THE POLYGRAPHS PROJECT: *SIMULATING EPISTEMIC INJUSTICE*

Dr Brian Ball
Associate Professor of Philosophy
AI & Information Ethics Lead
Northeastern University - London

INTRODUCTORY OVERVIEW

The informational environment has changed dramatically since the advent of the internet at the end of the last century, giving rise to various concerns about attitudes and opinions within contemporary societies, and the behaviours they may bring about.

The PolyGraphs project investigates the influences on public opinion of social network *structures* and information consumption *strategies*: our first public engagement workshop (in November 2022) was concerned primarily with the former; the second (in June 2023) was concerned with the latter.

Our research team uses computer simulations of communities of inquiring agents, who learn from their own observations, and from the testimony of their network neighbours; and we analyze and interpret the results (forthcoming a) using various measures of group opinion (in progress a).

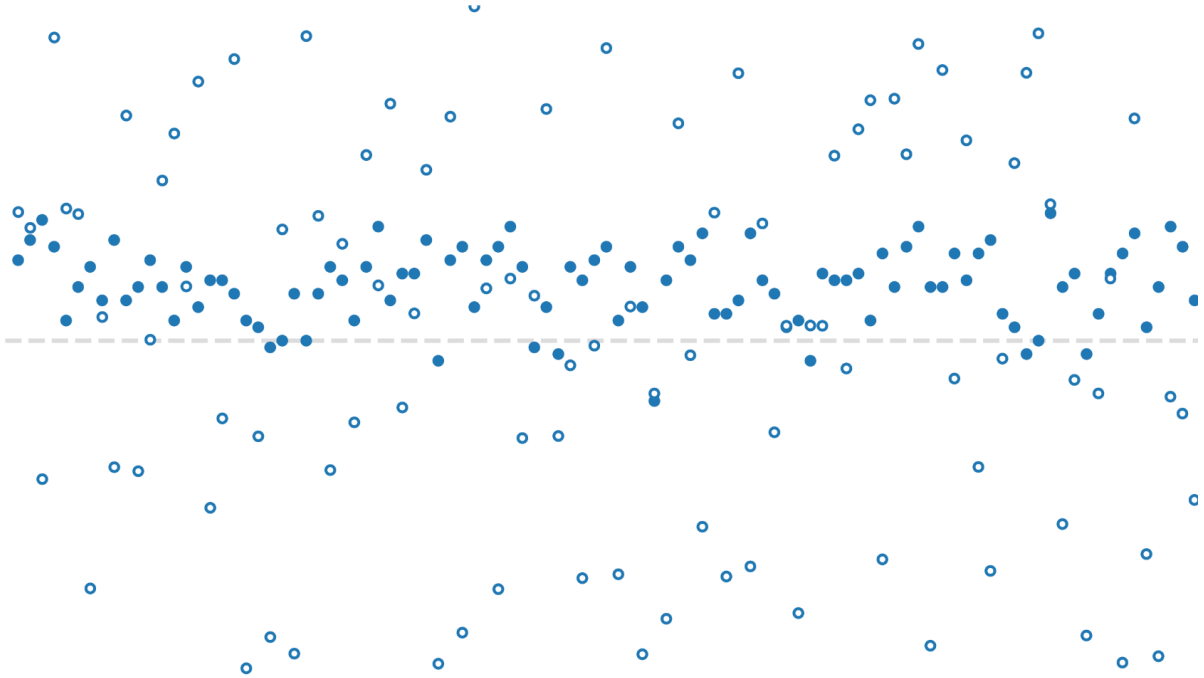
The computational framework we employ has been built to be efficient, customizable, and suitable for integration with machine learning. It is available open source on GitHub.

Our data can also be experienced – see here for a guide to interpreting our visualizations.

OUR MODEL AND ITS IDEALIZATIONS

Our simulations model communities as networks of agents investigating a hypothesis and communicating relevant evidence with their neighbours (cf. Bala and Goyal, 1998; Zollman, 2007) .

We make a number of idealizations: we assume the opinions concern factual matters, and so are either *correct* (true) or *incorrect* (false); our agents are *rational*, in the sense that they are appropriately responsive to evidence; and the *evidence* is stochastic, or chancy – it can be thought of as the results of coin tosses, to determine whether there is a bias towards heads or tails.



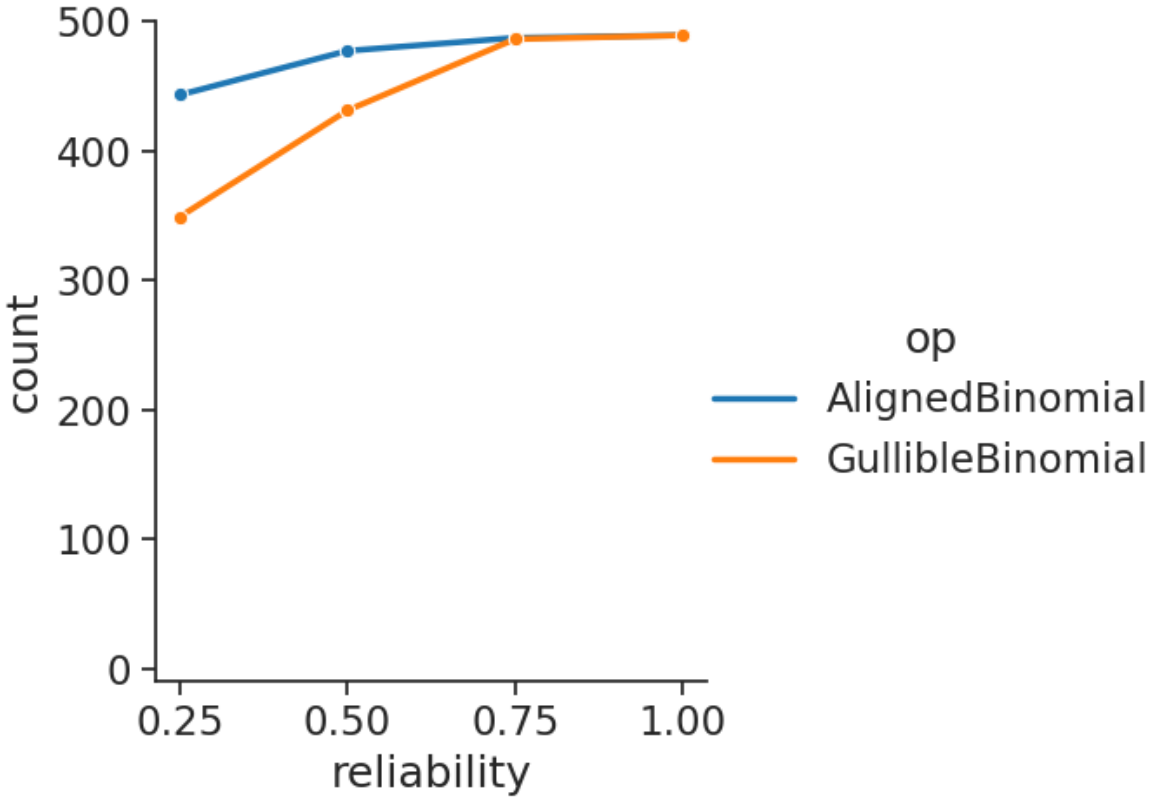
HOMOPHILY AND HIGHER-ORDER EVIDENCE

Others have found that, even in an environment comprising only accurate information, distrust of those with divergent opinions can lead to polarization (O'Connor & Weatherall, 2019). We showed that, in their (homophily-based) models, *more trust led to more knowledge* (forthcoming, b).

In our own (higher-order evidence) models (in progress b), we simulate the effects of introducing mis- and disinformation into the environment when agents pursue various information processing strategies.

We distinguish mere *misinformants* from *disinformants*. The former provide 'evidence' that is neutral overall with respect to the underlying question, which may therefore be thought of as 'noise'; whereas the latter present testimony that is biased away from the truth.

We also distinguish a trusting, or '*gullible*', strategy for processing the information available, from a more sceptical strategy in which the level of trust is '*aligned*' with the level of reliability of the informants in the networked community.



OUR (HOE) FINDINGS

We find that, whether agents pursue the gullible or aligned strategies, the more misinformants are present in the network, the less likely it is that a correct consensus will emerge in the community, and when it does, it takes longer to arrive at this opinion (i.e. the truth).

We also find that, for a given level of misinformation, the aligned strategy is more likely to achieve a correct consensus than the gullible one, but it takes longer to arrive at that consensus (when there is a significant difference in the number of simulation steps required). In short, when we compare the two strategies, there is a trade-off between *accuracy* and *efficiency*.

In the presence of disinformation, the ability of gullible agents to discern the truth plummets, collapsing almost entirely when levels of disinformation are high. Agents pursuing the aligned strategy do better in this regard, but are nevertheless significantly delayed in arriving at the truth.

EPISTEMIC INJUSTICE

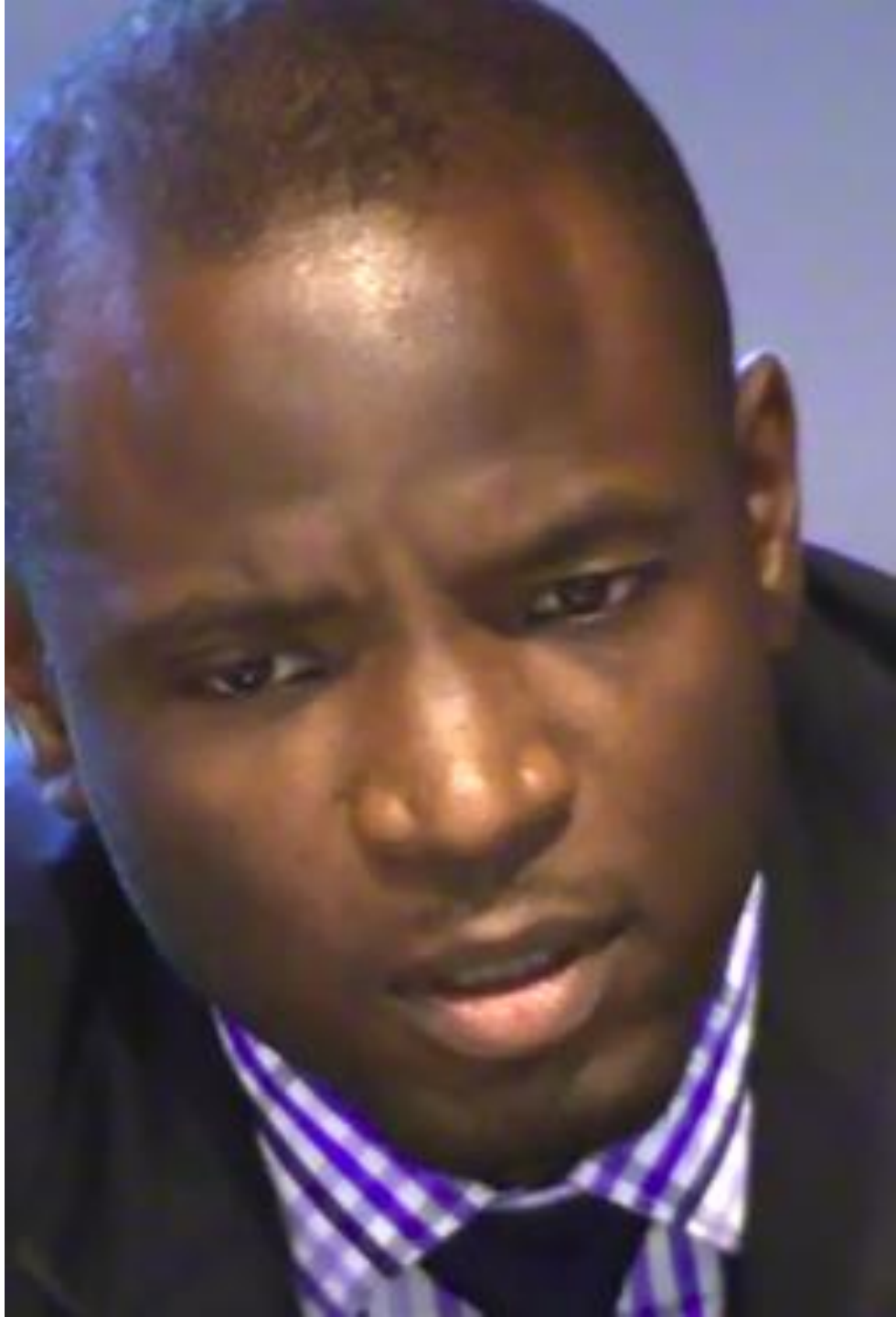
It is in this context that the prospect of simulating epistemic injustice arises.

Epistemic injustice consists in 'a wrong done to someone specifically in their capacity as a knower' (Fricker, 2007: 1).

One especially important variety of epistemic injustice for our purposes is testimonial injustice, which occurs when testimony is not believed due to the speaker's identity.

Fricker gives the example of the police investigating the 1993 racially aggravated murder of Stephen Lawrence in London, who did not trust his friend and witness Duwayne Brooks (later a councilor for Lewisham, pictured) to provide relevant evidence.

A subsequent public inquiry into the murder investigation found the Metropolitan Police to be 'institutionally racist'.



MODELLING TESTIMONIAL INJUSTICE

We can model testimonial injustice as involving a misalignment – in the form of either a credibility deficit or excess (Medina, 2011) – between levels of trust and reliability (or trustworthiness) based on group identity. For example, there is a credibility deficit for group G if reliability in G is 75% but members of G only receive 50% trust; and there is an excess if the figures are reversed.

And we can combine this with a homophily assumption: we might allow that members of G1 trust other members of the same group to a degree that aligns with, or exceeds, the level of trustworthiness in the group, while trusting members of other groups (G2, …) to a level below the level of trustworthiness in those communities.

Our aim is to look at the ill-effects on community opinions of such epistemic injustices.

POTENTIAL FOR COLLABORATION

Our attempts at simulating epistemic injustice have encountered setbacks.

First, we hoped to link our PolyGraphs simulation framework to a Twitter data set hosted at the Lazer lab; but Elon Musk's X is charging for access to data.

Second, we are short of human computing resources. In short, I need a data scientist with graph computing skills to move things forward!

I would also welcome input from those with relevant socio-cultural expertise.

For those interested in collaborating, I can offer expertise in the background philosophical theory, as well as a powerful computational framework for simulations.

FUTURE WORK

Plans for future work in relation to the PolyGraphs project include:

more sophisticated models of rational agents (e.g. in terms of Bayes nets, or automated reasoners);

applications in other areas (e.g. disinformation and democracy, networks of climate disinformation, organizational structures and information flow in a business setting); and

deep graph learning on simulation data (e.g. to better understand group belief).