Number 116 January 2019

ERCIM NEWS

www.ercim.eu

Special theme:

Transparency in Algorithmic Decision Making

Research and Society:
Ethics in Research

Public Opinion and Algorithmic Bias

by Alina Sîrbu (University of Pisa), Fosca Giannotti (ISTI-CNR), Dino Pedreschi (University of Pisa) and János Kertész (Central European University)

Does the use of online platforms to share opinions contribute to the polarization of the public debate? An answer from a modelling perspective.

The process of formation of opinions in a society is complex and depends on a multitude of factors. Some relate to personal preferences, culture or education. Interaction with peers is also relevant: we discuss important issues with friends and change or reinforce our opinions daily. Another significant effect is that from the media: external information reaches and influences us constantly. The choice of the people we interact with, and of the news we read, is thus crucial in the formation of our opinions. In the last decade, the patterns of interaction with peers, and of news consumption, have changed dramatically. While previously one would read the local newspaper and discuss with close friends and neighbours, nowadays people can interact at large distances and read news across the world through online media. Social media in particular is increasingly used to share opinions, but also, as the Reuters 2018 Digital News Report shows, to read and share news. This means that the peer and external effects in the dynamics of opinions are becoming susceptible to influences from the design of the social media platforms in use. These platforms are built with a marketing target in mind: to maximise the number of users and the time they spend on the platform. To achieve this, the information that reaches the users is not randomly selected. A so called 'algorithmic bias' exists: we see the news related to topics that we like and the opinions of friends that are close to our opinions, so that we are driven to read them and come back to the platform. However, a question arises: does that interfere in any way with the formation of our opinions? Do we ever change our minds any more, or we just keep reinforcing our positions? Do we ever see things in a different perspective, or we are now closed in our little information bubbles?

A group of researchers from the Knowledge Discovery and Data Mining Laboratory in Pisa, and the Department of Network and Data Science of the Central European University, funded by the European Commission, are trying to answer some of these questions by building models of opinion dynamics that mimic formation of opinions in society. Based on the evolution of opinions, a group of people can reach consensus, i.e. agreement on a certain topic, or fragmentation and polarisation of society can emerge. This process can be modelled by representing opinions with continuous numbers, and simulating interactions with specific rules to change opinions at regular intervals. The population typically starts from a random configuration and evolves in time to either consensus or a fragmented state. One very popular model for such simulations is the 'bounded confidence' model, where peers interact only if their opinion is close enough. In this model, clusters of opinion appear if the confidence is low, while for large confidence consensus emerges. This model has been modified to include algorithmic bias. Instead of selecting peers to interact with in a random way, they are selected with a bias: a person is more likely to choose to interact with a peer that has an opinion close to their

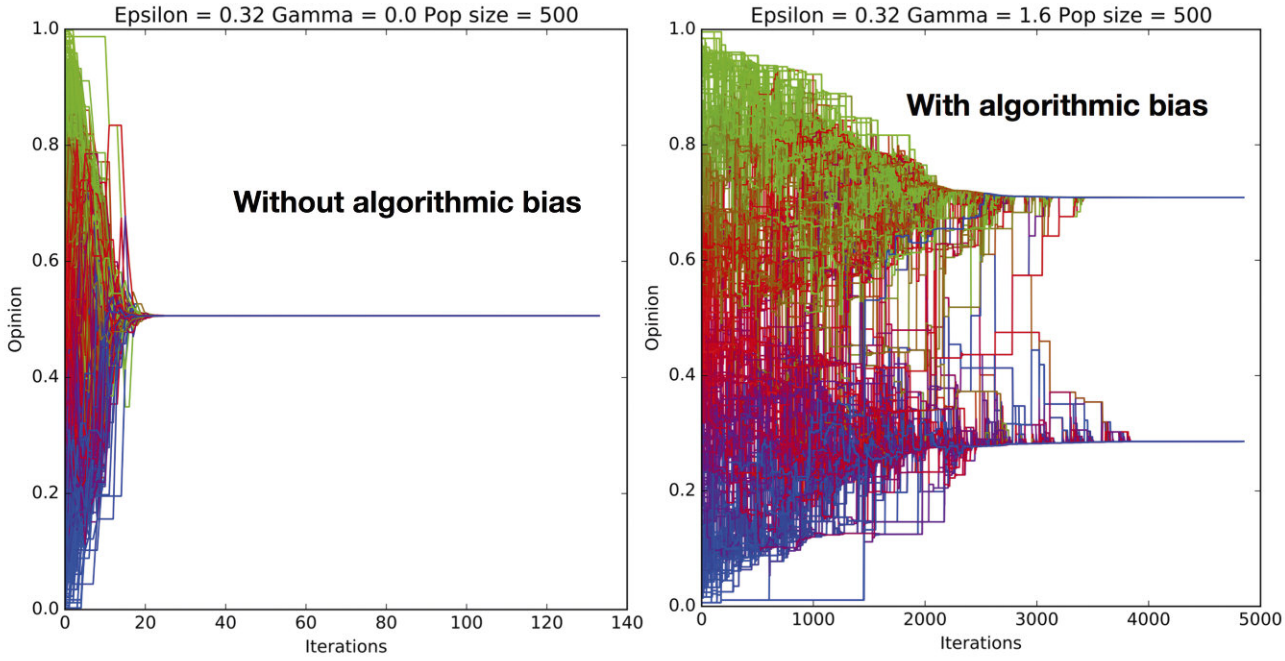

Figure 1: Simulation of opinion formation with and without the bias: The bias slows down the process and leads to the formation of two clusters instead of one.

own, while it will have a low probability of interaction with opinions far from their own.

Simulations of the algorithmic bias model show several results that suggest that online platforms can have important effect on opinion formation and consensus in society. First, the number of opinion clusters grows when algorithmic bias grows (see illustration). This means that online platforms can favour fragmentation of opinions. Second, this leads also to polarisation, where the distance between the opinions of the people is larger compared to the situation without algorithmic bias. Third, the changes in opinion are much slower when the bias is in operation. Even when consensus is obtained, the time to reach it becomes very long. In practice, this means that it could take years for people to agree on an issue, being in a highly fragmented state while this occurs.

These results bring important evidence that algorithmic bias may affect outcomes of public debates and consensus in society. Thus, we believe measures are required to at least stop its effects, if not reverse them. Researchers are investigating means of promoting consensus to counteract for the algorithmic bias effects. In the meantime, users could be informed of the way platforms feed information and the fact that this could affect their opinions, and maybe the mechanisms implemented by the platforms could be slowly withdrawn.

Reference:
[1] Alina Sîrbu, et al.: "Algorithmic bias amplifies opinion polarization: A bounded confidence model", arXiv preprint arXiv:1803.02111, 2018. https://arxiv.org/abs/1803.02111

Please contact:
Alina Sîrbu,
University of Pisa, Italy
alina.sirbu@unipi.it

Detecting Adversarial Inputs by Looking in the Black Box

by Fabio Carrara, Fabrizio Falchi, Giuseppe Amato (ISTI-CNR), Rudy Becarelli and Roberto Caldelli (CNIT Research Unit at MICC – University of Florence)

The astonishing and cryptic effectiveness of Deep Neural Networks comes with the critical vulnerability to adversarial inputs — samples maliciously crafted to confuse and hinder machine learning models. Insights into the internal representations learned by deep models can help to explain their decisions and estimate their confidence, which can enable us to trace, characterise, and filter out adversarial attacks.

Machine learning and deep learning are pervading the application space in many directions. The ability of Deep Neural Network (DNN) to learn an optimised hierarchy of representations of the input has been proven in many sophisticated tasks, such as computer vision, natural language processing and automatic speech recognition. As a consequence, deep learning methodologies are increasingly tested in security- (e.g. malware detection, content moderation, biometric access control) and safety-aware (e.g. autonomous driving vehicles, medical diagnostics) applications in which their performance plays a critical role.

However, one of the main roadblocks to their adoption in these stringent contexts is the diffuse difficulty to ground the decision the model is taking. The phenomenon of adversarial inputs is a striking example of this problem. Adversarial inputs are carefully crafted samples (generated by an adversary — thus the name) that look authentic to human inspection, but cause the targeted model to misbehave (see Figure 1). Although they resemble legitimate inputs, the high non-linearity of DNNs permits maliciously added perturbations to steer at will the decisions the model takes without being noticed. Moreover, the generation of these malicious samples does not require a complete knowledge of the attacked system and is often efficient. This exposes systems with machine learning technologies to potential security threats.

Many techniques for increasing the model's robustness or removing the adversarial perturbations have been developed, but unfortunately, only a few provide effective countermeasures for specific attacks, while no or marginal mitigations exist for stronger attack models. Improving the explainability of models and getting deeper insights into their internals are fundamental steps toward effective defensive mechanisms for adversarial inputs and machine learning security in general.

To this end, in a joint effort between the AIMIR Research Group of ISTI-CNR and the CNIT Research Unit at MICC (University of Florence), we analysed the internal representations learned by deep neural networks and their evolution throughout the network when adversarial attacks are performed. Opening the "black box" permitted us to characterise the trace left in the activations throughout the layers of the network and discern adversarial inputs among authentic ones.

We recently proposed solutions for the detection of adversarial inputs in the context of large-scale image recognition with deep neural networks. The rationale of our approaches is to attach to each prediction of the model an authenticity score estimating how much the internal representations differ from expected ones (represented by the