Title of the PhD Proposal: A Model for Reasoning About Human - Bot Interactions in Social Networks.

Name(s) of PhD Advisor(s): Frank Valencia (Supervisor) and Jean-Claude Dreher (Co-supervisor)

Host Laboratory: LIX, Ecole Polytechnique

Short abstract:

Recent behavioral experimental studies have revealed a feedback loop where human–AI interactions alter processes underlying human perceptual, emotional and social judgements, subsequently amplifying biases in humans [Glickman 2025]. This amplification is significantly greater than that observed in interactions between humans, due to both the tendency of AI systems to amplify biases and the way humans perceive AI systems. This multidisciplinary PhD project, involving multi-agent modelling and social science, aims to extend a model of opinion formation under cognitive bias by the PhD supervisor Frank Valencia [Alvim et al., 2024] to explicitly incorporate AI-driven social bots (i.e., *automated AI agents that participate in discussions*) and analyze their role in bias amplification within social networks. The goal is to develop a mathematical framework to simulate and analyze how interactions with intelligent bots contribute to bias reinforcement, test the model using real-world social network data, and design and evaluate strategies to mitigate and reduce bias reinforcement by social bots.

This PhD proposal aligns primarily with PC3 MATCHING and also with PC4 CONGRATS. It is a multidisciplinary approach combining multi-agent models and social science and the co-supervision by Frank Valencia and Jean-Claude Dreher will allow us to combine their complementary expertise in computational modeling and group-behaviour experimentation. Furthemore, the strategies to mitigate bias reinforcement using social bots address a collaborative aspect in social networks: How can social bots be used to reduce the impact of cognitive biases, rather than amplify these biases, thus allowing for a more rational civil discourse? The modelling aspects of users with cognitive biases contrasting opinions (i.e., competing about a given subject) while interacting with intelligent bots also relates to PC3 MATCHING Themes 2 and 3.

Short description of hosting research group / lab:

The Computer Science Laboratory of the École Polytechnique (LIX) is a joint research unit (UMR 7161) under two supervisory authorities, the École Polytechnique, a member of the Paris Institute of Technology, and the National Center for Scientific Research (CNRS), as well as a partner, the Inria Saclay Ile-de-France center. Within the CNRS, LIX is part of the National Institute for Information Sciences and their Interactions (INS2I) and within the Paris Institute of Technology, it belongs to the Department of Computer Science, Data, Artificial Intelligence. Frank Valencia is a permanent member of the INRIA group COMETE where he leads one of the three main research axes: *Bias and polarization in social networks*. He is also responsible for the scientific pole on *Data Analytics and Machine Learning Research* at LIX.

Table of Contents

Description of the PhD proposal (3 pages max)

Context (and scenarios if any)

Social networks have had a strong impact in opinion formation and amplification of cognitive biases [Rizou 2018]. Such impact is not only caused by interaction between humans but also by the *interaction between humans and intelligent software agents* known as social bots [Aldayel 2022]. These (social intelligent) bots are used for customer service but also to persuade people, boost social media engagement as well as for algorithmic curation, algorithmic radicalization. This may amplify cognitive, in particular confirmation bias, by selectively reinforcing existing beliefs. Therefore, while these systems can enhance user experience, they may also reinforce cognitive biases thus impacting critical and rational thinking.

This multidisciplinary PhD project, involving computer and social science, aims to extend a model of opinion formation under cognitive bias by the PhD supervisor Dr Valencia [Alvim et al., 2024] to explicitly incorporate social bots and analyze their role in bias amplification within social networks. The goal is to: (1) Develop a mathematical framework to simulate and analyze how interactions with intelligent bots contribute to bias reinforcement, (2) Test the model using real-world social network data (3) Design and evaluate strategies to mitigate and reduce bias reinforcement by social bots.

Problem and Objective

The increasing presence of intelligent social bots in social networks may affect opinion formation and reinforce cognitive biases [Aldayel 2022]. While these bots are designed to enhance user engagement, curate content, and facilitate interactions, they also introduce significant risks:

- Manipulation of Opinions: Intelligent bots can selectively reinforce certain viewpoints, making users more susceptible to biased or algorithmically-driven narratives, reducing exposure to diverse, independent reasoning.
- Cognitive Bias Reinforcement: By personalizing content and interactions, bots can amplify cognitive biases (e.g., confirmation bias, authority bias, back-fire effect, and groupthink), leading to rigid belief systems that resist counter-evidence.

Existing opinion dynamics models—such as DeGroot, Bounded Confidence, and Deffuant-Weisbuch—do not adequately capture the long-term cognitive effects of AI-driven interventions on cognitive bias and critical independent thinking. There is a lack of mathematical models that formally represent the impact of intelligent bots on bias reinforcement. A generalization of the classic DeGroot model recently proposed by the author of this proposal [Alvim et al., 2024], here referred to as Bias-DeGroot, provides a more nuanced representation of opinion formation, where each agent has their own cognitive biases. This tractable model offers theoretical results on consensus and opinion evolution in the presence of biases, which play a crucial role in shaping opinions within social networks.

Objective: This project aims to generalize the Bias-DeGroot model to reason about human-bot interactions and their role in amplifying cognitive biases. The model will allow us to study under what conditions social bots may lead to bias reinforcement and to design strategies to reduce such deviation from rationality. We expect that the strategies will also lead us to answer how social bots can help reduce cognitive biases, rather than amplify them, thus allowing for more rational civil discourse.

Brief overview of the state of the art.

Humans are known to be subject to a number of cognitive biases and AI systems can also exhibit biased judgements in domains ranging from perception to emotion. Recent studies have revealed a feedback loop where human–AI interactions alter processes underlying human perceptual, emotional and social judgements, subsequently amplifying biases in humans [Glickman 2022]. This amplification is significantly greater than that observed in interactions between humans, due to both the tendency of AI systems to amplify biases and the way humans perceive AI systems. The work [Aldayel 2022] shows that while social bots engage less frequently with users than influential figures, they still play a role in shaping opinions.

None of the work above conducts their studies with opinion models. There is an increasing interest in studying the effect of bots using opinion models. The work [Siahkali 2024] explores user-bot interactions in social networks using the Stochastic Bounded Confidence Model (SBCM) to study opinion shaping through agent-controlled bots and targeted advertising. By integrating Deep Deterministic Policy Gradient (DDPG) and Deep Reinforcement Learning (DRL), the study demonstrates that these strategies can efficiently influence opinions and optimize advertising resource allocation on social platforms. In [Vega-Oliveros 2021] introduces a variant of the standard voter model to characterize the number of bots needed to achieve specific opinion-shaping targets as a function of various system parameters in a fully connected network. In [Keijzer 2021], the authors extend Axelrod's cultural dissemination model to incorporate bots and evaluate the impact of network structure on opinion dynamics through simulations. The models in [Luo 2020] shows that with the increase of the number of social bots, the public opinion is lead out of balance more severely, and as the proportion of social bots reaches 20%, the public opinion of human agents are mislead, and the opinion held by social bots will be the final opinion in the human population. The model in [Ross 2019] indicates that, in a highly polarised setting, depending on their network position and the overall network density, bots' participation by as little as 2–4% of a communication network can be sufficient to tip over the opinion climate in two out of three cases. These findings demonstrate a mechanism by which bots could shape the norms adopted by social media users.

The above models do not include a mechanism to specify cognitive bias thus they cannot adequately capture the long-term cognitive effects of bots' AI-driven interventions on cognitive bias and critical independent thinking.

Research Questions

The first research question focuses on model design:

- How can the Bias-DeGroot model be extended to formally reason about opinion manipulation by social bots?
  What modifications are necessary to represent bots as agents with their own influence strategies?
  What are the mechanisms through which bots selectively reinforce cognitive biases (e.g., confirmation bias, authority bias)?

Once the model is defined, it can be used to address key questions regarding collaboration of human-bot interactions in the model, such as:

- How can social bots be used to reduce the effect of cognitive biases, rather than amplify these biases, thus allowing for more rational civil discourse?

Theoretical foundations

Complex social interactions and collective behavior in social networks have been studied using mathematical models of social learning that aim to capture opinion dynamics [Wasserman 94, Golub 2010, Acemoglu 2011, Alvim 2019, 2023-2024, Aranda 2024, Paz 2024]. This proposal is founded upon models for complex social interactions and collective behavior in social networks. The classic DeGroot multi-agent model [DeGroot 1973] is one of the most prominent formalisms for opinion formation dynamics in social networks. In the DeGroot model, a social network is represented as a weighted directed graph, where the edges indicate how much individuals (called agents) influence one another. Each agent i has an opinion (or belief) B_i, represented as a value in the interval [0,1], indicating the strength of their agreement with an underlying proposition (e.g., "vaccines are safe"); the higher the value, the stronger the agreement. Agents repeatedly update their opinions using the weighted average of their opinion differences (level of disagreement) with those who influence them (e.g., their contacts, neighbors, or friends).

In recent work [Alvim 2024], we introduced a generalization of the DeGroot that focuses on arbitrary biases on disagreement. Each agent i now has their own individual cognitive biases B(i,j) on levels of disagreement with another agent j: i.e., they express how agent i reacts to a disagreement with j. These biases are represented as arbitrary functions on opinion differences in the unit square region and they are classified into four sub-regions: M,R,B,I based on the cognitive reactions they may cause in an agent during instances of opinion disagreement. Agents that are malleable, easily swayed, exhibit fanaticism or prompt to follow authoritative figures can be modeled with biases in M. Agents that are receptive to other opinions, but unlike malleable ones, can exhibit some skepticism to fully accepting them can be modeled with biases in the region R. Individuals that become more extreme when confronted with opposing opinions can be modeled by biases in B. Finally insular, stubborn agents can be modeled with the bias in I. A noteworthy theoretical result we obtained for the Bias-DeGroot model states that assuming that each bias of every agent is a continuous function in the region R, the society converges to a consensus if that society is strongly connected [Alvim 2024].

By extending the Bias DeGroot with a principled approach to modeling the adaptive behavior of intelligent bots, we plan to provide a robust framework for analyzing and simulating how intelligent bots can manipulate opinions, reinforce biases, and potentially degrade expertise in social networks.

Approach and Methods

The project will be executed in three main phases—model development, simulation and analysis, and empirical validation and mitigation strategies

1. **Model Development:** Develop a mathematical extension of the Bias-DeGroot model to formally incorporate intelligent bots as agents with their own influence strategies. This will involve:

■ Integrating Bots Dynamics: Introduce additional parameters and update rules that represent bot behavior (e.g., reinforcement learning–driven adaptations, targeted messaging, and bias amplification).

■ Cognitive Bias Representation: Formalize how cognitive biases (e.g., confirmation bias, authority bias) affect the opinion update process of human agents when interacting with bots.

2. **Simulation and Analysis**: Construct a simulation framework to model a social network composed of both humans and bot agents. We will simulate various network structures (e.g., echo chambers, heterogeneous communities) to observe differences in opinion dynamics and implement different bot strategies (e.g., reinforcement learning, targeted messaging) and study their impact on opinion formation and cognitive bias over time.
3. **Empirical Validation and Mitigation Strategies**: We validate our simulation results using social media datasets and survey data on cognitive biases and opinion polarization. We will explore modifications to AI-driven systems (e.g., incorporating diversity-promoting algorithms, adaptive trust mechanisms) to mitigate user bias.

This multi-phased approach—spanning model innovation, rigorous simulation analysis, and empirical validation—will provide a robust framework to understand and mitigate the adverse effects of AI-driven opinion manipulation in social networks.

Evaluation of the contributions

The contributions of this project will be evaluated through a combination of theoretical analysis, simulation experiments, and empirical validation, ensuring that the proposed model not only advances theoretical understanding but also offers practical insights into reducing bias amplification in human-bots interactions. The evaluation aspects include:

1. Theoretical Analysis: Assess the mathematical robustness of the extended Bias-DeGroot model by proving key properties such as consensus, stability in the presence of social bots.
2. Simulation-Based Evaluation: Implement large-scale simulations across various network topologies (e.g., echo chambers, heterogeneous communities) to measure how well the model replicates observed social phenomena. Systematically vary model parameters (e.g., bots influence strength, bias intensity, network connectivity) to quantify their impact on opinion diversity and degradation of critical thinking.
3. Empirical Validation: Validate simulation outcomes using real-world data from social media platforms and survey data on opinion dynamics, cognitive biases, and expertise levels under the presence of social bots.

The above evaluation strategy will ensure that the extended Bias-DeGroot model accurately reflects the complexities of AI-driven opinion manipulation, sheds light on the erosion of critical thinking, and explores practical ways to counter its negative social impact.

Nature of Digital Collaboration

Digital collaboration has transformed the way people interact, exchange ideas, and form opinions, especially in online spaces. Social networks, recommendation systems, and AI-driven tools have made it easier to connect, but they also shape the way information spreads and influence how people engage with different perspectives. One of the biggest changes in digital collaboration is the increasing role of intelligent social bots—automated agents that participate in discussions, amplify content, and sometimes manipulate opinions. While these bots can enhance engagement and streamline communication, they also raise concerns about bias reinforcement. Many AI-driven systems tailor content to individual users, creating feedback loops where people are repeatedly exposed to viewpoints that align with their existing beliefs. This can limit critical thinking and reduce opinion diversity over time. At the same time, digital collaboration offers opportunities to design AI systems that counteract bias rather than reinforce it. This project explores how human-bot interactions in social networks influence opinion formation and whether social bots can be used to promote rational discourse rather than deepen ideological divides. By combining multi-agent modeling, cognitive science, and strategies to mitigate the effect of bias on opinion formation, this work seeks to better understand how AI can support more constructive and informed digital collaboration.

Social networks involve large-scale interactions where thousands of users engage with both human peers and AI-driven social bots. These interactions shape opinion dynamics, cognitive biases, and the spread of information, often in complex and unpredictable ways. Given the role of social bots in influencing discussions, filtering content, and reinforcing narratives, this PhD will focus on understanding the group-level effects of human-bot interactions and their impact on bias amplification and rational discourse.

The research will specifically target asynchronous communication and influence dynamics, which are characteristic of interactions in social media platforms, online forums, and algorithmically curated spaces. Unlike face-to-face discussions, where feedback is immediate, online exchanges occur over extended time frames, allowing biases to gradually shape discourse and opinion shifts. Social bots, through content recommendations, engagement incentives, and personalized interventions, can significantly impact how information is perceived and shared, leading to long-term changes in public opinion.

Given the large size of the groups involved (ranging from hundreds to thousands of users) and the distributed, remote nature of interactions, the evolution of opinions and bias reinforcement mechanisms will be studied over an extended period, typically months. The project will examine how continuous exposure to AI-driven interventions affects belief formation, polarization, and expertise degradation in realistic, large-scale network settings. By simulating these dynamics using multi-agent models, the research aims to provide insights into the conditions under which bias is reinforced or mitigated and offer potential strategies for designing social bots that promote rational, balanced discourse rather than deepening ideological divisions.

Contribution to digital collaboration: Expected results and Impact (1 page max)

This PhD project will advance the understanding of human-AI interactions in digital collaboration, particularly in opinion formation, bias reinforcement, and strategies for promoting rational discourse. By extending the Bias-DeGroot model to incorporate social bots, the research will provide a formal framework to analyze how AI-driven agents influence collective decision-making in social networks. This work will help establish conditions under which AI systems reinforce cognitive biases versus when they contribute to more balanced discussions.

The main expected result is the development of a formal model of human-bot opinion dynamics, capturing how interactions with social bots affect bias amplification. This model will allow for simulation-based analysis, revealing how factors such as network topology, bot influence strategies, and user susceptibility to bias shape digital discourse.

Additionally, empirical insights gained from real-world social network datasets will help validate the model's predictions, offering a deeper understanding of how AI-driven interventions impact long-term belief evolution. Beyond theoretical contributions, the project aims to develop bias-reducing interventions using social bots. These strategies will be tested through simulations and data-driven experiments to assess their effectiveness in counteracting bias reinforcement and fostering rational discussions. By integrating AI-driven solutions into digital collaboration, this research will propose ways to design social bots that enhance, rather than degrade, cognitive autonomy in social networks' interactions.

Positioning in the eNSEMBLE program (½ page max).

This PhD project closely aligns with PC3 MATCHING and also has strong connections to PC4 CONGRATS.

Within PC3 MATCHING, Theme 3 focuses on the long-term effects of intelligent systems on human expertise, critical thinking, and decision-making. This project directly contributes to this theme by modeling how social bots influence cognitive biases in online discussions. Prolonged exposure to AI-driven interactions can amplify biases, reduce critical thinking, and shape belief formation, raising concerns about expertise degradation and opinion manipulation. By extending the Bias-DeGroot model, this research provides a mathematical framework to analyze these effects and determine the conditions under which social bots either support informed decision-making or contribute to cognitive stagnation and polarization. Additionally, this project proposes bias-reducing strategies, exploring how AI-driven interventions can be designed to promote balanced discourse rather than reinforce ideological silos. By investigating both the risks and opportunities of AI in opinion formation, this research contributes to defining the desirable conditions for intelligent system deployment in social networks.

Furthermore, this project aligns with Theme 2 of PC3, which focuses on the dynamics of human-AI collaboration and competition, particularly in environments where users interact with intelligent systems. This research contributes by studying how social bots function as either collaborative agents, promoting rational discourse, or competitive agents, reinforcing biases and polarization. By extending multi-agent models to incorporate adaptive social bots, this study examines how AI intervention strategies influence group opinion dynamics and long-term interactions in digital environments.

By addressing the impact of AI on cognitive biases and opinion formation, this project directly supports PC3 MATCHING's mission to understand how intelligent systems shape human collaboration. The interdisciplinary approach—combining multi-agent modeling, cognitive science, and AI-driven interventions—positions this research as a valuable contribution to the study of human-AI interactions and their societal implications.

Finally, since this PhD proposal focuses on human-bot interactions in social networks, it is also highly relevant to PC4 CONGRATS, which addresses large-scale community collaboration via digital platforms. PC4 aims to develop tools for understanding community interactions, aligning with this project's primary objective: to develop a model for reasoning about human-bot interactions in social networks. By providing a mathematical and computational framework for analyzing these interactions, this research directly supports PC4's goal of improving digital collaboration and community dynamics.

References

(Highlighted in yellow are the publications by Frank Valencia related to this proposal)

1. Alvim M, Amorim B, Knight S, Quintero S, Valencia F: A Formal Model for Polarization under Confirmation Bias in Social Networks. Log. Methods Comput. Sci. 19(1) (2023)
2. Alvim M, Knight S, Valencia F: Toward a Formal Model for Group Polarization in Social Networks. The Art of Modelling Computational Systems 2019: 419-441(2019).
3. Alvim M,Gaspar da Silva A, Knight Sophia, Valencia, F: A Multi-agent Model for Opinion Evolution in Social Networks Under Cognitive Biases. FORTE 2024: 3-19 (2024)
4. Aranda J, Betancourt S, Díaz JF, Valencia Frank: Fairness and Consensus in an Asynchronous Opinion Model for Social Networks. CONCUR 2024: 7:1-7:17 (2024)
5. Paz J, Rocha C, Tobòn L, and Valencia Frank. Consensus in Models for Opinion Dynamics with Generalized-Bias. COMPLEX NETWORKS 2024. To Appear (2024)
6. Wasserman, S., & Faust, K.. Social network analysis in the social and behavioral sciences. In *Social Network Analysis: Methods and Applications* (pp. 1–27). Cambridge University Press. ISBN: 9780521387071 (1994).
7. Acemoglu, D., Ozdaglar, A. Opinion Dynamics and Learning in Social Networks. *Dyn Games Appl* **1**, 3–49 (2011).
8. Golub, Benjamin, and Matthew O. Jackson. "Naïve Learning in Social Networks and the Wisdom of Crowds." *American Economic Journal: Microeconomics*, 2 (1): 112–49. (2010)
9. Aldayel, A., Magdy, W. (2022). Characterizing the role of bots' in polarized stance on social media. *Social Network Analysis and Mining, 12*(1), 30.
10. Rizoiu, M.-A., Graham, T., Zhang, R., Zhang, Y., Ackland, R., & Xie, L. (2018). #DebateNight: The Role and Influence of Socialbots on Twitter During the 1st 2016 U.S. Presidential Debate. *arXiv:1802.09808*.
11. Glickman, M., Sharot, T. (2025). How human–AI feedback loops alter human perceptual, emotional and social judgements. *Nature Human Behaviour, 9*, 345–359.
12. Siahkali, F., Samadi, S., & Kebriaei, H. (2024). Towards Opinion Shaping: A Deep Reinforcement Learning Approach in Bot-User Interactions. *arXiv:2409.11426*.
13. Vega-Oliveros, D.A., Grande, H.L.C., Iannelli, F., et al. (2021). Bi-layer voter model: modeling intolerant/tolerant positions and bots in opinion dynamics. *Eur. Phys. J. Spec. Top. 230*, 2875–2886.
14. Keijzer, M. A., & Mäs, M. (2021). The strength of weak bots. *Online Social Networks and Media, 21*, 100106.
15. Luo, Y., Cheng, C., & Yu, C. (2020). Discrete Opinion Dynamics with Social Bots on Signed Network. *2020 39th Chinese Control Conference (CCC)*, Shenyang, China, 6690-6694.
16. Ross, B., Pilz, L., Cabrera, B., Brachten, F., Neubaum, G., & Stieglitz, S. (2019). Are social bots a real threat? An agent-based model of the spiral of silence to analyse the impact of manipulative actors in social networks. *European Journal of Information Systems, 28*(4), 394–412.