On the fragility of pluralistic ignorance

Kevin Ferneding s242823@dtu.dk
Søren Stange s204229@dtu.dk
Freja T. Dollas s204248@dtu.dk
Jeppe B. Weikop s253874@dtu.dk
Technical University of Denmark
Lyngby, Denmark

Abstract

We analyze and explore the fragility of pluralistic ignorance, focusing on the paradigmatic case of *The Emperor's New Clothes*. Firstly, we argue that an assumption of independence of beliefs is unhelpful in the fragility analysis. Then, we propose a Bayesian approach to modeling belief revision in a state of pluralistic ignorance where beliefs are non-independent. Finally, we develop a simulation framework and show that the dissolution of pluralistic ignorance can indeed be modeled computationally. We finish with future research directions.

1 Introduction

Pluralistic ignorance is a term originating from the social sciences, coined in 1931 by social psychologist Floyd H. Allport and his student Daniel Katz [2]. In colloquial terms, it describes a social situation in which *no one believes, but everyone believes that everyone else believes*; that is, everyone in the social group privately hold the same belief regarding a particular matter, but mistakenly assume that they are in the minority. In order to conform to the perceived norm of the group, the group members outwardly express support for the opposing belief, and this insincere outward conformity then reinforces the very misperception that produced it. As a result, the social group as a whole becomes organized around a belief that none or few of the group members genuinely hold.

The study of pluralistic ignorance is of interest partly because of the potential consequences of social groups acting against their own personal conviction. When people coordinate around a norm they each privately reject, it distorts collective decision-making and allows inaccurate or even harmful beliefs to persist far longer than they should. This is reflected in some of the most well-known cases of pluralistic ignorance where whole communities are steered towards outcomes that no member of the community actually desires. Three such classical examples are described below.

> **College drinking**: At a college welcome party for freshmen, each freshman expects that other freshmen thoroughly enjoy drinking. To conform to this perceived social norm, each freshman consumes a large amount of alcohol at the party, despite that they do not find drinking enjoyable at all. Upon observing the excessive drinking of other freshmen, each freshman forms the belief that they are the only freshman that does not enjoy drinking.
>
> **Classroom**: After presenting some very challenging material in class, a teacher turns to the students and asks whether they have any questions. Every student is confused about the material, but since no one else raises their hand, each student forms the belief that no one else is confused. To avoid the stigma of being the only confused student, no student asks a question.

Submitted to:
02287 Logical theories for Uncertainty and Learning

The Emperor's New Clothes: In the classical folktale *The Emperor's New Clothes*, two traveling con-men offer to weave the Emperor a magnificent set of clothes that is invisible only to those who are either incompetent or very stupid. Upon his acceptance, the con-men pretend to weave and dress the Emperor in a suit, and the Emperor – not wanting to disclose that he does not see the cloth – pretends to admire the non-existent suit and parades through town wearing it. The townsfolk initially pretend admiration to avoid looking foolish. Eventually, a child cries out that the Emperor is wearing nothing at all, causing the rest of the crowd to follow.

Collectively, we refer to these examples of pluralistic ignorance as the *paradigmatic cases*. They are of special importance when considering definitions of pluralistic ignorance, because any attempt to define or characterize the phenomenon should capture these quintessential instances.

2 Background

In recent decades, the social phenomenon of pluralistic ignorance has gathered interest in scientific communities concerned with formal epistemology, and several authors have approached the topic with formal methods like logic, game theory, and Bayesian learning. A prerequisite for the application of such methods is naturally that pluralistic ignorance must be precisely defined, and – importantly – that the definition allows pluralistic ignorance to occur among rational agents.

Significant efforts to develop a formal definition that adequately captures the social-psychological subtleties of the phenomenon have been made, and a taxonomy was established by Bjerring et. al. in 2014 defining four levels of definitions, in increasing order of strictness [2]. These are relayed below.

PI_1: *a situation where an individual member of a group believes some proposition P and mistakenly believes that the other members of the group believe* $\neg P$.

PI_2: *a situation where a group of individuals* ***all*** *believe some proposition P and they* ***all*** *mistakenly believe that the other members of the group believe* $\neg P$.

PI_3: *a situation where a group of individuals* ***all*** *believe some proposition P and they* ***all*** *mistakenly believe that the other members of the group believe* $\neg P$ ***and*** *this situation is brought about by the group being in a special kind of social context*.

PI_4: *a situation where a group of individuals* ***all***

(i) *privately believe some proposition P;*

(ii) *believe that everyone believes* $\neg P$*;*

(iii) *act contrary to their private belief that P (i.e. act as if they believe* $\neg P$*);*

(iv) *take the actions of the others as strong evidence for their private beliefs about P.*

Bjerring et. al. argue that PI_4 most precisely describes pluralistic ignorance, as one can develop scenarios that fit the other definitions, but do not seem to correspond to paradigmatic cases of pluralistic ignorance [2]. In other words, PI_1-PI_3 are over-generalizations. This is of course partly a matter of interpretation; there is no single correct way to capture the characteristics common to all cases of pluralistic ignorance, just as social psychologists may disagree whether a particular social group is in a state of pluralistic ignorance. In this work, we apply the PI_4 definition proposed by Bjerring et. al.

2.1 Rationality of pluralistic ignorance

Bjerring et al. make a three-part argument to show that there is reason to believe in the rationality of pluralistic ignorance in some cases, and thereby that pluralistic ignorance can arise in groups of rational agents [2]. Firstly, they argue that it can be pragmatically rational for agents to conform to (iii) of PI_4 if the reasons for coordinating with the group against their own belief outweigh the consequences of the alternative, e.g. in the Emperor's example risking looking unfit for ones office. Secondly, they argue that it can be rational to conform to (iv) of PI_4. In many cases the observational evidence (that an agent observes that other agents act as $\neg P$) might outweigh introspective evidence (that she herself believes P and acts as $\neg P$). When that is the case, the agents act epistemically rational when conforming. Thirdly, they argue that it may be rational to conform to (iii) and (iv) of PI_4 *all things considered*, since conforming to both when the considerations of doing so outweighs not doing so are shown to be rational.

Thereby, they show that it is possible to have rational agents and still experience cases of pluralistic ignorance. This is an important observation, because it allows instances of pluralistic ignorance to be modeled as systems of rational agents.

2.2 Fragility of pluralistic ignorance

Pluralistic ignorance is often portrayed as a *fragile* social state, in the sense that if only a few members of the social group decide to publicly express their true belief, many more are likely to follow [3]. Indeed, this is exactly what occurs in the classical fairy tale *The Emperor's New Clothes* in which a young child's blunt remark triggers a cascade of other group members admitting to their private belief, shattering the shared illusion and dissolving the state of pluralistic ignorance. However, Hansen showed in 2011 that fragility is not inherent under PI_2 [3]; nothing in the definitions PI_1-PI_4 prevents the incorrect beliefs of agents from being robust. *Robust* in this context means that one agent announcing their private belief does not change the other agents' beliefs about each other (except the announcer), and thus pluralistic ignorance persists among the rest of the group. Note that we consider pluralistic to be dissolved only when no agent holds false beliefs about the private belief of any other agent [3].

In [3], Hansen operates under a PI_2 definition and concisely formalizes it in the following way:

$$\bigwedge_{a \in \mathbb{A}} \left(\neg B_a \phi \wedge \left(\bigwedge_{b \in \mathbb{A} \setminus \{a\}} B_a B_b \phi \right) \right). \tag{1}$$

Indeed, nothing in this definition makes pluralistic ignorance an inherently fragile phenomenon, nor do the two additional conditions in PI_4 change this property. An agent a can hold beliefs about other agents b and c such that $B_a B_b \phi$ and $B_a B_c \phi$ are completely separate beliefs. Intuitively, agent a might hold each belief for a different reason, and in that case learning agent b's true belief changes nothing about agent a's belief about agent c.

This assumption of independence of beliefs, however, is not useful in the analysis of the fragility of pluralistic ignorance, because it fails to capture the dynamics present in paradigmatic cases. Consider for instance the *Classroom* example. If Alice knows Bob to be a very diligent and capable student, and Bob announces that he did not fully understand the teacher's presentation, then surely Alice will reason that the presentation was in fact difficult to follow. Similarly, in *The Emperor's New Clothes*, agents will

start to doubt the Emperor's invisibility claims if a highly competent and respected townsman announces that he perceives the Emperor as unclothed, but perhaps not if a local fool makes a similar announcement.

3 Fragility analysis under non-independence

In order to more accurately describe the fragility of pluralistic ignorance in cases where beliefs are *not* independent, but rather tied to a certain property of the agents, we need to consider a system of agents in which agents hold beliefs about each other in regard to this property in addition to the beliefs required by PI_4. These additional beliefs dictate how agents internally update their private beliefs as a result of truthful announcements made by other agents.

We take a Bayesian approach to modeling this system of agents. As a guiding example, we consider *The Emperor's New Clothes* and borrow the terminology from there. Initially, all agents perceive that the Emperor is unclothed $\neg\phi$ and believe that all other agents perceived that the Emperor is clothed ϕ.

$$\bigwedge_{a \in \mathbb{A}} \left(B_a \neg\phi \wedge \left(\bigwedge_{b \in \mathbb{A}\setminus\{a\}} B_a B_b \phi \right) \right) \tag{2}$$

Additionally, agents initially trust the Emperor's testimony Ω that any competent agent believes that the Emperor is clothed. This is expressed below with ψ_b denoting "agent b is competent".

$$\bigwedge_{a \in \mathbb{A}} B_a \Omega \;=\; \bigwedge_{a \in \mathbb{A}} \left(B_a \bigwedge_{b \in \mathbb{A}} (\psi_b \to B_b \phi) \right) \tag{3}$$

Since agents associate belief in ϕ with competence, an agent a would be surprised to learn that another agent b believes $\neg\phi$ if agent a considers agent b highly competent. In this case, agent a finds herself with conflicting beliefs: she believes that agent b is competent (denoted $B_a\psi_b$) while her other beliefs imply that agent b is not competent (denoted $B_a\neg\psi_b$), as demonstrated below.

$$\begin{aligned}
& K_a B_b \neg\phi \wedge B_a(\psi_b \to B_b\phi) && \\
\equiv\; & K_a B_b \neg\phi \wedge B_a(\neg B_b\phi \to \neg\psi_b) && \text{(by contraposition)} \\
\Rightarrow\; & B_a B_b \neg\phi \wedge B_a(\neg B_b\phi \to \neg\psi_b) && \text{(knowledge entails belief)} \\
\Rightarrow\; & B_a \neg\psi_b && \text{(by distribution of } B\text{)}
\end{aligned}$$

To resolve this conflict, the agent quantifies her beliefs. We denote agent a's regard for agent b's competence with a number $\psi_{a,b} \in [0,1]$. For n agents, we construct an $n \times n$ matrix C where entry $C_{i,j} = \psi_{i,j}$ describes agent i's regard for agent j's level of competence. Entries along the diagonal of the matrix are zero if agents do not consider their confidence in their own competence (hereafter *self-competence*), as seems to be the assumption in paradigmatic cases. However, agents can take self-competence into account with no further change required. Note that the competence matrix C can be considered an adjacency matrix representation of the social graph, and it is generally sparse, as agents only have an opinion about the competence of a small subset of the other agents. If agents have a base-level regard for the competence of agents with which they have no social relation, this value can be used instead of 0 in matrix entries corresponding to non-edges.

Initially, all agents trust Ω with a prior probability $P(\Omega) = K$ close to 1. When an agent b announces $[!B_b\phi]$, this triggers a belief revision in the other agents that have a non-zero regard for agent b's competence. For instance, consider the belief revision in agent a after agent b's announcement with $K = 0.95$ and $\psi_{a,b} = 0.72$.

$$\begin{aligned} P(\Omega \mid B_b\phi) &= \frac{P(B_b\phi \mid \Omega) \cdot P(\Omega)}{P(B_b\phi \mid \Omega) \cdot P(\Omega) + P(B_b\phi \mid \neg\Omega) \cdot P(\neg\Omega)} \\ &= \frac{(1-\psi_{a,b}) \cdot K}{(1-\psi_{a,b}) \cdot K + (1-K)} \approx 0.8417 \end{aligned}$$

Note that $P(B_a\phi \mid \neg\Omega) = 1$, because if the testimony Ω is false, then all agents perceive the Emperor to be unclothed, regardless of competence. As agents announce their belief in $\neg\phi$, other agents gradually lose confidence in Ω and become more inclined to express their own private belief. In this way, pluralistic ignorance is eventually dissolved.

It is important to note that agents *do not* reason about other agent's beliefs on the basis of their competence. For example, if Alice considers Bob significantly more competent than Charlie and Bob announces his belief $B_{Bob}\phi$, Alice does *not* start to doubt if Charlie's private belief matches their public belief, as this would be in violation of condition (iv) in the definition of PI_4. Whether this is a desirable property in a formal definition of pluralistic ignorance is a discussion beyond the scope of this work.

4 Simulation

To gain an understanding of how pluralistic ignorance is dissolved in practice, we attempt to simulate a state of pluralistic intelligence in a network of Bayesian agents that operate as described in the previous section. Firstly, we construct a realistic social network using the Barabási-Albert model for scale-free networks [1], imitating how social connections are formed in the real world. Vertices in the graph correspond to agents, and the weighted directed edges indicate the source node's *opinion* regarding the competence of the destination node. These numeric opinions are drawn from a uniform distribution on $[0, 1]$. An edge with a low base-level value is added between agents without a relation in the network.

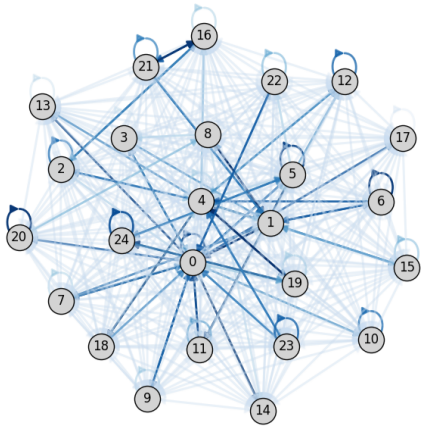


Figure 1: Social network of 25 agents randomly generated with the Barabási-Albert algorithm. The faint edges encode the perceived competence between nodes without a social relation. As expected in a social network, the graph contains *'hubs'*, i.e. nodes with a much higher degree than the graph average.

Each agent in the simulation maintains an internal state consisting of a private belief, a public belief, a belief in the Emperor's testimony Ω, a list of opinions on the competence of all other agents, and a record of the public beliefs of all other agents. When an agent changes their public belief, all other agents perform Bayesian belief revision based on their individual opinion of the announcer (as described in previous section). If an agent's belief in Ω drops below a certain threshold, that agent also publicly expresses their private belief in the next round.

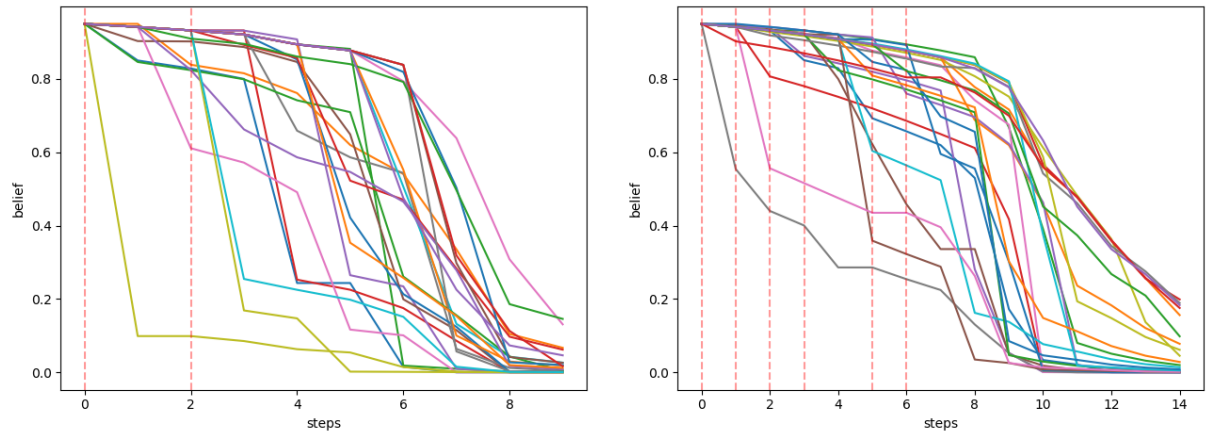
In the simulation, we distinguish between *spontaneous public announcements* and *attitude shifts*. The former describes a bold or reckless agent publicly expressing their private belief even though the agent is not at all certain that the Emperor's testimony is false – much like the child does in the fairy tale. The latter describes an agent that on the basis of other announcements has grown sufficiently certain of the falsehood of the Emperor's statement and therefore chooses to outwardly express their private belief. Spontaneous announcements are only made in rounds where no agents shift their attitude. The spontaneous announcer is picked at random.

The key metrics describing how quickly pluralistic ignorance is dissolved is then the number of rounds and the number of spontaneous announcements. If pluralistic ignorance is truly fragile (as in the tale), we expect a small number of spontaneous announcements.

4.1 Simulation results

The simulation framework contains several configurable parameters, including the number of agents, the number of social connections per agent (see [1]), the initial Bayesian prior $P(\Omega) = K$, the attitude shift threshold, and the perceived base-level competence of unknown agents.

Additionally, there is an element of randomness in the simulation: the dissolution is significantly accelerated by public announcements by well-connected agents, as shown in the simulations in Figure 2.



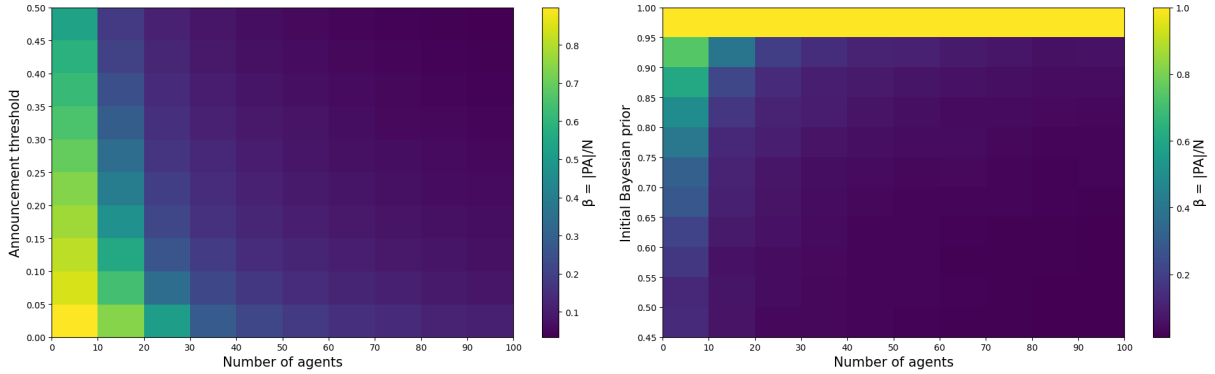
(a) Pluralistic ignorance dissolved in 9 steps.

(b) Pluralistic ignorance dissolved in 14 steps.

Figure 2: Dropping belief in Ω among a network of 25 agents in two simulations with identical parameters. In (a), only two spontaneous announcements by two well-connected agents are required for dissolution. In (b), a total of six spontaneous announcements are required.

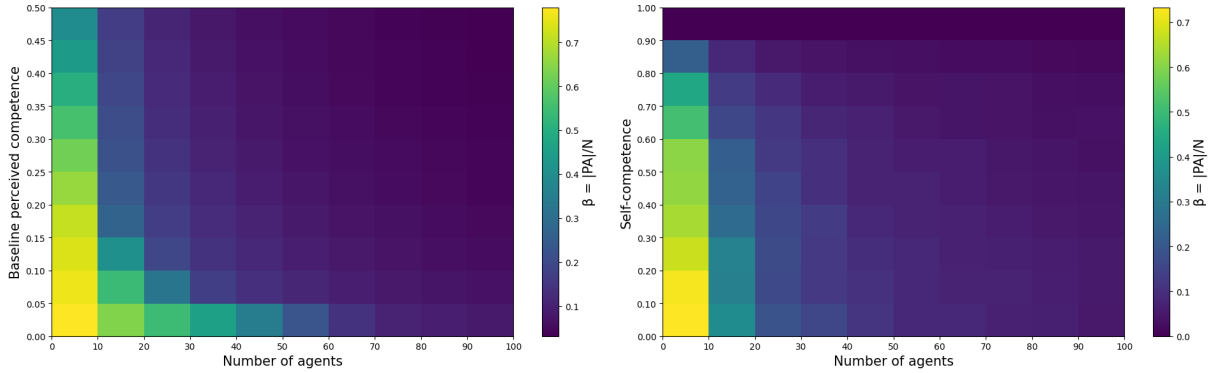
It should be emphasized that the dropping belief in the Emperor's testimony is *not* what defines the dissolution of pluralistic ignorance; dissolution is defined by what agents believe that other agents believe, which in turn is defined by the public beliefs of agents. However, these graphs nicely illustrate how the shared illusion collapses under spontaneous public announcements.

To explore and compare how the each of the simulation parameters affect the rate of dissolution, we define the metric $\beta = \frac{|\mathcal{P}A|}{N}$, i.e. the ratio of number of public announcements to number of agents. It captures how large a proportion of the agents must spontaneously announce their private belief in order to dissolve the state of pluralistic ignorance (on average). In the heatmap illustrations below, the effect of each parameter is visualized. Each β-value is computed as the mean β-value of 1000 simulation runs.



(a) Effect of announcement threshold.

(b) Effect of initial prior (K).



(c) Effect of baseline competence.

(d) Effect of perceived competence of self.

Figure 3: Heatmap visualizations of the impact of simulation parameters.

Overall, we find that only a small proportion of agents are required to perform a spontaneous announcement for dissolution in most configurations, particularly in systems with more than 20 agents. This result supports the observation that pluralistic ignorance is fragile. However, it should be emphasized that these simulations presuppose a set of assumptions, namely that agents wish to publicize their private beliefs once sufficiently certain, that all such announcements are truthful, and that all agents are equally likely to perform a spontaneous announcement. One can certainly imagine similar models where these assumptions are not made.

5 Further work

While the proposed simulation key aspects of pluralistic ignorance, it relies on some simplifying assumptions. We assumed fixed values for announcement thresholds and agent stubbornness. In reality, social thresholds might shift after public announcements and people's stubbornness might vary over time. Agents might have different social thresholds due to stronger or weaker consequences of extraordinary beliefs. Additionally, the competence matrix C is assumed to be static. For a more precise simulation of social dynamics, future work should examine how trust changes over time, how public recognition emerges and how unexpected behavior by connected agents affects perception. Furthermore, we assumed that agents never change their private belief. It remains to be investigated if pluralistic ignorance could lead to the adaption of an actual false belief. Such a dynamic could prevent pluralistic ignorance from dissolving in social networks. The public and private beliefs were modeled as binary values. Future works could explore more nuanced representations of beliefs, for example in the form of Dempster-Shafer belief functions.

In section 3, we mentioned that agents do not reason about other agents' beliefs on the basis of their competence, as this would violate the definition of PI_4. While this belief inference could capture social relations such as expertise and presumptuous behavior better, it would break the conceptual definition of pluralistic ignorance as a situation in which nobody suspects others to misrepresent their beliefs. It should be further investigated if this is a desirable property of a formal definition. If it is not, that should be reflected in the aforementioned dynamic updates of the social network.

6 Conclusion

In this work, we examined pluralistic ignorance with a focus on its fragility. We combined epistemic logic with a computational simulation of belief dynamics in a social network. To address limits of traditional independence assumptions, we introduced a Bayesian belief update framework to formalize cascading belief revisions after dissenting belief announcements. We implemented a simulation environment inspired by *The Emperor's New Clothes* and analyzed the dissolution of pluralistic ignorance. Our results have demonstrated its fragility successfully; in most simulations, only a few announcements triggered collective attitude shifts. Further analyses have looked into the disproportional impact of highly connected agents and the impact of parameters such as threshold and initial priors. Overall, our findings indicate that pluralistic ignorance is highly dependent on private conviction, social conformance and perceived competence. This complex interplay makes pluralistic ignorance inherently unstable under realistic conditions. Several social dynamics were ignored for conceptual reasons. Future work could incorporate richer dynamics to further enrich the link between theoretical understanding and practical emergence and dissolution of pluralistic ignorance.

References

[1] Albert-László Barabási & Réka Albert (1999): *Emergence of Scaling in Random Networks*. *Science* 286(5439), pp. 509–512, doi:10.1126/science.286.5439.509. arXiv:https://www.science.org/doi/pdf/10.1126/science.286.5439.509.

[2] Jens Christian Bjerring, Jens Ulrik Hansen & Nikolaj Jang Lee Linding Pedersen (2014): *On the rationality of pluralistic ignorance*. *Synthese: An International Journal for Epistemology, Methodology and Philosophy of Science* 191(11), pp. 2445–2470, doi:10.1007/s11229-014-0434-1.

[3] Jens Ulrik Hansen (2012): *A Logic-Based Approach to Pluralistic Ignorance*, pp. 67–80.