Constructing and deconstructing bias: modeling privilege and mentorship in agent-based simulations

Andria L. Smith* **(asmith@is.mpg.de)**
Max Planck Institute for Intelligent Systems, Heisenbergstr. 3
Stuttgart, Baden-Württemberg 70569 Germany

Simon Heuschkel* **(simon.heuschkel@student.uni-tuebingen.de)**
University of Tübingen, Maria-von-Linden-Str. 6
Tübingen, Baden-Württemberg 72076 Germany

Ksenia Keplinger (kkeplinger@is.mpg.de)
Max Planck Institute for Intelligent Systems, Heisenbergstr. 3
Stuttgart, Baden-Württemberg 70569 Germany

Charley M. Wu (charley.wu@uni-tuebingen.de)
University of Tübingen, Maria-von-Linden-Str. 6
Tübingen, Baden-Württemberg 72076 Germany

* These authors contributed equally

Abstract

Bias exists in how we pick leaders, who we perceive as being influential, and who we interact with, not only in society, but in organizational contexts. Drawing from leadership emergence and social influence theories, we investigate potential interventions that support diverse leaders. Using agent-based simulations, we model a collective search process on a fitness landscape. Agents combine individual and social learning, and are represented as a feature vector blending relevant (e.g., individual learning characteristics) and irrelevant (e.g., race or gender) features. Agents use rational principles of learning to estimate feature weights on the basis of performance predictions, which are used to dynamically define social influence in their network. We show how biases arise based on historic privilege, but can be drastically reduced through the use of an intervention (e.g. mentorship). This work provides important insights into the cognitive mechanisms underlying bias construction and deconstruction, while pointing towards real-world interventions to be tested in future empirical work.

Keywords: social influence; bias; privilege; social network; intervention

Introduction

Bias exists in how we pick leaders, who we are influenced by, and who we interact with. For instance, there are more CEOs named John or David than women CEOs in the S&P 1500 companies (Johnson, Hekman, & Chan, 2016). Despite increased interest in creating more diverse and inclusive organizational environments, there are many barriers in place, such as biases, preventing progress (Keplinger & Smith, 2022).

Although empirical research on the sources of biases and potential interventions for unlearning biases has a long tradition (Axelrod, 1997; Freeman, Penner, Saperstein, Scheutz, & Ambady, 2011; Serban et al., 2015; Schelling, 1971), it is still unclear when, why, and how privilege and bias arise (Colella, Hebl, & King, 2017). Specifically, we are interested in how privilege, defined as unearned access to rewards and resources for specific groups (Case, Iuzzini, & Hopkins, 2012; Crevani, 2019), hinders the emergence of marginalized leaders (Badura, Galvin, & Lee, 2022). Here, to integrate theories on leader emergence and social influence, we use agent-based simulations which are a computational approach still rarely applied to the leadership context (but see Cao et al., 2020). These simulations add precision to previous verbal theories (Samuelson et al., in press; Vancouver, Wang, & Li, 2020) and shed light on the cognitive mechanisms underlying the learning of biases towards arbitrary agent features (e.g., race, gender, age, etc.).

This study aims to 1) demonstrate how biases are recreated through rational principles of multi-agent learning when certain agents are placed in privileged locations in the environment and 2) investigate the impact of an intervention, where we create temporary social network connections between high and low performing agents modeled as external agents (e.g., mentors or role models), to unlearn the bias. Our simulation shows how we can systematically reduce bias across all agents, thus leading to an increase in diverse representation of emergent leaders.

Methods

We use agent-based simulations, where a team of 7 agents collectively optimize a two-dimensional fitness landscape with multiple local optima. We use the Ackley and Drop Wave environments as common test functions for optimization algorithms (Surjanovic & Bingham, 2013), in addition to the Mason and Watts (2012) environment, with previous work characterizing these environments as having similar average payoff and

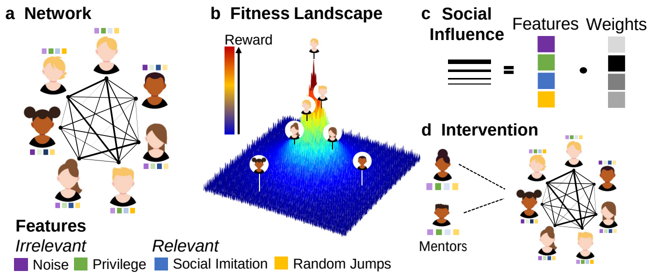

Figure 1: Agent-based simulations. **a**) Network structure and agent features. **b**) Fitness landscape. **c**) Social influence learning as function of learned feature weights. **d**) Intervention through mentorship.

number of local optima (Barkoczi, Analytis, & Wu, 2016). We define each landscape with $1000 \cdot 1000$ discrete locations.

Agents are connected by a weighted social influence matrix A and defined by a set of features (Fig. 1a). On each iteration, agents update their position on the fitness landscape $\mathbf{x}_i$ and the according fitness value y_i using a search policy combining individual and social learning (Fig. 1b), and then update social influence based on learned feature weights W (Fig. 1c).

Social influence. Each agent i is defined by a set of features $\mathbf{f}_i = [\gamma, \eta, \rho, \nu]$ representing personal attributes that are *policy relevant* (i.e., capturing characteristics of learning strategy: γ, η), indicate *privilege* (i.e., starting condition ρ) or are completely *irrelevant* (i.e., noise ν).

At every iteration each agent tries to estimate the performance of other agents j with a linear weighted sum of their features $\hat{r}_{i,j} = \mathbf{w}_i^\top \mathbf{f}_j + \varepsilon$, where $\varepsilon = \min_j r_j$ is an offset for the lowest current reward across all agents in the group. The performance of other agents is measured using temporally discounted past rewards relative to the current time point T: $r_{i,T} = (\sum_{t=0}^{T} y_{i,t} \cdot \lambda^{T-t}) / \sum_{t=0}^{T} \lambda^t$, where we set the temporal discount $\lambda = 0.9$. Weights are updated by minimizing the mean squared prediction error between the actual rewards and their predictions through gradient descent:

$$\mathbf{w}_i \leftarrow \mathbf{w}_i - \alpha \cdot \frac{\partial}{\partial \mathbf{w}_i} \mathcal{L}_{MSE}(\mathbf{r}, \hat{\mathbf{r}}_i) \tag{1}$$

with learning rate $\alpha = 0.1$.

Thus, the learning of weights captures feature-specific biases of social influence, where agents with highly weighted features will exert more social influence. We update the social influence matrix every iteration as a function of predicted performance: $A \leftarrow A + \beta \cdot W \cdot F^\top$ with $\beta = 0.5$ controlling the update rate.

Individual and social learning. Agents use a combination of social and individual learning policies to optimize their position in the fitness landscape. Each agent i first uses social imitation with probability $P(\gamma_i)$, whereby it uses a softmax imitation policy as a function of social influence weights:

$$\pi_{\text{imitation}}(\mathbf{x}_j) \propto exp(\frac{a_{i,j}}{\sum_k a_{i,k}}/\tau) \tag{2}$$

Intuitively, agents are more likely to imitate others with higher perceived influence $a_{i,j}$.

If the social policy is not enacted, with probability $1 - P(\gamma_i)$, the agent uses individual learning. First, the agent tries a random jump with a probability of $P(\eta_i)$, whereby it evaluates a random position in the landscape and jumps there if it increases its fitness value. If the agent does not use a random jump, it performs local optimization using stochastic hill climbing (SHC) over all neighbouring positions:

$$\pi_{\text{SHC}}(\mathbf{x}') \propto \exp(y_i'/\tau), \tag{3}$$

where each y_i is the fitness value of a neighboring solution $\mathbf{x}'$, and with higher-valued solutions more likely to be selected. In all cases, we set $\tau = 0.01$.

Policy relevant features γ and η are sampled uniformly from $\mathcal{U}(0, 0.1)$, whereas privilege ρ and noise ν are sampled from a multimodal Gaussian with two different means $\in (0.03, 0.07)$, each with variance of 0.01, and truncated between $[0, 0.1]$. Both policy relevant features influence performance by aiding in escaping local optima, but γ depends on the quality of social information.

Privilege and intervention. To model *privilege*, we use ρ to define the starting position of an agent, by placing them in a location within the top $(\rho \times 10)$-th quantile ± 0.005 of rewards. This initialization leads to a positive correlation between privilege and performance at the beginning of the simulation (e.g., $\rho = 0.5$ will start near the median reward in the landscape), but has no bearing on learning capabilities.

We hypothesize that agents will develop a bias towards learning large feature weights for privilege. Therefore, we introduce an intervention in form of *mentorship*. Mentors are modeled as agents who have less privilege $\rho \sim \mathcal{N}(0.03, 0.01)$, but have high policy relevant traits $\eta, \gamma \sim \mathcal{N}(0.08, .005) \in [0, 0.1]$ and high performance $r_{\text{mentor}} > 90\%$ of all fitness values. At each iteration, a less privileged agent ($\rho < .05$) gets a mentor assigned with probability 0.2. The features and r_{mentor} of this mentor are used to optimize the social weight of the mentee $\mathbf{w}_i$, additionally to the group members. Thus, mentors do not participate in the collective optimization and therefore are not targets for social imitation or part of the social influence network, but only support social feature learning (Eq. 1). However, mentors signal awareness of noise features to the group that certain biases can be broken (Freeman Jr & Kochan, 2019; Ivey & Dupré, 2022; Raza & Onyesoh, 2020; Williams et al., 2020).

Results

We ran 1000 simulations with 150 iterations per environment. Figure 2 shows that even though agents start out unbiased, by iteration 25 they learn strong weights for the privilege feature ρ. As the simulation continues, random jumps η prove to be useful and the agents learn increasingly strong weights for this feature. Social imitation in form of γ does not seem to be valued initially, with decreasing weights until after the inter-

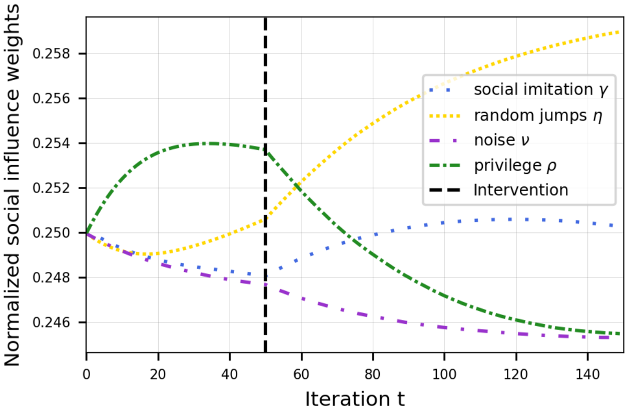

Figure 2: Results. Impact of the intervention on the mean feature weights across fitness landscapes. The normalized weights indicate how much performance is credited to each feature.

vention, which may be due to a masking effect of the strong privilege weights.

After the intervention (vertical dashed line), privilege weights decrease strongly, while social imitation weights begin to increase. This shows that mentors do not only reduce the influence of policy irrelevant features (i.e., privilege), but also help agents learn to rely more on policy relevant features γ and η. Although privilege and noise weights decrease faster after the intervention, agents still rely more on privilege ρ than noise ν, showing how difficult it is to get fully rid of a learned bias. Simulations without the intervention also result in decay in privilege weights that is much less pronounced and fails to encourage increasing weights for social imitation.

Conclusion

Our computational approach provides a tool for understanding how rational principles of learning can shape the formation of biases in which features are assigned credit for performance. We show how biases naturally arise based on historic privilege, but can be mitigated through an intervention by creating mentoring relationships between high and low performing marginalized agents. This work provides important insights into the cognitive mechanisms underlying how biases can develop and be unlearned, while pointing towards real-world interventions to be tested in future empirical work.

Acknowledgements

This work is supported by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039A and funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy–EXC2064/1–390727645.

References

Axelrod, R. (1997). The dissemination of culture: A model with local convergence and global polarization. *Journal of conflict resolution*, *41*(2), 203–226.

Badura, K. L., Galvin, B. M., & Lee, M. Y. (2022). Leadership emergence: An integrative review. *Journal of Applied Psychology*, *107*(11), 2069.

Barkoczi, D., Analytis, P. P., & Wu, C. (2016). Collective search on rugged landscapes: A cross-environmental analysis. In *Proceedings of the 38th annual conference of the cognitive science society* (pp. 918–923).

Cao, S., MacLaren, N. G., Cao, Y., Dong, Y., Sayama, H., Yammarino, F. J., ... others (2020). An agent-based model of leader emergence and leadership perception within a collective. *Complexity*, *2020*.

Case, K. A., Iuzzini, J., & Hopkins, M. (2012). Systems of privilege: Intersections, awareness, and applications. *Journal of Social Issues*, *68*(1), 1–10.

Colella, A., Hebl, M., & King, E. (2017). One hundred years of discrimination research in the journal of applied psychology: A sobering synopsis. *Journal of Applied Psychology*, *102*(3), 500.

Crevani, L. (2019). Privilege in place: How organisational practices contribute to meshing privilege in place. *Scandinavian Journal of Management*, *35*(2), 101035.

Freeman, J. B., Penner, A. M., Saperstein, A., Scheutz, M., & Ambady, N. (2011). Looking the part: Social status cues shape race perception. *PloS one*, *6*(9), e25107.

Freeman Jr, S., & Kochan, F. (2019). Exploring mentoring across gender, race, and generation in higher education: An ethnographic study. *International Journal of Mentoring and Coaching in Education*.

Ivey, G. W., & Dupré, K. E. (2022). Workplace mentorship: A critical review. *Journal of Career Development*, *49*(3), 714–729.

Johnson, S. K., Hekman, D. R., & Chan, E. T. (2016). If there's only one woman in your candidate pool, there's statistically no chance she'll be hired. *Harvard Business Review*, *26*(04), 1–7.

Keplinger, K., & Smith, A. (2022). Stigmatization of women in the workplace: Sources of stigma and its consequences at the individual, organizational and societal level. In *Diversity in action* (pp. 23–38). Emerald Publishing Limited.

Mason, W., & Watts, D. J. (2012). Collaborative learning in networks. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(3), 764–769. doi: 10.1073/pnas.1110069108

Raza, A., & Onyesoh, K. (2020). Reverse mentoring for senior nhs leaders: a new type of relationship. *Future Healthcare Journal*, *7*(1), 94.

Samuelson, H. L., Lee, J., Wessel, J. L., Grand, J. A., Samuelson, H., Lee, J., ... Grand, J. (in press). Computational modeling in organizational diversity and inclusion.

Schelling, T. C. (1971). Dynamic models of segregation. *Journal of mathematical sociology*, *1*(2), 143–186.

Serban, A., Yammarino, F. J., Dionne, S. D., Kahai, S. S., Hao, C., McHugh, K. A., ... Peterson, D. R. (2015). Leadership emergence in face-to-face and virtual teams: A multi-level model with agent-based simulations, quasiexperimental and experimental tests. *The Leadership Quarterly*, *26*(3), 402–418.

Surjanovic, S., & Bingham, D. (2013). *Virtual library of simulation experiments: Test functions and datasets.* Retrieved March 31, 2023, from http://www.sfu.ca/ ssurjano.

Vancouver, J. B., Wang, M., & Li, X. (2020). Translating informal theories into formal theories: The case of the dynamic computational model of the integrated model of work motivation. *Organizational Research Methods*, *23*(2), 238–274.

Williams, N., Ravenell, J., Duncan, A. F., Butler, M., Jean-Louis, G., & Kalet, A. (2020). Peer mentor development program: lessons learned in mentoring racial/ethnic minority faculty. *Ethnicity & Disease*, *30*(2), 321.