Disclaimer: this is a working paper. Do not cite without the author's permission.

Emergence of Pluralistic Ignorance of Social Norms in the Absence of Enforcement – An Agent-Based Model

Linda Urselmans

University of Essex

30 July 2016

Working Paper

Abstract

Pluralistic ignorance occurs when the majority of a social group adopts a social norm that no single individual would adopt, if given a choice in the absence of information about other peoples' norms. It can occur when people incorrectly believe that the other members of the group are in favour of a given norm, and thus follow suit. How do these situations arise? I propose that emergence of PI requires no external norm enforcement and may arise because people want to feel comfortable with those surrounding them, but also make inaccurate inferences about neighbours' private beliefs on the basis of their behaviour. The results support the hypothesis. They highlight the importance of density of social interaction and whether agents intelligently seek more comfortable surroundings when they move.

Introduction

Social norms are upheld by behaviour and assumptions about behaviour rather than formal norms enshrined in a legal code. They are different from formal norms in that they entail the presumption of practices. These presumptions may be wrong (Brennan et al., 2013). Pluralistic ignorance (PI) results when people wrongly assume that the behaviour they observe in others is a reflection of their attitudes, and thus an endorsement of the norm (Bjerring et al., 2014). The study of PI dates back to Katz and Allport (1931) who are widely credited with first developing the notion.

A commonly used example of PI is the case of drinking by university students. In their first week at university (and throughout their university years), many students drink excessive amounts of alcohol, even though they would rather not do so given the choice. As every student observes the others drinking obsessively, they assume that everyone else enjoys the excessive drinking behaviour. As a result, a lot of students continue to drink heavily as they seek to avoid standing out from the group (Bjerring, Hansen, and Pedersen, 2014). Typically, a small group of students genuinely agrees with drinking obsessively, and their public behaviour is consistent with their private beliefs. But to the outsider, it cannot be established whether students' public behaviour is a reflection of a private attitude or, rather, a reflection of norm conformity in spite of private attitudes. So out of a group of 20, it may be that five of them really do prefer heavy drinking, and the remaining 15 find themselves in a situation of PI. The five 'genuine drinkers' contribute to PI despite not experiencing a discrepancy between their public behaviour and private attitudes.

Typically, social norms are modelled as spreading or emerging by means of contagion, imitation, learning, and coercion (Beheshti and Sukthankar 2014; Centola et al. 2005; Wang et al. 2013). However, the occurrence of PI presents a puzzle for theories of social norms, because it is unclear how it arises in the first place, since all these processes presuppose that some already conform to the norm. It is relatively easy to see, having already arisen, how PI can erode. In the university drinking situation, PI could collapse if everyone drunk less or even shared their private beliefs. Wang et al. (2013) introduce an opinion-dynamics model adapted to incorporate PI. In their model, a single agent rejecting the status of PI can lead to a complete change in opinions of the entire group, if the agent's opinion is firm and its neighbours are unsure about their own attitude (2013, p. 247). However, this model presupposes the norm was initially in existence. The central contribution of this paper is to show how PI can emerge in the first place, as agents move in social space to increase how comfortable they feel with their neighbours.

Agent-based modelling enables me to look at the emergence of PI from a bottom-up perspective. By giving the individuals rules of behaviour on a micro level, the model can give rise to macro-level patterns of behaviour that illustrate how emergence occurs, i.e. which preconditions are needed and what rules can recreate the stylised facts that have been established in the literature. In my model rather than switching private attitudes or bending to other kinds of pressure such as global opinion or the strength of others' attitudes, agents solely seek to minimise the difference between their surrounding neighbours and themselves. They do so by changing their social location. PI can emerge even though no agent actively strives to make a norm dominant. Rather it arises because agents try to reduce internal strain that comes from conforming to a norm they do not privately agree with combined with the desire to be with others with similar public behaviour. A specific prediction from my model is that it is more common if society is "dense" so there are few social locations that do not have close "social neighbours". For example, this implies it might be less common in anomic big cities than in small communities. I show also show that it is less common if people “move” (i.e. re-align their social relations) in an intelligent way. A virtue of the model is that it contains so few parameters, yet they are sufficient to give rise to PI. Alternative (agent-based) models of PI usually feature notions of “true-believers” and varying degrees of pressure exerted by different kinds of agents (see for example Centola et al, 2005). I show that pressure is not necessary for PI to arise.

PI and its many faces

One intuitively appealing, often-cited example of PI is Hans Christian Anderson’s tale of the Emperor’s new clothes. The respected and acclaimed emperor is fooled by a group of rogues into believing that their (in fact) non-existent new robe for the emperor is actually real. The rogues persuade the emperor that stupid people can’t see the robe- and thus the emperor, afraid to admit that he can’t see the robe, pretends that he can see it and thus, that the robe exists. The citizens share the same fear and pretend to be amazed by the garments- until an innocent child comes along and laughs at the naked emperor, and the spell of PI is broken.

The fable is appealing in that it is very easy to imagine situations resembling the emperor’s fate, for example the highly acclaimed professor in an academic setting who is perceived to have brilliant insights which, in private, many scholars find to be lacking (Centola et al, 2005). As strong as the intuition is, as is the case with the various others examples often presented (such as the University drinking) it is also vague. Not all examples share the same defining attributes for PI. As Centola et al (2005) note, PI theory has several shortcomings, including failure to make explicit assumptions about the group element of PI.

Because PI is a group phenomenon in that it is analytically distinct from the psychological process of deriving attitudes from behaviour (O’Gorman 1986, Moy 2008), the conditions that groups have to fulfil to be pluralistically ignorant warrant further attention. There is no single agreed-on definition of PI in the literature (Centola et al, 2005). What follows is a brief overview of PI. Broadly speaking, definitions of PI can be categorized as either focussing on the collective (the group and the environment in which PI occurs) or those focussing on the individual, trying to explain the psychological reasons for individual failings.

Since Katz and Allport (1931), definition of PI, the concept has been widened, leading to many different nuanced interpretations. Brennan and Goodin (2013) cite Allport saying that PI is “[…] a situation where a majority of group members privately reject a norm, but assume (incorrectly) that most others accept it”. This definition implies that a precondition for PI is that a majority of a group has to suffer from the cognitive error. This detail is not present in most other definitions. Moy (2008) mentions that PI has come to include underestimation of majority opinion, but also minority opinion, stating that PI refers to “perceptual inaccuracies of the collective, by the collective.” (p. 164). The notion of majority and minority here refers to the public opinion collective, not how many people actually have to be wrong in order for PI to occur (which is alluded to with “by the collective”, indicating at least a majority). Groeber and Rauhut (2010) leave out any condition on the group within which PI occurs and write: “Social psychologists primarily studied PI, which explains herding behaviour by the agents’ false assumption that most others will approve of what the majority publicly complies with.” (2010, 2).

Lastly, Van Boven (2000) cites (Miller & McFarland, 1987, 1991; O’Gorman, 1986; Prentice & Miller, 1996): “[p]luralistic ignorance occurs when people overestimate a group’s endorsement of an attitude or norm when, in fact, the attitude or norm enjoys little support among group members". This definition focuses on the collective attributes of PI, but makes no implicit mention of any majority, minority or plurality endorsement of norms.

Definitions focussing on the individual include for example "[p]luralistic ignorance describes the "belief that one's private attitudes and judgments are different from those of others, even though one's public behavior is identical" (Willer et al, 2009). Both definitions stress that people have to have a distinct private attitude and a public display of behaviour, and that it often involves wrong assumption about others. This interpretation is often attributed to Miller & McFarland (1991), stating "[p]luralistic ignorance is a psychological state characterized by the belief that one's private attitudes and judgements are different from those of others, even though one's public behaviour is identical […]".

Perhaps the most specific definition was presented by O'Gorman (1986) stating that "[p]luralistic ignorance refers to erroneous cognitive beliefs shared by two or more individuals about the ideas, feelings, and action of others." According to O'Gorman, any pair of individuals (sharing their cognitive error) within any group would constitute PI. This is not compatible with Brennan & Goodin's notion that a majority of a group has to share the cognitive error, unless a pair of people constitutes a majority in a group of three. The notion of group size is another aspect of PI that varies between definitions and research areas. From a public opinion perspective, the group can encompass a large fraction of society, or society as a whole (see for example Moy, 2008; O'Gorman, 1975; and Kuran, 1995). Other studies mention smaller groups, such as social groups at university (Prentice and Miller, 1993), inmates in prisons (O'Gorman 1986) or a religious community in a localized space (Schank, 1932).

Bjerring et al (2014) offer an in-depth discussion on the variety of PI definitions and conclude that many extant definitions lead to overestimation of PI as they leave too much theoretical room for manoeuvre. They offer the following alternative definition of PI, which as far as I know, is the most comprehensive one yet:

> *[…] "PI" refers to a situation where the individual members of a group*
> *(i) all privately believe some proposition P;*
> *(ii) all believe that everyone else believes ¬P;*
> *(iii) all act contrary to their private belief that P (i.e. act as if they believe ¬P);*
> *and where*
> *(iv) all the actions of the others as strong evidence for their private beliefs about P.*

As far as the specification of beliefs is concerned, the definition is certainly more detailed than previous ones. However, the notion of "all" individuals and "all" actions is not justified in conditions (iii) and (iv). It is not necessary for *all* members of a group to act contrary to their beliefs, for that everyone, as stated in (ii), believes that anyway. If truly every single member of a relevant group has to follow the same pattern, these strict conditions are unlikely to ever be met in the real world. It is reasonable to assume that most authors will acknowledge that the absoluteness of all members is in fact more flexible once applied to the real world. That does not mean we should shy away from including that flexibility in any definition of PI.

There remains a debate to be settled before we can model PI as a group phenomenon. The first question is: how big does a group have to be? The implications of any possible answer are wide-ranging. The second question, inextricably linked to the first, is: do people need to have physical or otherwise visible social interaction with others as a prerequisite for PI? If yes, then groups will naturally be relatively small, not larger than the largest number of people a human mind can grasp at a time. It would mean that PI understood by public opinion scholars would be a distinct phenomenon.

Is direct social interaction a requirement for PI?

Bjerring et al argue that situations that are characterized by a “lack of observational interaction among agents in the relevant social group” (2014, p.12) do not count as PI. In other words, if there is no direct observation of other bystanders possible from which assumptions may be drawn, PI is not present. They support this claim by reference to a wealth of examples such as the Emperor’s New Clothes, which all feature direct observation. I disagree. It is true that the bystander effect requires direct observation and that many examples including the Emperors’ new clothes are characterized in this way, but it is not a justification for restricting PI to this narrow set of interactions. When social interaction is not possible, other types of information gathering can take its place. An additional factor to consider is that information might have been collected at an earlier point in time (whether through social interaction or not).

Consider a scenario in which the populace of a country believes that a certain norm is supported by the majority (say, opposing gay marriage) but actually only a minority of true believers follows the norm. According to the definition above, this is only PI if people have direct social interaction with others (say, sharing a pint in a pub) at which point certain relevant comments about gay marriage may be made, signalling to everyone else whether gay marriage is opposed or not. But the social interaction that is relevant for this norm might not be available to people. Perhaps they do not know any gay people, or people with opinions strong enough to signal approval or rejection of the norm. Perhaps they do not venture into social spaces in which any behaviour could be interpreted to relate to the norm of approving gay marriage. Yet it is likely that they will nonetheless have an idea of what others think. They will fall back on other means of information gathering. A collective idea of certain behaviour leads to a collective idea of what that behaviour means, even though the initial behaviour was not personally observed. What is so different about directly observed behaviour that it should constitute a prerequisite, but information about behaviour, attitudes or both, acquired in other ways should not count? Bjerring et al state that “[b]ehavior is typically a good, yet fallible guide to figuring out what people believe.” (p. 13). The reality might be even worse than that.

Consider the following example of an unobservable social norm: It is widely accepted that one should not urinate in public swimming pools. The collective benefit of that norm is that people get to swim in more hygienic conditions, however it means that additional effort is required of individuals by having to exit the water and use (potentially unhygienic) toilets. But breaking the norm cannot be observed, assuming norm-defectors are wise enough to conduct their business underwater. People who are not observed to use the provided toilets provided may not necessarily have urinated in the pool. Perhaps they did not need to relieve themselves at all. The ambiguous nature of observed behaviour makes it virtually impossible to determine whether the norm is actually widely followed. And children that are not potty-trained yet, who can be assumed to urinate underwater, cannot all be assumed to actually have urinated. Is the norm upheld, or not? And crucially, what leads people to believe that it is? They might assume that everyone is sufficiently disgusted by the idea of urine-spoilt pools that they will all adhere to the norm; that they find the mere act of urinating underwater unappealing; or that surely no situation is urgent enough having to relieve oneself at once. But the social interaction that people have at the swimming pool is mostly useless regarding the norm. Short of using scientific measurements of urine concentrations in the water, there is no straightforward way to determine norm rejection. If everyone at the swimming pool believed that the norm is upheld but actually it isn't, would that amount to PI?

Using Bjerring's definition, all pool-goers believe that one shouldn't urinate in the pool (i), but they also all believe that everyone else thinks the opposite (ii). And everyone acts contrary to their belief, so they all urinate in the water (iii), assuming that what others do is strong evidence for their private beliefs (iv). The final condition states that the behaviour must be *perceived to be* meaningful in regards to the attitude that is of interest. So in this case, do pool-goers derive their conclusions from what they observe others to do, or not? As with many norms and attitudes, doing something does not necessarily mean adherence or agreement, and not doing something certainly doesn't always equal non-adherence or disagreement. Going back to the gay-marriage example, not openly approving it does not mean rejecting it. The behaviour remains ambivalent. Completely meaningless behaviour (in regards to the "true" situation) can be falsely perceived to be meaningful. Miller and McFarland (1991) speak of the ambivalence of the social situation, but the fact of the matter is that most behaviour is ambivalent as well, *even if* it is indicative of true attitudes. A pool-goer might urinate in the pool once, but use the toilet on the second occasion because he or she was out of the water at that time, which would not change the attitude towards the norm. It is not always possible to derive true attitudes from even such seemingly related behaviour. Behaviour is thus a poor proxy for private attitudes. The direct observability precondition carries no additional analytical benefit. Information obtained through say, the TV, may just be as misleading in information content as information derived from direct observation.

Explaining why situations without direct social interactions are not PI, Bjerring mention the notion of invoking "relevant background information" (p.11) about certain stereotypes, which is different from information received through social interaction. In the public-pool scenario, people would derive their conclusions from the stereotypes and information gained in other ways ("people like these would not do this"). But this kind of information can just as well lead people to believe that everyone else believes something different. It is assumed that that background information is somehow a given, it exists prior to the situation of interest and is not subject to the same cognitive errors. I argue that the final assumption is unlikely. There are many things we do not know of others, yet we assume some social truth. Assuming someone doesn't know what their colleagues think about gay marriage, they will be likely still to make an assumption based on observed behaviour relevant to other topics. Thus, personal interaction is important and will likely to have played a role at some stage, but it is not necessary for every situation to be entirely derived from it.

If people believed that gay marriage was widely accepted, but actually is widely opposed, they are technically bystanders: only this time they do not physically observe others, but rather use other means of information gathering such as consuming media or build on previous experience (which can be based on observations of behaviour not relevant to the case). PI happens because people make assumptions based on inaccurate information, but the source of that inaccurate information can vary.

By definition PI cannot occur in any scenario with perfect information. Any discrepancy between public behaviour and private attitudes would be the result of an informed choice, not an uncertain assumption. At the public swimming pool, if everyone told everyone else what they think and what they did, there is no room left for assumptions. Norm persistence is not dependent on what members of the group do but rather what they *presume* other group members do. Norms are sustained so long as people reveal their private attitudes only to a small number of people (Kitts, 2003), but it has been shown that PI can persist even when there is widespread knowledge that most people in fact, misrepresent their private attitudes.

Even if we accept the strict precondition of direct observability, it raises the question of whether one must observe absolutely everyone else in a relevant group, or just a fraction: if social interaction is a prerequisite, PI cannot occur in groups larger than the maximum number of people that can be observed simultaneously at any point in time. Widening the time-span, it would include the sum of all people that were observed to do something relevant to the private attribute in question. Images of others in day-to-day life are utilized using various sources of knowledge (O'Gorman). There is no convincing argument as to why knowledge of known behaviour must *only* be obtained through direct observation. O'Gorman posits that the "visible social milieu" of individuals and the "more distant and less visible social world of which that milieu is part" are both relevant to PI, which is perhaps an analytically less pure but more realistic assumption. Therefore, this paper rejects the assumption of direct social observability. Instead, social interaction information can be obtained directly as well as indirectly. In addition, this information can be retrieved at a later point in time, as a time-delayed social interaction.

Consider the bystander scenario that Miller and McFarland (1991) as an illustration of PI: an accident, such as a car crash, occurs and people arrive at the scene. No one has sufficient information to determine whether or not the scene is an emergency and warrants further help on their behalf. It is embarrassing to overreact and thus safer to err on the side of caution and show composure, even if that leads to a bystander situation which is ultimately not desired by any participant. Whether or not a person truly believes that the situation is an emergency is ultimately not relevant. Imagine that first person to arrive at the scene is convinced that there is not an emergency. The next person to arrive will use social comparison and investigate what the first person is doing, and then conclude that there is no emergency because the first person is not helping, and not signalling any distress. But the same result would occur if the first bystander was not a true believer: the first bystander knows the norm of composure and acts accordingly. PI is not necessarily dependent on dispersal of information through observation.

PI as a group phenomenon

The previous discussion revolved around individuals and their perception of others, but not how many others there must be. As stated previously, PI is not simply one person being wrong about others. The concept captures the self-sustaining system of people feeding off other people's behaviour whilst signalling certain behaviour themselves. No one is a neutral observer to the system but always a participant, willing or not. The absolute notions of "all people in a group" were discarded as too strict to be applicable to real life. A single helping bystander might not convince others that they should help if many others just stand and watch; not everyone at the court must pretend the Emperor's clothes exist, as long as enough others do. So how many people within a group need to share their cognitive error for it to be PI?

Any answer must be assumed in relative terms. It would be unrealistic to set a fixed number, if the group size can vary so much. Going back to the definitions offered in the literature, O'Gorman suggests there has to be at least two people who are wrong in the same manner while Brennan & Goodin indicate there needs to be a majority within a group. Other definitions don't offer any concrete suggestions apart from the aforementioned ultimate condition of everyone.

Out of 30 people, is it sufficient for two people to share the same wrong assumption? So two people pretend the Emperor's clothes exist and assume that everyone else believes it. The remaining 28 do not share that social reality: they assume everyone does *not* believe the clothes exist, and act accordingly: they laugh. It seems counterintuitive to classify that as PI. What about 16 people out of a group of 30? The answer most probably is "it depends." It would, for example, depend on the visibility and validity of those 16 people. But assuming that all members of the groups are equal in all relevant respects, would 16 people persuade the remaining 14?; and more importantly, would it matter: if 16 out of 30 suffered from PI, it already constitutes a majority.

What is clear is that individuals will always tend to assume that there exists their attitude and then that of "all others", people don't tend to differentiate further (Gunther, 2008). It is very difficult to define relative numbers because the contexts of PI differ so widely. Assuming that two people alone constitute PI is viable for certain contexts (such as small groups) but hard to generalize. Thinking of society as a whole and their perception of say, public opinion on issues of health or immigration, having only two (interacting) people in the entire country sharing a wrong belief is so likely to happen that it is analytically meaningless.

The discussion on the preconditions of PI is of direct relevance not just for the theoretical concepts of PI, but also for the agent based model that I present in order to test the conditions for its emergence. In order to investigate whether one condition led to more PI than another, I need to define what a group exhibiting PI is characterized by in the first place. The definition that I offer might help to serve as a starting point towards a complete formal definition that includes the group conditions.

Having defined the group conditions of PI, I shall now describe the model in detail. The aim of the model is to simulate the emergence (and persistence) of PI, testing several conditions for their propensity to generate PI or stifle it.

An alternative definition

For purposes of defining PI, first I need to define what is meant here by a group. PI is seen here as a property of a group, though it is rooted in individual behaviour. The group is characterised partly by the fact that they behave the same way in public. So, first, I require that each member of the group exhibits the same public behaviour. However, I also require that there is group awareness. For my purposes, a set of individuals only form a group if, second, they are *directly or indirectly* aware of each other's public behaviour. For instance such a group may be constituted by a social network. Through this a network a group member may be directly aware of the behaviour of those to whom she has a direct network tie; but she may be indirectly aware of others' public behaviour through longer paths in the network, e.g. i may know about k's behaviour because she is linked to j who is linked to k. One natural interpretation of my agent-based model is that it concerns network ties, but social networks are not the only means whereby awareness can arise, as I have indicated above.

Pl is a relatively trivial social phenomenon is the group only has two members. Moreover, if PI requires a majority requirement in some form, this is ill-defined in a group of two. So, third, I require that the group has at least three members. Finally I require that at least a majority of the members of the group have the same inconsistency between their public behaviour and private beliefs. As I have said, PI is a property that is assigned to a group. It is not assigned to a minority of its members. Imagine a group member who has perfect information about public behaviour and private beliefs of all members. Would she say *the group* suffered from a inconsistency between public and private behaviour if *only a minority* did so? I think she wouldn't: she'd say 'a minority of the group' or 'subgroup' or 'some members'. The majority is a *necessary condition* for her to talk of *the group* suffering from PI, though she might require a higher proportion. So, if we defined PI in the way a group member, herself, might do so, the *at least a majority* requirement is necessary.

The Model

The agent-based model consists of a population of agents positioned on a regular two-dimensional grid. Each node of the grid is either empty (denoted by white space) or contains at most one agent (denoted by either blue or yellow colour). At the beginning of the simulation, agents are placed randomly on the grid. One interpretation of this grid topology is that agents on adjacent squares have direct network ties between each other. Moreover, two agents have an indirect network tie if it is possible to move between their squares in such a way that each square along the path is occupied and moves can only be made horizontally or vertically (but not diagonally) between squares on the path, i.e. by making a rook-wise move.

Each agent has a public behaviour and a private attitude, each of which can be either type A (yellow) or type B (blue), and are randomly initialised at the beginning of the simulation. An agent's public behaviour is directly visible to all its neighbours, whereas its private attitude is known only to itself. PI is not necessarily dependent on dispersal of information from observation. However, my model deals with the important sub-set of cases where this is possible. An agent's private attitude may or may not differ from its public norm. Here I assume that agents cannot switch behaviours or attitudes. In reality agents sometimes do these things. However, I want to show that PI can emerge even in situations where there is no pressure exerted other than the mere presence of others.

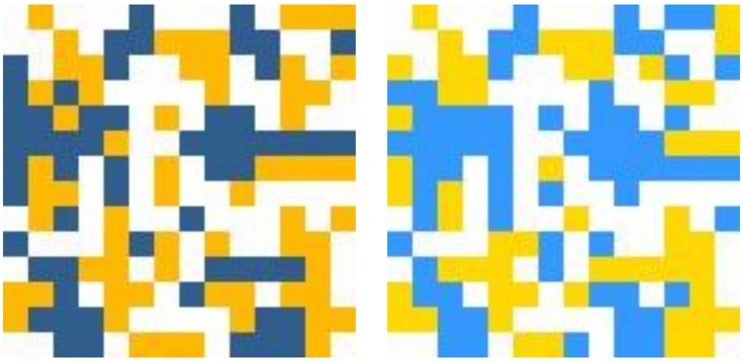


Figure 1: Visualisation of the space of norms: public (left), private (right)

The simulation runs iteratively for a set amount of rounds, or 'ticks'. At each tick, all agents are given the opportunity to move. For any agent the neighbouring positions are the eight cells that are directly contiguous to the cell the agent occupies; thus her neighbourhood is the 3X3 matrix of cells with her cell at the centre. Note that not all cells in an agent's neighbourhood need to be occupied at a given tick of the simulation. At the beginning of each agent's turn to move, it examines its neighbouring nodes and for each potential new location computes a score representing the discrepancy between its private norm and the public-norm of the new neighbourhood. This is defined as the total number of neighbours who have a different public norm from the focal agent's private norm. The agent then chooses a tile which minimises this score, breaking ties randomly. If the current location already minimises the discrepancy, then the agent remains stationary. Underlying this logic is the assumptions that agents want to reduce inconsistency between its private beliefs and the public behaviour of its social neighbourhood – they wish to be "comfortable" in their social position.

The way PI is captured is by examining the clustering of agents. As agents cannot flip attitudes (once assigned private and public attitudes, they will never change either one), they can only escape or enter PI through movement. The size of a cluster (i.e. how many agents it contains) and the numbers of clusters on the map are recorded. Say a group is indirectly connected if it is possible to move between any two members along a path where each cell is occupied and only rook-wise moves are possible. According to my definition above, an instance of PI occurs when there is a group of agents who share a common public behaviour, form a group in that they are directly or indirectly linked and there is at least majority of members whose private attitude are inconsistent with their public behaviour. In the simulation I push this majority requirement to an extreme, so all members of the group must exhibit this inconsistency. This makes it more difficult for PI to emerge. Thus my simulations show that PI can emerge from agents changing their social position *even under a strict definition.* I define the prevalence of PI as the number of such regions over the entire grid occupied by distinct groups satisfying this definition.

A key aspect of this model is the visualisation of the differing private and public norms. It is possible to see either all agents' public norms, or all their private norms. The view can be switched while the simulation is running. It brings a useful intuitive insight into the spatial dimension of PI. Fig. 1 shows a selection of the same part of the map after t = 2500 ticks. It is easy to see that the clusters that appear in the private norm view (right-hand side) are broken up when seeing the agents' public norms. Aside from the statistical analysis of the simulation results, the model seeks to make use of the spatial aspects of agent based modelling and use it to spot patterns.

Table 1 provides an overview of the movement logic described above, whereby agents move if they can move to a more "comfortable" social position. Each agent counts itself as 'same'. The same-count can thus never be lower than 1. The yellow tile indicates each agents default decision, which is 'stay' at 1=same and 0=different, which must assume that the agent has no neighbours.[1] The maximum number of 'sames' is 9, for 'differents' it is 8, as the agent itself is 'same' and thus has a slight bias in favour of like-minded people. The crucial question is what the agent does in cases of the same amount of 'sames' and 'differents', as indicated by the blue tiles. Due to the same-bias, for an equal ratio to occur, there need to be always one different agent more than a same-agent, i.e. for 3-3 situation, the agent needs 2 same-neighbours but 3 different neighbours. In summary, agents tend to move if they are "uncomfortable" because the 'different' count exceeds the 'same' count by at least one.

What are the implications of these rules? When an agent moves in the model, in general this alters all other previous neighbours' calculations of what would be the best behaviour to apply, resulting in a domino-effect. It is because of this non-linear feedback effect that the model exhibits complexity, meaning that simulations are needed to examine *emergence of macro-level patterns* from micro-level behaviour.

[1] If agents are isolated, they will always move, regardless of the movement decision graph shown above.

Table 1 Movement rule decision matrix for agents

	Different Count										
Same Count	**-**	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**-**
	1	**S**	**S**	**M**	**M**	**M**	**M**	**M**	**M**	**M**	
	2	**S**	**S**	**S**	**M**	**M**	**M**	**M**	**M**		
	3	**S**	**S**	**S**	**S**	**M**	**M**	**M**			
	4	**S**	**S**	**S**	**S**	**S**	**M**				
	5	**S**	**S**	**S**	**S**	**S**					
	6	**S**	**S**	**S**	**S**						
	7	**S**	**S**	**S**							
	8	**S**	**S**								
	9	**S**									

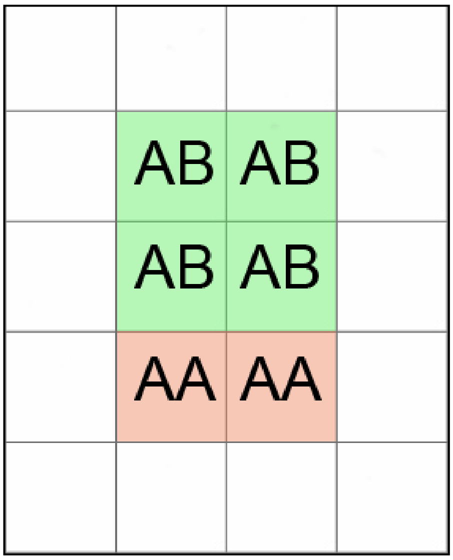
S = Stay
M = Move

Scenarios with and without PI

The following images are designed to further clarify what counts as a group subject to PI and what doesn't. Green fields indicate PI, red fields don't. The first letter is the private norm; the second letter is the public norm. White space denotes empty tiles that agents can move to if they so wish.

#1:

The top left green field shows an agent of private norm A and public norm B. The four AB-agents form a group suffering from PI in which the public norm is B, but the privately held norm is A. The two AA agents do not suffer from PI, because they do not form a *relevant* group: although they are connected, their group only has two members.

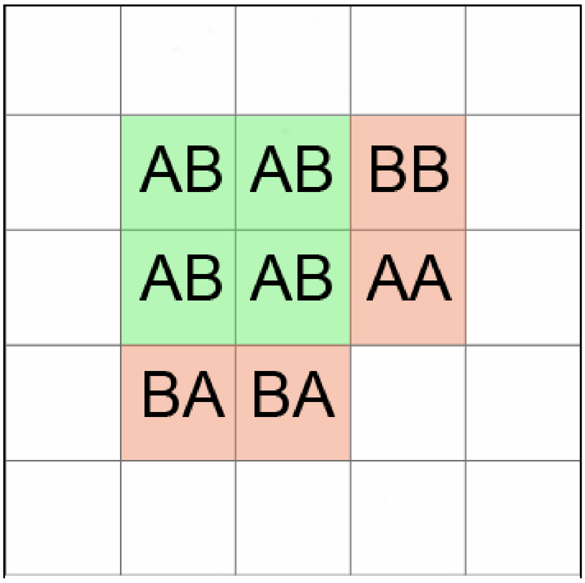


	AB	AB	
	AB	AB	
	AA	AA	

Figure 2 Scenario 1

#2:

The second scenario shows the same cluster of four agents of AB, who again are suffering from PI. Now consider the two bottom BA agents and the AA agent. For two reasons this group does not suffer from PI, although their public behaviour is the same. First agent AA is not connected to the two others by a horizontal or vertical 'rookwise move' pathway. Now suppose, though, that AA's position moved downward one space, so the group was connected. According to the theoretical definition above this group would suffer from PI because they are connected, have the same public behaviour, and a majority of two out of three exhibit an inconsistency with private beliefs. But recall that in the simulations I require that *all* group members suffer from this inconsistency; so for the purposes of the simulation, this group still would not count as an instance of PI.

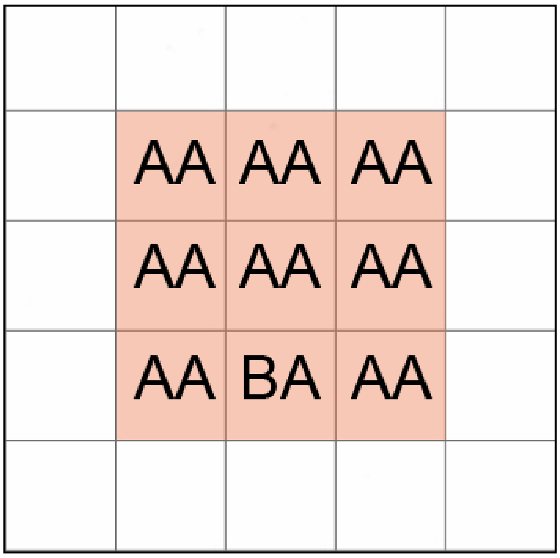


	AB	AB	BB	
	AB	AB	AA	
	BA	BA		

Figure 3 Scenario 2

#3:

The third scenario is a straightforward case where there is no PI. The group of nine agents all follow the public norm A. Eight out of nine agents are consistent, one agent is not (BA). PI is defined to require at least a majority of a group to suffer from that inconsistency, a condition that is not satisfied here.

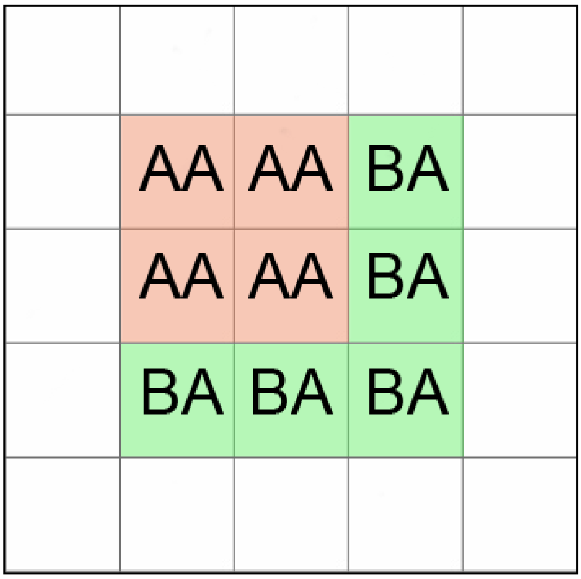


	AA	AA	AA	
	AA	AA	AA	
	AA	BA	AA	

Figure 4 Scenario 3

#4

The fourth scenario serves to show that the definition of a group in the model, which requires connectivity, does not require a rectangle-shaped cluster. The five BA agents form a group with PI as each agent borders at least one other agent with the necessary requirement.



	AA	AA	BA	
	AA	AA	BA	
	BA	BA	BA	

Figure 5 Scenario 4

Experimental Setup

A single simulation runs for a set number of rounds with a set number of agents – hence a set ratio of agents to spaces on the board. To examine the emergence of PI, three movement rules have been devised. They are tested against the control treatment, in which agents move randomly across the board without any interactions. A movement rule is separated into two steps of reasoning. The first question an agent asks is “Should I move?” Intelligently reasoning agents will determine the public norms of their immediate neighbours, and compare them to their own private norm. If the agent feels outnumbered, it will move, otherwise it will stay. The second line of reasoning happens once an agent has determined “Yes, I should move.” The question is now “Where should I move to?” Intelligent reasoning implies that agents detect all free tiles surrounding them, and of those tiles determine which one would be best to move to, using the same norm-comparison process from the previous step. In other words, it can see their neighbours’ neighbours and determine what the majority norm for each surrounding tile is, and pick the location that is best for it it terms of maximising its “comfort”.

The control movement rule is ‘Random’, in which both reasoning steps are randomized. Conversely, ‘Intelligent’ movement rules use intelligent reasoning for both steps. Additionally, two more mixed movement rules were tested (see table 2). The reasons for including the two mixed rules were twofold. First, it was important to see whether perhaps one step of reasoning was sufficient to evade PI, making the two-step process redundant. Secondly, the movement rules aim to emulate settings that can be found in the real world: for example, an agent under the ‘Risky’ treatment might be a person who cannot make the decision to move based on norms and social comparison, but once they discover they have to move, they will try to do so intelligent such as relocate to the best space that they can see.

Movement rule	Should I move reasoning	Where to move to reasoning
Random	Random	Random
Intelligent	Intelligent	Intelligent
Semi-intelligent	Intelligent	Random
Risky	Random	Intelligent

Table 2 all movement treatment conditions

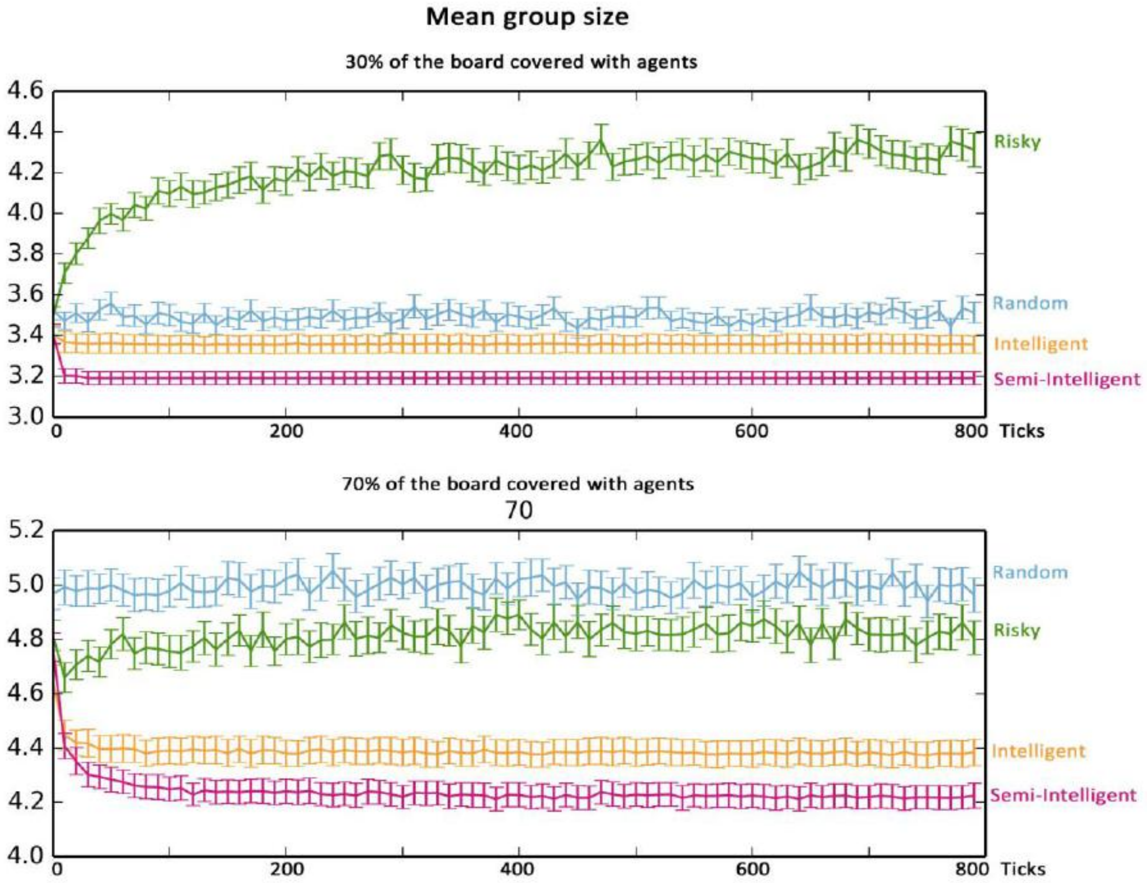
My hypothesis is that on average, all three movement conditions containing intelligent elements should result in less PI compared to the random control, because agents will seek out neighbours with public behaviour corresponding to their own private attitude. The ratio of agents to the board size will influence the results because more agents mean fewer spaces to escape to and thus potentially more occurrences of PI. The Null Hypothesis is that none of the intelligent treatments display any difference in behaviour from the completely random movement of agents across the board.

Results

In line with the hypothesis, the ratio between tiles on the board and the number of agents makes a big difference for both the mean group size as well as the number of groups that are classified as pluralistically ignorant. Figure 6 shows the mean group size of PI groups over time (8000 rounds, with data collected every 10th round). Treatment conditions where agents occupy 30% and 70% of the board are compared. The difference between the two treatments is intuitive: the more agents on the map (70%), the greater the number of agents that constitute a group with PI.

For each density treatment all four movement conditions are compared. As predicted, intelligent and semi-intelligent agents suffer less from PI than random agents do. After only a short period of time of roughly 20 rounds, their numbers settle and remain low over the course of the simulation. It should be noted that the graph constitutes the combined 150 simulation results, not a single run. Thus, each data point at time t is an average of 150 data points at time t. What is striking is that random agents break the pattern completely. Not only are they bad at avoiding PI, they even increase their chances over time. The average group size increases steadily over many rounds, with considerable difference to all other movement conditions. This is perhaps less surprising if we reconsider their behaviour: they will seek neighbours like them, but based on the assumption that their neighbours' public behaviour is an accurate reflection of private beliefs.

Figure 6 Mean group size of P.I. over 8000 rounds



If that is true, then why don't other intelligent agents suffer the same fate? Their secret is simple: they simply do not move as much. As they satisfy their needs by becoming more "comfortable", they stop moving. Risky agents do not have that luxury, as their movement decision is determined randomly.

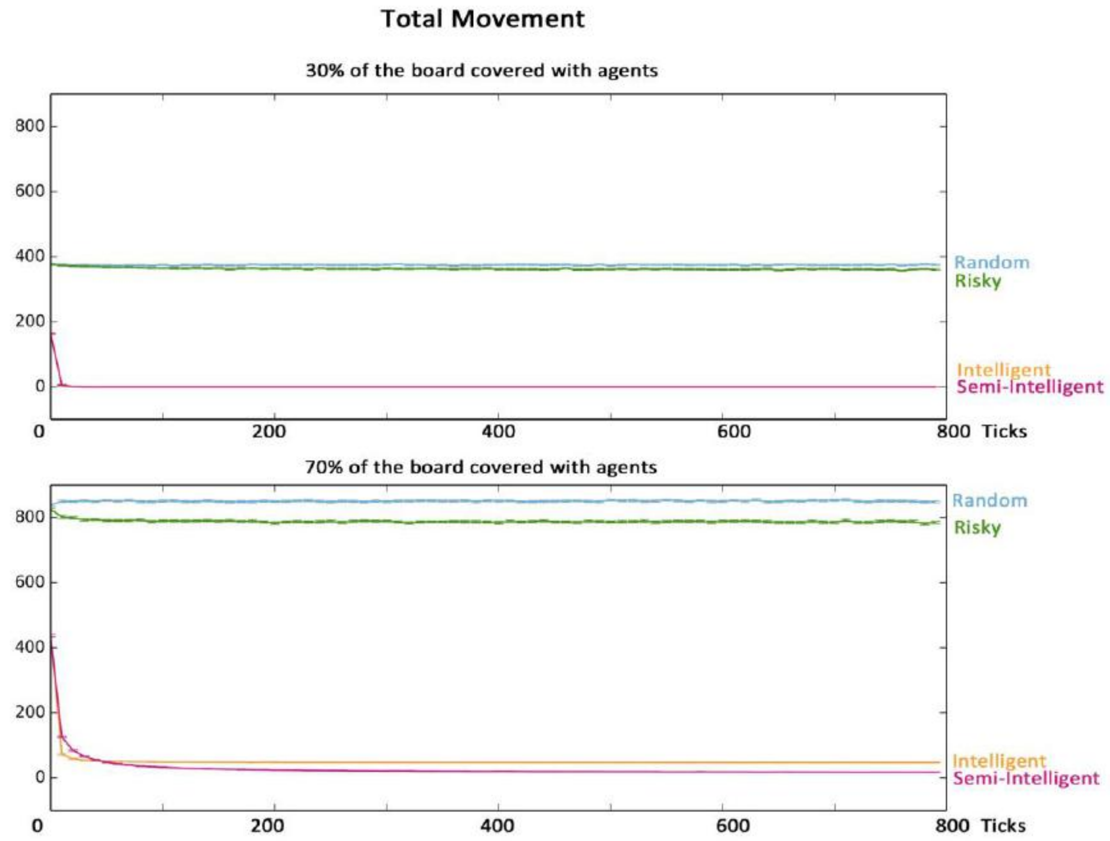


Figure 7 Total number of steps taken at each round, over 8000 rounds

Figure 7 illustrates the movement differences. Note how similar random and risky agents are in total number of steps taken each round. In line with the notion of "the best way to win the game is not to play it", both intelligent and semi-intelligent agents are satisfied quickly and have the same neighbours for the remaining rounds. This also explains why the decision about where to move to can be completely arbitrary, as semi-intelligent agents prove- because they decide so rarely in the first place. The story gets all the more interesting once we look at the 70% treatment. Random and risky movement conditions have now swapped first place and runner up for the highest number of agents in a group, but their movement remains similarly high. The differences in means are still significant between the two. It seems that when the place is much more crowded, the intelligent reasoning that risky agent possess for their second reasoning step prevents PI more successfully than completely random movement does.

Considering that with a crowded board, the decision to move in the first place becomes less meaningful (as they are likely to end up in a bad position again) and thus the random movement decision of risky agents less risky, their PI size decreases.

One interesting observation remains: semi-intelligent agents once again exhibit the least PI members per group. It seems thus most prudent to choose the next step randomly, as long as you reason intelligently whether or not to move in the first place. Why? The answer again can be found in the erroneous assumption that agents make about private and public attitudes. Because intelligent agents are prone to misinterpreting their neighbours' true intentions, picking a next destination by chance is actually less likely to result in such an error. But that only works in combination with intelligent first-step reasoning, as the stark contrast between semi-intelligent and random agents show.

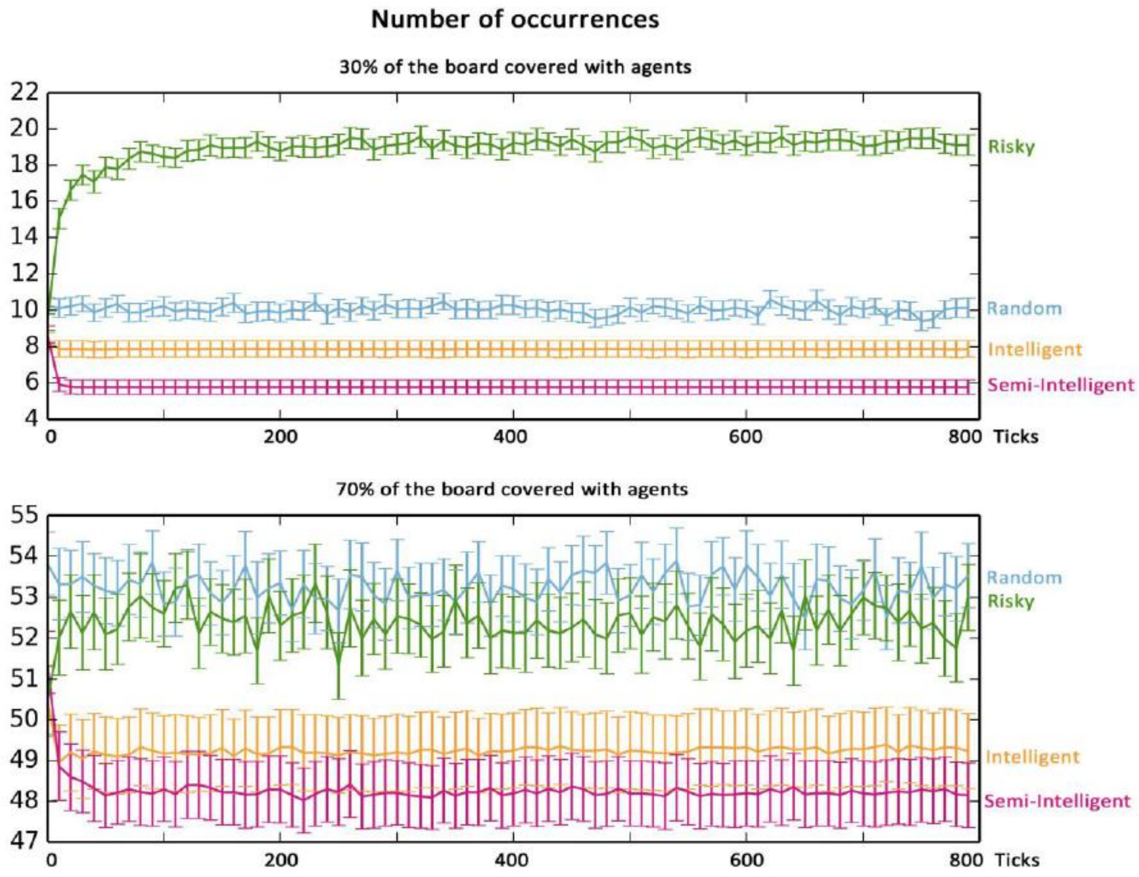


Figure 8 Number of occurrences at each round over 8000 rounds

The same pattern is true for the number of group metric (see Figure 8), with the difference being that at 70% treatment, the random and risky movement conditions differences are no longer statistically significant; their confidence intervals overlap visibly, and the behaviour more variant over time.

Interesting is that risky movement results in higher PI across all treatments until more than 60% of the board are covered with agents: from that moment on, risky movement is more PI averse than random movement. This change in outcome is reflected both in the number of groups as well as the mean group size. The mean group size of risky agents at 90% drops even below that of semi-intelligent agents, although the three non-random movements are difficult to distinguish at that point. The confidence intervals show great overlaps- with the exception of the random control which has significantly larger groups.

To test the differences in means between the treatments at every board-to-agent ratio, a one-way ANOVA was conducted. The Null hypothesis (that the intelligent movement makes no difference to the random treatment) would not be rejected if the differences are too marginal. However, the results are significant for every treatment at every agent-to-board ratio. To see which of these differences made the biggest impact, a Tukey-test was conducted, pairing every combination of the four treatments and testing their differences. The number of groups metric shows significance throughout, but as the agent-to-board ratio passes 50% and higher, the differences in mean between the semi-intelligent and intelligent; and random and risky treatments lose significance. Whereas the p-value for all pairs at 50% of the board are 0.00, they are 0.3 and 0.8 respectively at 60% board coverage. The extreme conditions of 90% board coverage seem to disrupt the trend, however. The difference between random and risky movement is now significant again. The mean group-size metric is more robust and shows high p-values only at the 90% board coverage condition, none of which include a pairing with the random treatment- in other words, the Null-hypothesis can be rejected for all treatments, most of them at the 99% confidence interval. It is unsurprising to see higher p-values for the semi-intelligent and intelligent treatment pairs, given their similarities – the differences are in most cases still significant, however.

Figure 9 plots values of the mean group size at 60% board coverage, comparing the four movement treatment. Intelligent and Semi-Intelligent treatments show great similarity, and even similarity of in-group variance. The risky treatment has the highest values overall and slightly larger spread of data compared with random movement, which has fewer outliers.

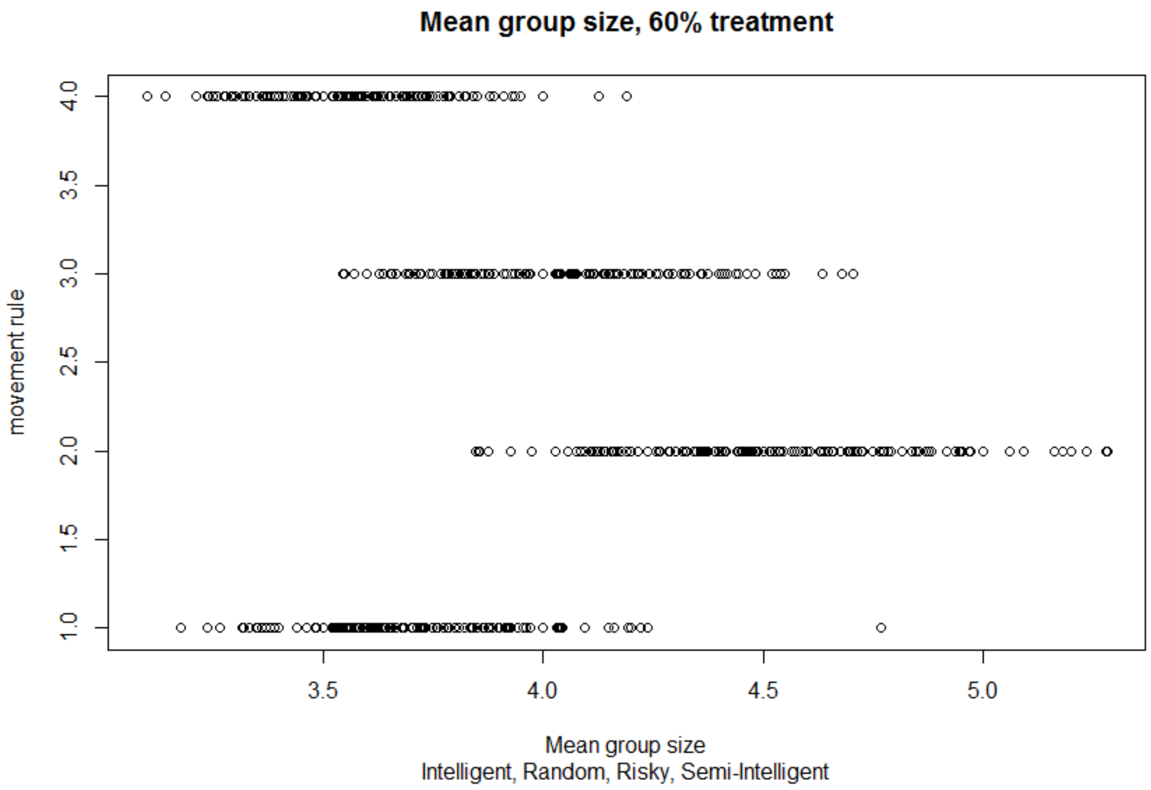


Figure 9 Plotted values of group sizes across the treatment

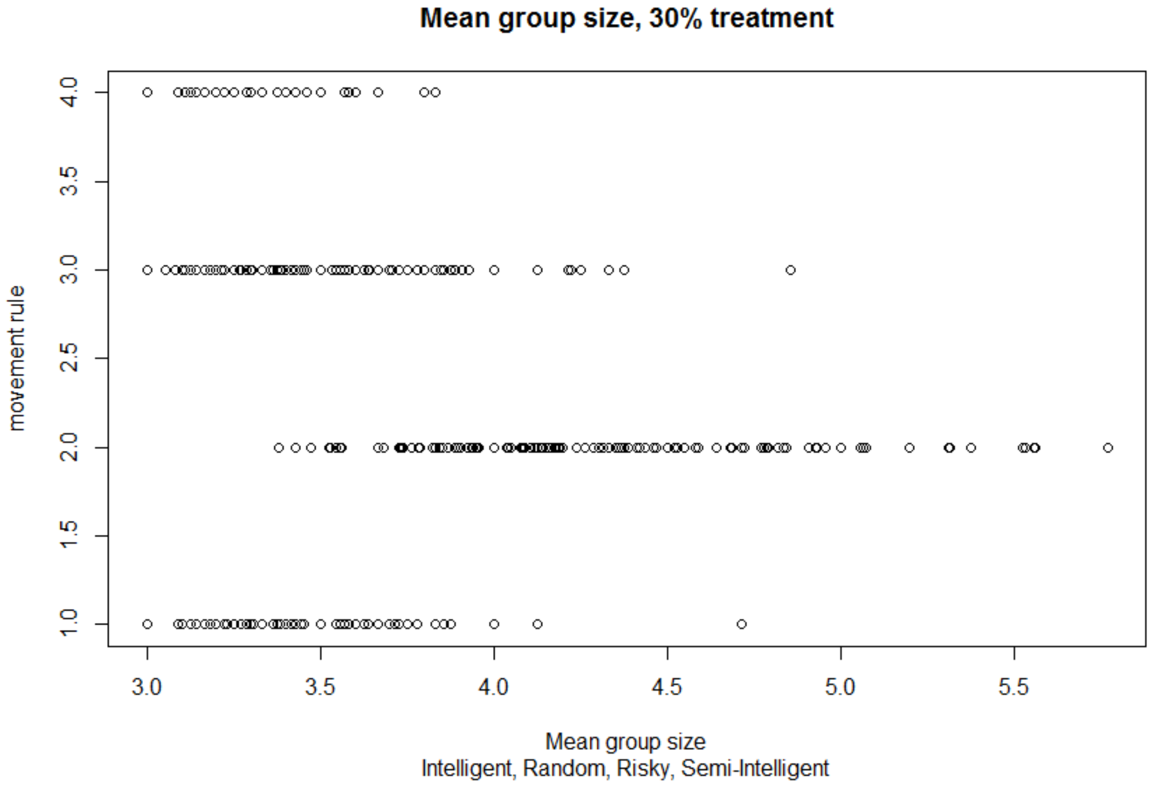


Figure 10 Plotted values of mean group sizes across the 30% treatment

Comparing a low board coverage (30%, see Fig. 10 above) with a high one (80%, see Fig. 11 below), we see quite a few differences in the in-group variance as well.

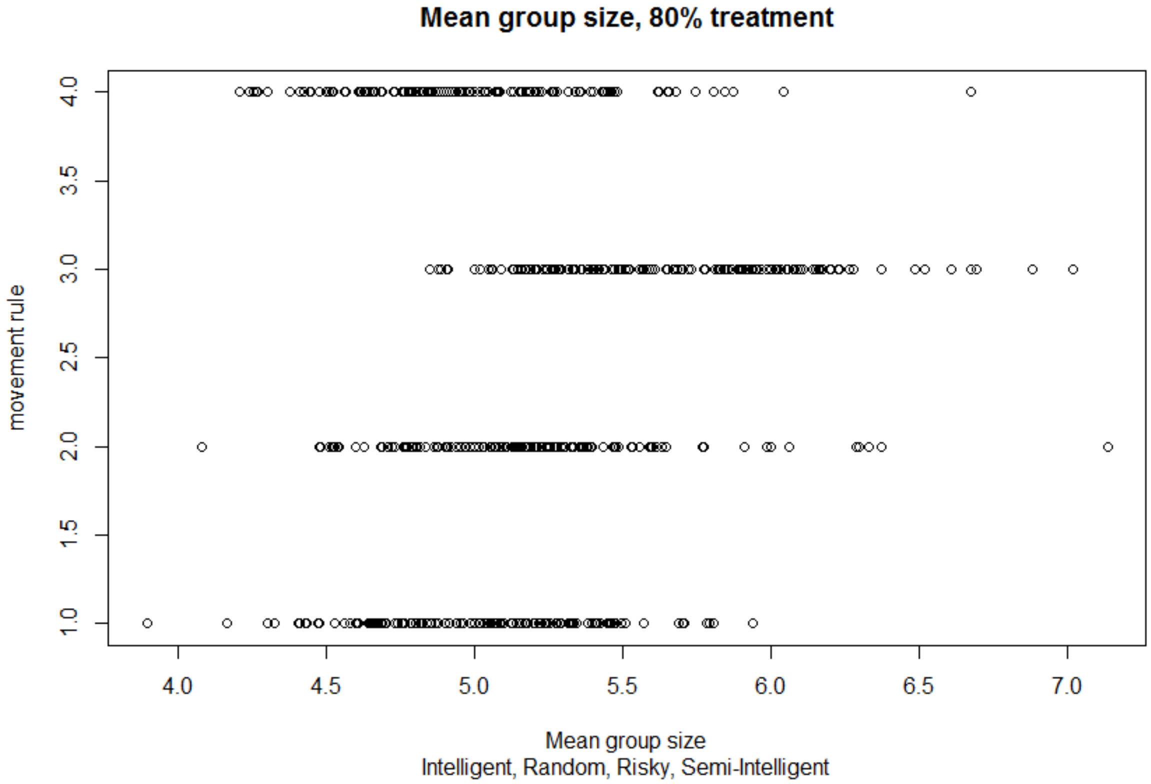


Figure 10 Plotted values of mean group sizes across the 30% treatment

The risky treatment shows a large spread of data in lowly populated conditions, but is much denser when the board is densely covered. This is intuitive given that agents pick their target tile randomly, but will not move if no free tile is available. As the board fills up, the choices are limited and thus, the agents have less options.

Conclusion

In this paper I sought to show that the emergence of PI can be explained by a simple rule of agent behaviour. Rather than switching private attitudes or bending to other kinds of pressure such as global opinions or strength of attitudes, agents solely seek to minimise the difference between their surrounding neighbours and themselves.

The model shows that PI can occur without people actively striving to enforce a norm, but because every person tries to reduce any possible internal friction that might arise from conforming to a norm they do not privately agree with. These micro-behaviours lead to the macro pattern of PI. The answer to how PI emerges is thus: people bring it onto themselves. There is no need for any external pressure through coercion or even legal enforcement; and no active exchange between agents (that might result in peer pressure type settings) is required either.

The model fits well within the literature of a range of issues that deal with discrepancies between people's behaviour and their attitude. In the context of political science, it could be used to study the prevalence of income inequality effects on student results- a phenomenon frequently associated with 'undesirable' norms that students nonetheless accept, thus hindering their own future prospects.

The pattern of PI is a value-neutral phenomenon which can occur as a catalyst for both positive and negative norm-change and preservation. In the case of student drinking and peer pressure, or rules that harm individuals, PI is a hindrance to morally superior norms, if we assume that individual happiness or opportunities are the goal. But it can also occur to bring change such as the changing norms about racist thought- white supremacy was once a widely accepted norm, but in most western countries today this has shifted. Racism still exists, but the attitudes underlying racism are no longer the accepted norm. For such large changes to happen, it is more than plausible to imagine that there has to exist a period of PI in which dissidents of the new norm find themselves in a position in which they are ostracised, and knowing what the new norm has become, will shift to identifying non-breaching of the new norm as an implicit compliance of the new norm, despite not knowing what the others' private attitude may be.

Finding more stringent conditions for groups of PI theory is important to distinguish the phenomenon from social acquiescence bias and other related concepts. I have proposed one such possible definition here, which hopefully can serve as a starting point for further discussion on this matter.

Bibliography

Beheshti, Rahmatollah and Gita Sukthankar. 2014. “A normative agent-based model for predicting smoking cessation trends” Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems (AAMAS '14). International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 557-564.

Bjerring, Jens Christian, Jens Ulrik Hansen, and Nokolaj Jang Lee Linding Pedersen (2014) On the Rationality of PI. *Synthese* 191, pp. 2445–2740.

Boven, van (2000) PI and political Correctness. *Political Psychology* 21(2)

Brennan, Geoffrey, Lina Eriksson, Robert E. Goodin and Nicholas Southwood. Explaining Norms (Oxford: Oxford University Press, 2013).

Centola, Damon, R. Willer and M. Macy. 2005. “The Emperor's Dilemma: A Computational Model of Self-Enforcing Norms” American Journal of Sociology, 110(4): 1009-1040.

Chong, Dennis. Rational Lives: Norms and Values in Politics and Society (Chicago: Chicago University Press, 2000).

Corrigan, Patrick. 2004. “How Stigma Interferes With Mental Health Care” American Psychologist, Vol 59(7), Oct 2004, 614-625. http://dx.doi.org/10.1037/0003-066X.59.7.614

Corrigan, P. W. (2000), Mental Health Stigma as Social Attribution: Implications for Research Methods and Attitude Change. Clinical Psychology: Science and Practice, 7: 48–67

Downs, Anthony. An Economic Theory of Democracy. (Boston: Addison-Wesley Publishing, 1954)

Gunther, Albert C., Richard M. Perloff and Yariv Tsfati “Public opinion and the third-person effect” in The Sage Handbook of Public Opinion Research pp. 184-191

John R. Logan, Jennifer Darrah, and Sookhee Oh (2012) The Impact of Race and Ethnicity, Immigration and Political Context on Participation in American Electoral Politics. *Social Forces* 90 (3): 993-1022

Katz, Daniel and Floyd H. Allport. "Introduction." Chapter I in Daniel Katz and Floyd H. Allport, Students' Attitudes: A Report of the Syracuse University Reaction Study. Syracuse, NY: Craftsman Press (1931): 1-8.

Krosnick, Jon A. ‘Survey Research’ *Annual Review of Psychology* (1999) No. 50, pp. 537-567.

Kuran, Timur. Private Truths, Public Lies: The Social Consequences of Preference Falsification. (Cambridge: Harvard University Press, 1995)

Lerman, Kristinam Xiaoran Yan and Xin-Zeng Wu. The Majority Illusion in Social Networks. *Computers and Society* arXiv:1506.03022

Mellon, Jonathan and Posser (2015) Investigating the Great British Polling Miss: Evidence from the British Election Study *Draft paper*

Miller, Dale T.; McFarland, Cathy in Suls, Jerry (Ed); Wills, Thomas Ashby (Ed), (1991). Social comparison: Contemporary theory and research. , (pp. 287-313). Hillsdale, NJ, England: Lawrence Erlbaum Associates

O'Gorman, Hubert J (1986). The Discovery of PI: An ironic lesson. *Journal of the History of the Behavioural Sciences* 22

Prentice, Deborah A., and Dale T. Miller. (1993) "PI and Alcohol Use on Campus: Some Consequences of Misperceiving the Social Norm." *Journal of Personality and Social Psychology* 64 (2): 243–56

Rauhut, Groeber (2010) Does ignorance promote norm compliance? *Computational and Mathematical Organization Theory* 16 pp. 1-28

Sanford Labovitz and Robert Hagedorn. Measuring Social Norms. The Pacific Sociological Review Vol. 16, No. 3 (Jul., 1973), pp. 283-303

Wang, Sheng-Wen, Cung-Yuan Huang and Chuen-Tsai Sun (2013) "Modeling self-perception agents in an opinion dynamics propagation society". Simulation, 90(3): 238-248.

Savitsky, Kenneth, Nicholas Epley, and Thomas Gilovich (2001) "Do Others Judge Us as Harshly as We Think? Overestimating the Impact of Our Failures, Shortcomings, and Mishaps." Journal of Personality and Social Psychology 81 (1): 44–56

Schank, R. L. (1932) "A Study of Community and Its Group Institutions Conceived of as Behavior of Individuals." *Psychological Monographs* 43 (2): 1–133.

Shaw, John R. QuickFill: An efficient flood fill algorithm Code Project, 12 March 2004. URL: http://www.codeproject.com/Articles/6017/QuickFill-An-efficient-flood-fill-algorithm last retrieved: 12 September 2014.

Willer, Kuwabara and Macy (2009) The False Enforcement of Unpopular Norms. *American Journal of Sociology* 115(2), pp. 451-490