	-4.60	<.001	0.20*** (0.03)	6.67	<.001
$\epsilon_{mean} \times I_1$	-0.004 (0.02)	-0.15	.879	-0.04 (0.03)	-1.33	.185	0.02 (0.02)	1.34	.181
$\epsilon_{mean} \times I_2$	-0.01 (0.02)	-0.32	.748	-0.11*** (0.03)	-3.98	<.001	0.11*** (0.02)	6.15	<.001

**Table 3: Effects of interventions on pluralistic ignorance and WWOH action at the final time step (continued)**

	$P_U$			$P_I$			$P_A$		
	$\beta$	$t$	$p$	$\beta$	$t$	$p$	$\beta$	$t$	$p$
$\epsilon_{mean} \times I_3$	-0.01 (0.02)	-0.54	.586	-0.20*** (0.03)	-7.62	<.001	0.21*** (0.02)	12.09	<.001
$\alpha \times I_1$	-0.01 (0.02)	-0.57	.569	-0.04 (0.03)	-1.57	.116	0.03* (0.02)	1.95	.052
$\alpha \times I_2$	0.03 (0.02)	1.48	.139	-0.11*** (0.03)	-4.18	<.001	0.15*** (0.02)	9.02	<.001
$\alpha \times I_3$	-0.01 (0.02)	-0.36	.722	-0.28*** (0.03)	-11.04	<.001	0.32*** (0.02)	19.43	<.001
$s_{mean} \times I_1$	-0.02 (0.02)	-0.77	.443	-0.04 (0.03)	-1.56	.118	0.02 (0.02)	0.99	.321
$s_{mean} \times I_2$	-0.02 (0.02)	-0.98	.329	-0.11*** (0.03)	-4.27	<.001	0.10*** (0.02)	5.85	<.001
$s_{mean} \times I_3$	0.001 (0.02)	0.05	.960	-0.18*** (0.03)	-6.96	<.001	0.20*** (0.02)	12.00	<.001

*Note.*  $P_U$ ,  $P_I$ ,  $P_A$  denote the rate of underestimation, the rate of inconsistency, and the rate of WWOH actions, respectively.  $I_1$ ,  $I_2$ , and  $I_3$  denote whether the intervention is implemented once, twice, or three times (vs. no intervention), respectively. Standard errors are in parentheses. \* $p < 0.1$ , \*\* $p < 0.05$ , \*\*\* $p < 0.01$ .

To show the above results more clearly, the effects of the intervention under different combinations of parameter values are plotted (Figure 3 and 4). As Figure 3 shows, the intervention has an estimated effect decreasing the rate of inconsistency under most parameter combinations. The magnitude of the decreasing effect increases with the probability of communication ( $p_c$ ), mean perceived credibility ( $s_{mean}$ ), convergence parameter ( $\alpha$ ), and mean bounded confidence threshold ( $\epsilon_{mean}$ ). The magnitude also increases with the number of times the intervention is implemented, and the increase is especially pronounced when the times increase from twice to three times. Likewise, the intervention has an estimated effect increasing the rate of WWOH actions under most parameter combinations. The magnitude of the increasing effect increases with the probability of communication ( $p_c$ ), mean perceived credibility ( $s_{mean}$ ), convergence parameter ( $\alpha$ ), and mean bounded confidence threshold ( $\epsilon_{mean}$ ) (Figure 4). The magnitude increases with the number of times the intervention is implemented. When the intervention is implemented once, its effect remains close to 0 under all parameter combinations.

For the sake of illustrating the process of how the intervention affects the outcomes, one set of parameter combination that sustains PI and WWOH at the empirical level ( $p_r = 0.5$ ,  $p_c = 0.3$ ,  $\epsilon_{mean} = 0.5$ ,  $\alpha = 0.3$ ,  $c_{mean} = 0.1$ ,  $r_{mean} = 0.1$ ) is selected and the changes of PI and WWOH actions are plotted over time. As Figure 5 shows, the intervention primarily changes the rate of inconsistency and WWOH actions instead of the rate of underestimation, which is aligned with the results from the regression analysis. In general, the rate of inconsistency drops and the rate of WWOH actions surges immediately after the intervention is implemented. However, this effect is gradually dampened over time, that is, the rate of inconsistency tends to increase and the rate of WWOH actions tend to decrease, as the simulation continues after the intervention. When the mean perceived credibility is held equal, the intervention effect seems to be enhanced every time it is implemented (see the facet columns). When the implementation number is held equal, the intervention effect seems to get more pronounced when the mean perceived credibility increases, especially when the intervention is implemented twice or three times (see the facet rows).

## 4 Discussion (1000-1500)

By building an agent-based model (ABM) of pluralistic ignorance (PI) regarding women working outside the home (WWOH) among men in Saudi Arabia, the present study seeks to explore the factors associated with the persistence of PI regarding WWOH. It also aims to investigate whether social norm interventions can reduce PI and encourage WWOH behaviours when PI regarding WWOH persists at the empirical level, and if so, under what conditions. The results reveal that among all parameters included in the model, the probability of communication with neighbours ( $p_c$ ) and mean resistance to norm ( $r_{mean}$ ) significantly affect PI and WWOH actions. Higher probability of communication with neighbours and higher mean resistance to norm decreases PI (defined as the rate of inconsistency) and increases the rate of WWOH actions. PI (defined as the rate of inconsistency) persists if the probability of communication with neighbours is low or medium ( $p_c = 0.1$  or  $p_c = 0.3$ ) and the mean resistance to norm is low ( $r_{mean} = 0.1$ ). When PI regarding WWOH persists at the level identified by the empirical study (**Bursztyn et al., 2020**), a social norm intervention does not significantly change the rate of underestimation but decreases the rate of inconsistency and increases the rate of WWOH actions under most parameter combinations.

The magnitude of the intervention effects positively depend on the probability of communication ( $p_c$ ), mean bounded confidence threshold ( $\epsilon_{mean}$ ), convergence parameter ( $\alpha$ ), mean perceived credibility ( $s_{mean}$ ), and the number of times the intervention is implemented.

The results regarding the factors associated with persistent PI does not support a conjecture proposed by the empirical study about PI among men in Saudi Arabia. Bursztyn and colleagues (2020) proposes that a lack of communication may be the reason why a misperception of social norms can become an equilibrium in a population, which explains the PI they observed among Saudi men. The present study finds that all else equal, the probability of communication is not significantly associated with the rate of underestimation, which suggests that communication may not be the primary factor contributing to the persistence of norm misperception. However, it is negatively associated with the rate of inconsistency and positively associated with the rate of WWOH actions. This suggests that promoting communication can reduce the proportion of individuals with inconsistent beliefs and actions as well as encouraging WWOH actions in the population. This further suggests that even though the literature has used PI to refer to both the rate of underestimation and the rate of inconsistency, these two definitions may not coincide in terms of what they refer to. If they did, what affects one would also affects the other in a highly similar manner, but this is not what the present study finds <sup>2</sup>. Therefore, future studies should pay close attention to the operational definition of the broad concept of PI since different ones cannot be used interchangeably.

The mean resistance to norm ( $r_{mean}$ ) being significantly associated with two of the outcomes suggests that aspects of individuals can influence group-level phenomena. The extent to which each individual is influenced by social norms on average in making decisions about WWOH actions can influence the overall the rate of inconsistency and the rate of WWOH actions. This is not surprising given the decision making rules programmed in the model: agents' WWOH action depends on their preference, which is a weighted average of their belief and norm perception. When an agent's belief is high (representing higher support for WWOH), the higher the weight given to its belief (resistance to norm  $r_i$ ), the greater extent to which its supporting belief will be reflected in preference and therefore action. It is important to notice that the above conclusion may not gener-

---

<sup>2</sup>The finding that the social norm intervention does not significantly affect the rate of underestimation but decreases the rate of inconsistency further speaks to this point.

alise when the weighted average assumption no longer applies. Past research has also adopted an exponential function to model the relation between norm perception and action (**e.g., de Mooij, 2022**) or more complex functions the solutions to which (and therefore the updating “rule” for belief, norm perception, and action) require mathematical optimisation (**Gavrilets, 2021; Tverskoi et al., 2023**). Further investigations assuming different functional relations are needed to corroborate the finding of the present study.

If individuals’ resistance to norm does affect the overall rate of inconsistent belief and action and WWOH actions regardless of the functional assumptions, this conclusion can be taken as generally true and may be used to guide empirical research. For example, the concept of resistance to norm coincides with the concept of susceptibility to social norms that has been discussed in existing studies of cultural differences in human behaviours (**e.g., Savani et al., 2015; Muthukrishna & Schaller, 2020**). Findings of the present research suggests that lower resistance to norm, that is, higher susceptibility to norm, are more likely to lead to PI (defined as the rate of inconsistency). If, as suggested by the findings of existing research, there are cultural differences in individuals’ susceptibility to social norms (**Savani et al., 2015**), we would also expect to see cultural differences in PI regarding a certain issue. This hypothesis has not, to my best knowledge, been tested by existing study and can be a possible direction to future research.

As for the effects of social norm intervention, the intervention decreases the rate of inconsistency and increases the rate of WWOH actions, the magnitude of which depends on the implementation times as well as other parameter values. First of all, as expected, the more times the intervention is implemented, the more pronounced the intervention effect is. Based on the process through which the intervention impacts the outcome variables, this may be because the intervention at a later time enhances the effect of a previous implementation before its effect is completely “washed out”. Due to this enhancement, even though the effect is still gradually dampened over time after the final implementation, the multiple times of interventions together lead to statistically significant effects on the two outcome variables.

Second, credibility of the information source also has a positive effect on the intervention effects. This is not surprising given that the magnitude of the influence of the intervention on agents’ norm perception is programmed to be proportional to its perceived credibility. All else equal, the higher

the mean perceived credibility is in the population, the more influence the intervention has on agents (if any), and the closer agents' norm perception is to the true norm on average. If the true norm is that most agents support WWOH, this norm will be more likely to be reflected in WWOH actions when the credibility of the intervention is high.

Moreover, the intervention effect also depends on the two parameters that control the updating of agents' belief about WWOH, the mean bounded confidence threshold ( $\epsilon_{mean}$ ) and the convergence parameter ( $\alpha$ ). This is because the intervention via summary information is conceptualised as a persuasive message and is programmed to impact the norm perception in a way similar to how agents update their belief about WWOH based on those of others. For each agent, the higher the bounded confidence thresholds, the less stringent the requirement on the closeness between its norm perception and the true norm so that the agent will modify its own perception. The higher the convergence parameter, the closer the agent's modified norm perception will be to the true norm, given it does modify its perception (i.e., when the absolute difference between its own perception and true norm does not exceed the bounded confidence threshold).

An acute reader might have already realised that the explanations of the model results have been largely confined to the particular assumptions adopted by the model, especially those on how the updating depends on the parameters. This points to an important limitation of the present study, that is, the conclusions drawn from the model hold true if the model assumptions hold true. This is a limitation in two senses. First, the findings of the current model alone are not sufficient to support any general claim regarding PI about WWOH. They need to rely on models with different assumptions for corroboration. The previous discussion on the finding about the relation between individuals' resistance to norm and PI is written out of this consideration. Second, the findings are also not sufficient to explain the PI regarding WWOH in Saudi Arabia or predict the effect of social norm intervention that utilises summary information. This is because the present model is an extremely simplified one with model assumptions mostly drawn from existing modelling studies. Further improvements can be made to adjust the assumptions using empirical studies on how people in Saudi Arabia reason about beliefs and norm perceptions as well as how they make decisions on actions. Additional processes can also be added to make the model more realistic, the most important of which, in my opinion, is the process for a job market. This is because actual

decisions on WWOH actions presumably not only depend on people's preferences but also on whether jobs outside the home are available to women in Saudi Arabia.

In addition to tackling the previously mentioned limitations, future studies can also explore the following directions. First, further development of the model can calibrate the parameter setting using empirical data on WWOH in Saudi Arabia such as Arab Barometer (**Carrella et al., 2020; Grazzini & Richiardi, 2015**). In choosing parameter combinations that give rise to PI and WWOH actions close to the empirical data, the present study investigates a coarse set of parameter settings and can only provide a rough approximation to the summary statistics. Systematic calibration techniques can help find parameter settings that minimise the distance between the empirical data and simulated data, which, together with better model assumptions, may provide more realistic simulation environment to test policy interventions. Second, due to time limit, the present study is only able to test one type of social norm intervention and vary the number of times it is implemented. Future studies can benefit from exploring and modelling a diverse kit of existing social norm interventions. For example, an intervention technique that is increasingly being explored is to change the public actions of salient individuals in order to create large-scale norm change (**Paluck et al., 2012; Paluck et al., 2016**). This method has been shown to be effective in reducing harassment and conflict behaviours in schools (**Paluck et al., 2012; Paluck et al., 2016**) and can be easily programmed in an ABM so that its effect on PI regarding WWOH can be tested. Different social norm interventions can be compared and the most effective intervention strategy can be selected to guide policy making (**Bailey et al., 2019**).

## 5 Conclusion

## Reference



## Appendix: The Derivation of Equation 3 and 4

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

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

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

$$f(E_i | N_i) \propto \exp\left(-\frac{(E_i - N_i)^2}{2c_i^2}\right).$$

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

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

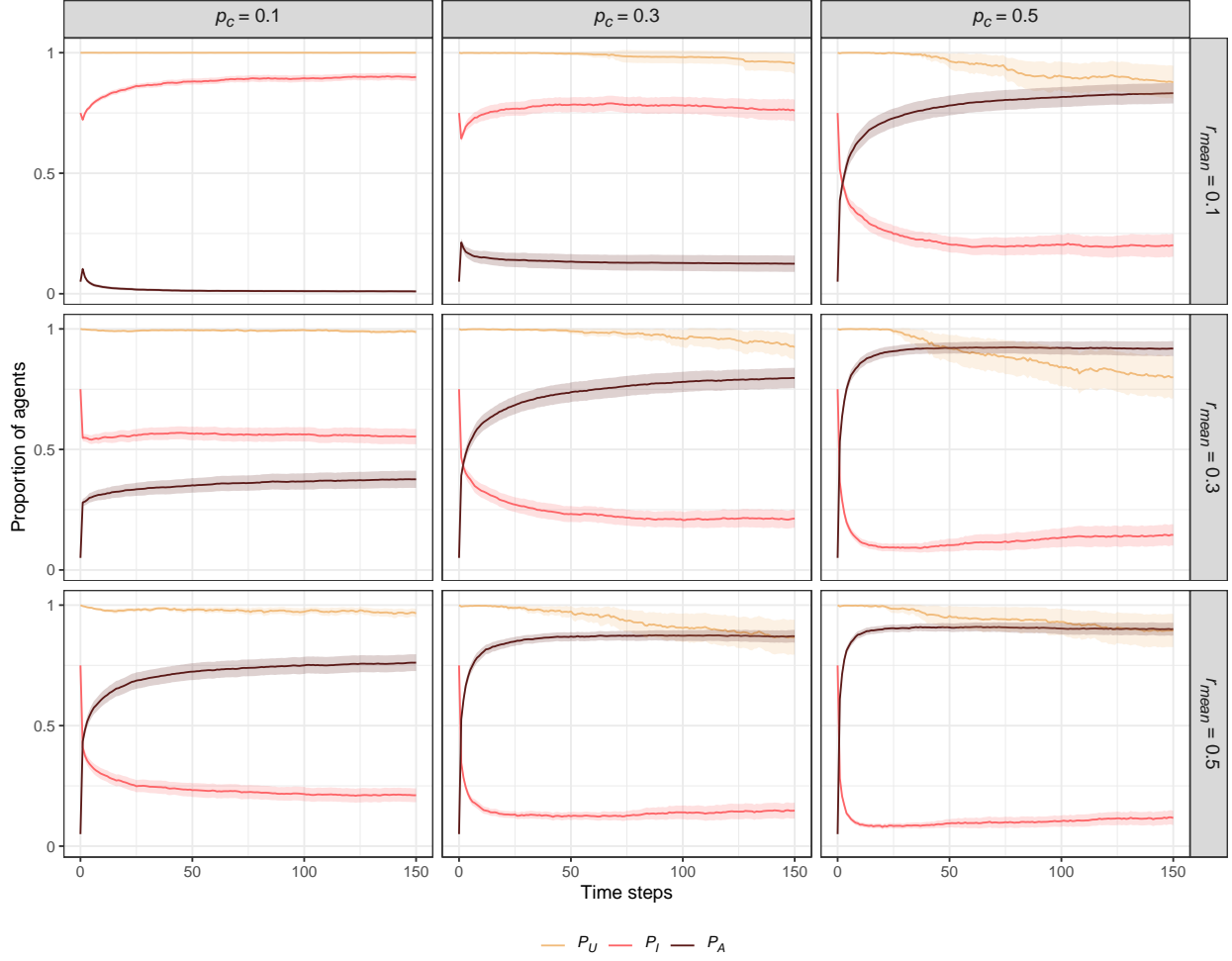
where  $C$  is a normalising constant. Since the second term in the exponent does not involve the random variable  $N_i$ , we can incorporate it into the constant and rewrite the expression as

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

where

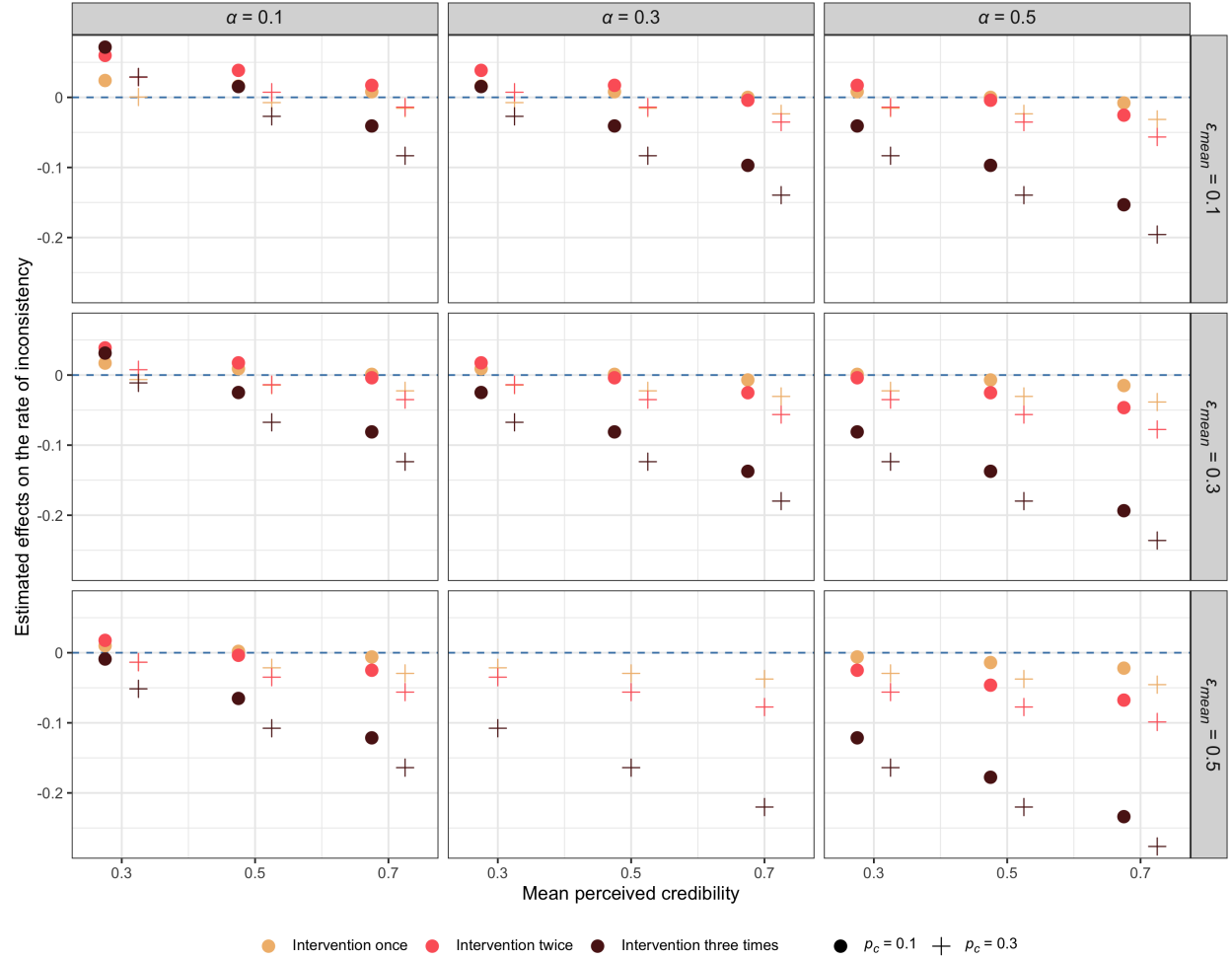
$$\mu'_{N_i} = \frac{E_i \sigma_{N_i}^2 + \mu_{N_i} c_i^2}{\sigma_{N_i}^2 + c_i^2} \quad \text{and} \quad \sigma'^2_{N_i} = \frac{\sigma_{N_i}^2 c_i^2}{\sigma_{N_i}^2 + c_i^2}.$$

Substituting  $\frac{c_i^2}{c_i^2 + \sigma_{N_i}^2}$  with  $\lambda$  gives the equations in the main text.

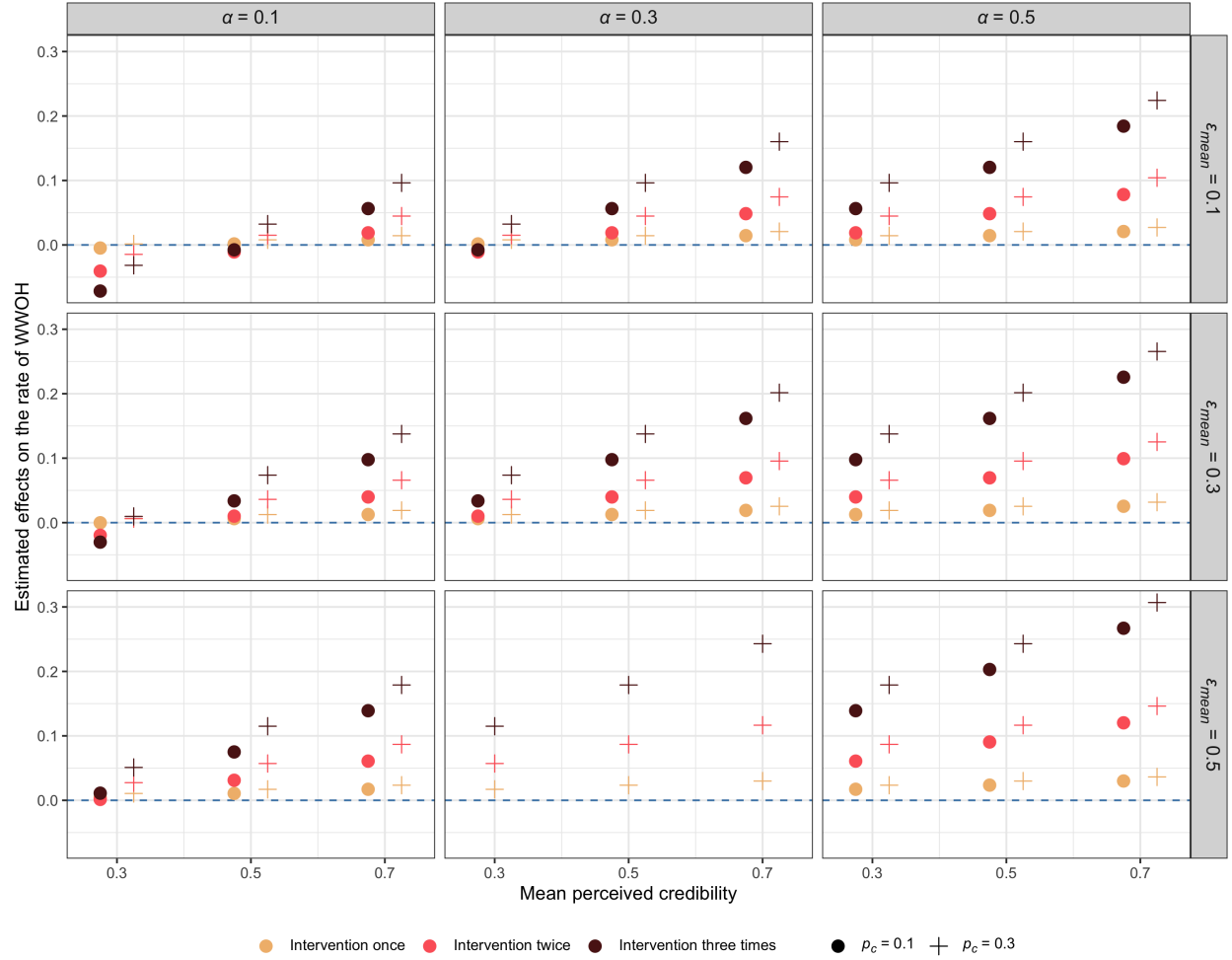


**Figure 2: Pluralistic ignorance and WWOH action over time when no intervention is implemented.**

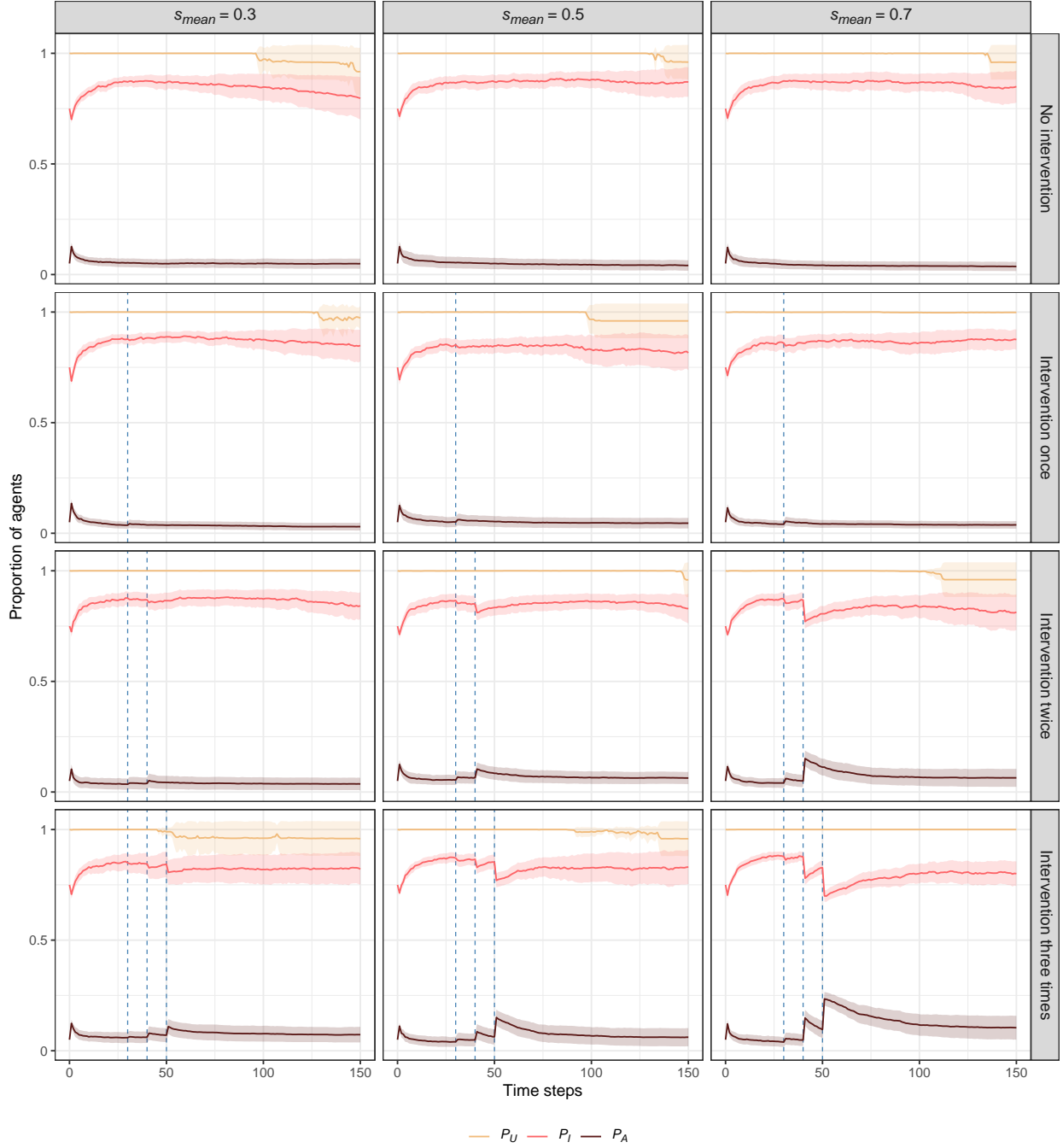
$P_U$ ,  $P_I$ ,  $P_A$  denote the rate of underestimation, the rate of inconsistency, and the rate of WWOH actions, respectively. Central lines are average values of simulations with the same parameter setting, and shaded areas are the 95% confidence intervals. Probability of communication with neighbours  $p_c$  is shown on facet columns, and the mean resistance to norm  $r_{mean}$  is shown on facet rows. Parameters not shown on the plot are set at the following values:  $p_r = 0.5$ ,  $\epsilon_{mean} = 0.5$ ,  $\alpha = 0.3$ ,  $c_{mean} = 0.3$ .



**Figure 3: Estimated effects of the intervention on the rate of inconsistency at the final time step.** Point estimates of the effects of the intervention are plotted under different combinations of parameter values. Blue dashed lines indicate 0 effect. Convergence parameter  $\alpha$  is shown on facet columns, and the mean bounded confidence threshold  $\epsilon_{mean}$  is shown on facet rows. Parameters not shown on the plot are set at the following value:  $p_r = 0.5$ .



**Figure 4: Estimated effects of the intervention on the rate of WWOH actions at the final time step.** Point estimates of the effects of the intervention are plotted under different combinations of parameter values. Blue dashed lines indicate 0 effect. Convergence parameter  $\alpha$  is shown on facet columns, and the mean bounded confidence threshold  $\epsilon_{mean}$  is shown on facet rows. Parameters not shown on the plot are set at the following value:  $p_r = 0.5$ .



**Figure 5: Effects of the intervention on pluralistic ignorance and WWOH action over time.**  $P_U$ ,  $P_I$ ,  $P_A$  denote the rate of underestimation, the rate of inconsistency, and the rate of WWOH actions, respectively. Central lines are average values of simulations with the same parameter setting, and shaded areas are the 95% confidence intervals. Blue dashed lines indicate the time step when the intervention is implemented. Mean perceived credibility of the information source  $s_{mean}$  is shown on facet columns, and the intervention type is shown on facet rows. Parameters not shown on the plot are set at the following values:  $p_r = 0.5$ ,  $p_c = 0.3$ ,  $\epsilon_{mean} = 0.5$ ,  $\alpha = 0.3$ ,  $c_{mean} = 0.1$ ,  $r_{mean} = 0.1$ .