The emergence of discrimination due to miscategorization

M. Alperen Yasar*

Ca' Foscari University; and
Paris I Pantheon-Sorbonne University

May 14, 2024

Abstract

Purpose — This study explores the emergence of discrimination based on observable characteristics. In many instances, agents presume differences arising from traits such as race or gender, even when these parameters are irrelevant to the situation at hand. This paper intends to reveal an emergent behavior and a persistent culture of discrimination caused by miscategorization in strategic interactions.

Design/methodology/approach — We assume that agents occasionally engage in conflicts modeled as asymmetric hawk and dove games, where boundedly rational agents may categorize their opponents based on observable traits to make effective decisions. Three categorization strategies are considered: fine-grained, regular, and coarse-grained. Subsequently, an evolutionary agent-based model is employed to examine the performance of these strategies in a dynamic environment.

Findings — The results demonstrate that fine-grained categorization provides an advantage when the cost of fighting is low, while coarse-grained categorizers exhibit more peaceful behavior, gaining an advantage when the cost of conflict is high. Our primary finding indicates the emergence of discrimination based on non-relevant traits, manifested through consistent aggressive behavior towards individuals possessing these traits.

Originality/value — This paper is the first to investigate the emergence of discrimination without assuming prior differences between groups. Previous studies have assumed either an initial population difference or a homophily-based approach. In contrast, we demonstrate that discrimination can emerge even in the absence of such assumptions. Discrimination between two groups may arise as long as there are agents who label these categories.

Keywords — discrimination, asymmetric hawk and dove games, agent-based modeling, emergent behavior

Paper type — Research paper

1 Introduction

Discrimination based on observable characteristics, such as race or gender, remains a pervasive issue in various social interactions despite these traits often being irrelevant to the situation at hand. This phenomenon has been widely studied across multiple disciplines, including economics, psychology, and sociology (Bertrand and Duflo, 2017; Greenwald and Banaji, 1995; Pager and Shepherd, 2008). However, the emergence of discrimination in the absence of prior differences between groups has received limited attention.

Previous research has primarily focused on either the effects of discrimination (Becker, 2010), or how it arises from pre-existing biases between groups (Arrow, 1998; Phelps, 1972). These studies have provided valuable

*Funding and acknowledgements: This work has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 956107, "Economic Policy in Complex Environments (EPOC)". I greatly appreciate the comments of Marco LiCalzi and Paolo Pellizzari.

insights into the mechanisms perpetuating discrimination but have yet to fully address how discrimination can emerge even when there are no initial differences between the groups involved.

This paper aims to bridge this gap by exploring the emergence of discrimination based on observable characteristics in strategic interactions using an agent-based modeling approach. We simulate conflicts as asymmetric hawk and dove games, where boundedly rational agents categorize their opponents based on observable traits to make effective decisions. By introducing three categorization strategies—fine-grained, regular, and coarse-grained—we investigate how these strategies perform in a dynamic environment and their impact on the emergence of discrimination.

Our main contribution lies in demonstrating that discrimination can emerge even in the absence of prior differences between groups. We show that the mere presence of agents who label categories based on observable traits can lead to the emergence of discriminatory behavior. This finding highlights the importance of understanding the role of categorization in the formation and persistence of discrimination.

The remainder of this paper is structured as follows: Section 2 provides an overview of the relevant literature on discrimination and categorization in strategic interactions. Section 3 describes the methodology, detailing how an asymmetric hawk and dove game works. Section 4 presents the findings of our simulations, focusing on the performance of different categorization strategies and the emergence of discrimination. We present our results in Section 5 by defining a culture of discrimination and how it might arise without prior difference assumptions. Finally, Section 6 discusses the implications of our results, limitations, and avenues for future research.

2 Literature review

Discrimination in organizations has been a persistent issue, negatively impacting individuals, teams, and overall organizational performance. Understanding the factors that contribute to the emergence and perpetuation of discrimination is crucial for developing effective strategies to promote diversity, equity, and inclusion in the workplace. While various theories and approaches have been employed to study discrimination, recent research has highlighted the importance of examining the impact of categorization strategies in discriminatory behaviors (Flache and Mäs, 2008; Bruner, 2019; Stewart and Raihani, 2023).

To further understand the complex dynamics of discrimination, the review explores the contributions of agent-based modeling approaches in studying social phenomena, with a particular focus on models that investigate the emergence of discrimination (Martell *et al.*, 2012; Amadae and Watts, 2022). By examining these models, the review aims to identify the potential of agent-based modeling to provide novel insights into the processes underlying the formation and perpetuation of discriminatory behaviors in organizations.

2.1 Discrimination and categorization

The existence of discrimination in academia, corporations, and politics is now well documented (see Bertrand and Duflo (2017) for a literature review on field experiments of discrimination). Common bases for discrimination include gender, race, or political opinions. A recent paper has demonstrated that only 10% of board members consist of female executives (Brodmann *et al.*, 2022). Martell *et al.* (2012) suggest this low ratio could stem from male board members' perceptions of women's skills, which may further pressure female employees. In response, they propose new promotion policies to reduce gender discrimination in organizations. Bruner (2019) shows how being classified as a minority can decrease the payoff of a group by using replicator dynamics. Stewart and Raihani (2023), on the other hand, develop a model where stereotypes evolve over time.

Research has shown that the way individuals categorize others can have significant implications for their attitudes and behaviors toward those groups. For example, Fiske (1993) found that people tend to engage in

finer-grained categorization of individuals perceived as having more power than themselves while employing coarser categorization for those considered to be of lower status. This difference in categorization strategies may reflect a motivation to understand and emulate the characteristics of high-power individuals while paying less attention to those perceived as less important.

Furthermore, Fryer and Jackson (2008) suggest that the frequency of exposure to a particular category can influence the level of detail in the categorization process. Specifically, they argue that less frequently encountered objects or groups are more likely to be categorized coarsely, leading to a loss of accuracy in predictions about those categories. This finding has important implications for understanding how minority groups may be more susceptible to stereotyping and discrimination due to their lower frequency of representation in various social contexts.

Categorization is a very natural behavior that can be observed in many situations. Koriat and Sorka (2017) suggest using cues to associate an entity with another entity we encountered. Erickson and Kruschke (1998) explain how agents have descriptional rules about categories, which help them develop rules of thumbs to use in most situations. Chi (2009) claims that people categorize entities or processes to understand them better; it is easier to understand and teach when we divide a topic into subtopics. Fiske (1993)'s experiment shows that people tend to finely categorize individuals with more power than them and coarsely categorize others. This result is because people want to understand how to become like individuals from higher hierarchies, while they do not care as much for the people they consider below them. Fryer and Jackson (2008) suggest that agents categorize less frequent objects more coarsely, which causes them to lose accuracy in their predictions.

This paper follows Gibbons *et al.* (2021) by examining the connection between categorization, performance, and organizational culture. Our definition of categorization comes from Taylor *et al.* (1978), where agents categorize other agents. We try to understand how discrimination can emerge from categorization. Previous studies, such as Martell *et al.* (2012), have shown the emergence of segregation through an agent-based model using a spatial approach inspired by Schelling (1978). In contrast, this paper uses an evolutionary game model and is unique because it does not incorporate any initial bias against any groups, unlike the study mentioned above.

2.2 Agent-based approaches to discrimination

Agent-based modeling (ABM) has emerged as a valuable tool for studying complex social phenomena, including the emergence of discrimination in organizations (see Wall (2016) for an extensive literature review on the use of ABM in organizational sciences). ABM allows researchers to simulate the interactions and behaviors of individual agents within a defined environment, enabling the examination of how micro-level processes can give rise to macro-level patterns (Schelling, 1978).

In the context of discrimination research, ABM has been used to investigate how categorization strategies and power dynamics can contribute to the emergence of discriminatory outcomes. For example, Martell *et al.* (2012) developed an agent-based model to explore the role of gender stereotypes in shaping hiring and promotion decisions in organizations. Their model demonstrated how even small biases at the individual level can accumulate over time to create significant disparities in representation and advancement opportunities for women. Similarly, Amadae and Watts (2022) used an ABM to examine the impact of power imbalances and categorization strategies on the emergence of discrimination in a simulated organizational environment. Their findings highlighted the importance of considering the interplay between individual-level cognitive processes and structural factors in understanding the dynamics of discrimination. O'Connor (2017) state that one of the main advantages of agent-based models is that we do not have to model the interaction between all strategies. Hence, multiple strategies with heterogeneous agents can be conveniently implemented, meaning that interactions can be much more complex in agent-based models (Kallens *et al.*, 2018).

This paper will employ an evolutionary game design to compare different strategies. Evolutionary games are

common in studying organizational culture and its emergence from simple behaviors (Newton *et al.*, 2019). Evolutionary and computational studies have seen a resurgence in the last two decades (Newton, 2018; Carley, 2002). Many issues encountered in organizations, such as trust-building processes, Kantian morality, and assortative mating, are some of the behavioral topics studied with evolutionary games (Fujiwara-Greve *et al.*, 2012; Alger and Weibull, 2016; De Cara *et al.*, 2008). The advantage of evolutionary games is the fact that they allow agents with complex memories while still being loyal to theory (Adami *et al.*, 2016).

3 Asymmetric hawk-dove games

Let us imagine two members of a research team coming up with different ideas for a solution. Both members believe that their ideas are better than their colleagues. In such a situation, members may cooperate on a common solution. However, it is also possible for them to go into a power struggle either to show that they are smarter, to impress their superiors, or simply because they sincerely believe that their ideas are better. These power struggles are common occurrences (Magee and Galinsky, 2008; Greer *et al.*, 2017; Kang, 2022), and they might occur in all kinds of organizations such as governments, research teams, and schools (Caselli, 2006; Kang, 2022; Twemlow *et al.*, 2001). These fights might negatively affect organizational performance, while individuals might gain some benefits (De Dreu and Weingart, 2003; Greer *et al.*, 2017).

We use hawk and dove games to represent intra-organizational conflicts. In these games, if both players play an aggressive strategy (hawk), then a fight occurs where they both share the reward but also share the cost of fighting. If both of them play a peaceful strategy (dove), then they share the reward. Finally, if one of them plays an aggressive strategy while the other one becomes defensive, then the aggressive player gets the reward while the defensive agent gets nothing. A representation of a symmetrical game is illustrated in Table I.

		Player 2	
		Hawk	Dove
Player 1	H	$(V-C)/2, (V-C)/2$	$V, 0$
	D	$0, V$	$V/2, V/2$

Table I: A generic hawk and dove game, where V is the prize, and C is the fighting cost

In a symmetrical hawk-dove game, if the reward is greater than the cost of fighting ($V > C$), then being aggressive, or playing "hawk" (H), is the only Nash Equilibrium (NE) for both players. However, if the reward is less than the cost of fighting ($V < C$), there are two pure strategy NE (H, D) and (D, H) and one mixed strategy NE (MSNE). The MSNE involves playing "hawk" with a probability of V/C and "dove" with a probability of $(1 - \frac{V}{C})$.

The asymmetry in power levels between players in conflicts makes applying the symmetric version of the hawk-dove game difficult. To better represent the dynamics of power struggles, we adopt the asymmetric version of the game presented by Mesterton-Gibbons (1994). In this model, the outcome of a hawk-hawk confrontation is not equal for both players and depends on their power differences. In Table II, $\theta_{i,j}$ refers to the winning probability of an agent i against an agent j, where $0 \leq x_i \leq 1$ denotes the power of agent i.

	H	D
H	$\theta_{i,j}V - (1-\theta_{i,j})C$	V
D	0	$V/2$

Table II: An asymmetric hawk and dove game with row player's payoffs, where $\theta_{i,j}$ indicates the probability of winning for an agent i against an agent j

Following Doi and Nakamaru (2018), we calculate the winning probability of an agent i against an opponent

j using Equation 1:

$$\theta_{i,j} = \frac{1}{1 + e^{-(x_i - x_j) * \zeta}} \tag{1}$$

The asymmetric hawk-dove game we use considers that players in a conflict often have differing levels of strength, information, or prestige. The power reliability parameter, ζ, represents how much a player can rely on their strength, ensuring that a minimal difference in struggling power does not solely determine the outcome of a struggle. As ζ increases, the probability of winning for the stronger player approaches 1 even with a minimal difference in struggling power, while as ζ approaches 0, both players' winning probabilities approach one-half even with a significant difference in struggling power.

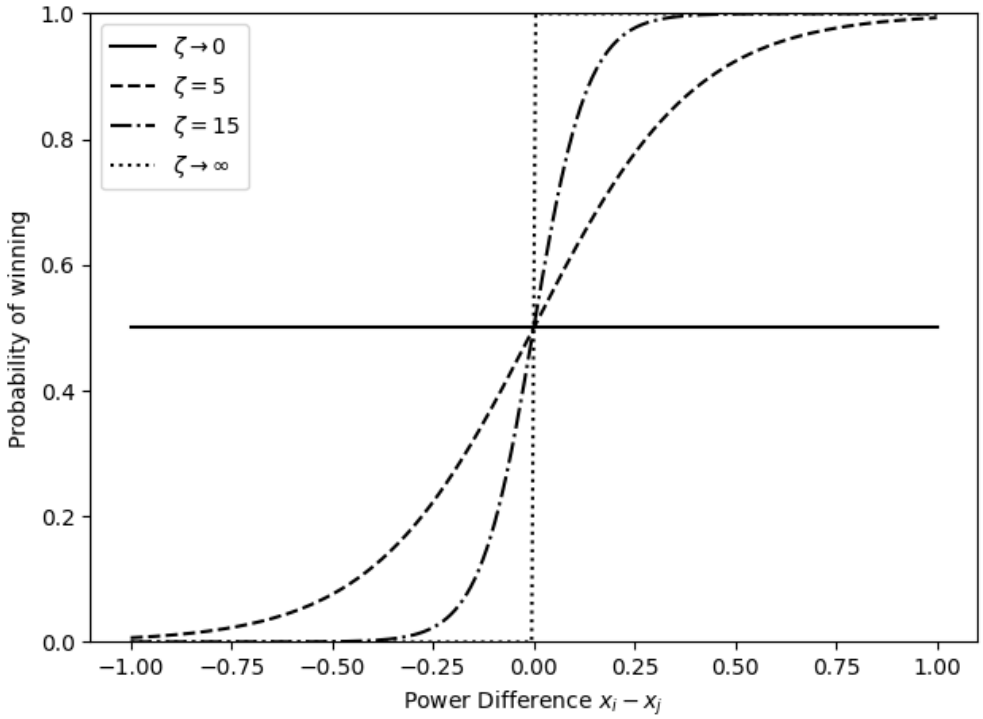


Figure 1: The effect of reliability parameter on the outcome of a hawk-hawk escalation

According to Smith and Parker (1976), if agents possess complete information and the outcome of a power struggle is not solely dependent on the difference in struggling power, mixed strategies may not be necessary. Thus, we denote $s_i = \{s_p, s_w\}$ as the set of best responses for an agent i against an opponent j. s_p is the best response of agent i when she believes that she is more powerful than an agent j (hence, $E_i[x_i] \geq E_i[x_j]$), and s_w is the best response of agent i when she believes that she is weaker than her opponent ($E_i[x_i] < E_i[x_j]$). An example of a best response for an agent i is $s_i = \{H, D\}$. Then, the agent i will choose to play "hawk" if she believes she is more powerful than her opponent and play "dove" otherwise. The possible best responses are the following:

$$s_i = \begin{cases} \{H, H\} & \text{if } \frac{C}{V} < \frac{1-\theta_{i,j}}{\theta_{i,j}} \\ \{H, D\} & \text{if } \frac{1-\theta_{i,j}}{\theta_{i,j}} \leq \frac{C}{V} < \frac{\theta_{i,j}}{1-\theta_{i,j}} \\ \{D, H\} & \text{if } \frac{\theta_{i,j}}{1-\theta_{i,j}} \leq \frac{C}{V} \end{cases} \tag{2}$$

The first condition states that when the cost of fighting is low, an agent opts for a hawk response regardless of her relative strength compared to her opponent. In contrast, the second condition states that when the cost of fighting is high, the agent only chooses to play hawk if she believes that she is stronger than her opponent and chooses the dove strategy otherwise.

The third condition arises when the cost of fighting is very high. In this scenario, there is an *evolutionary stable strategy* (ESS) where the weaker player adopts an aggressive approach, and the stronger player steps back. The stronger player cannot take the chance of adopting a hawk strategy due to the possibility of losing the conflict, given the high cost of fighting. On the other hand, the weaker player has no reason to change her strategy if she is aware that her opponent will retreat. This equilibrium is called paradoxical by Smith and Parker (1976). We do not impose the paradoxical equilibrium; hence, agents will play $\{H, D\}$ if $\frac{1-\theta_{i,j}}{\theta_{i,j}} \leq \frac{C}{V}$; however, we observe how it emerges in simulations due to misbeliefs about categories in Section 4.

The sigmoid functions in Figure 1 align with the findings of Yu *et al.* (2022) and demonstrate how a fight can arise from a conflict when the power difference between players is not too large or too small. When the players' powers are very similar, they should play dove as $V < C$. If one player is clearly stronger than the other, such as a significant power difference, the weaker player will back down, and the struggle will end in a hawk-dove outcome. According to Smith and Parker (1976), any uncertainty about the power difference can increase the chance of a paradoxical equilibrium, which is also the case in our model.

4 Agent-based model

Schelling (1978) showed that to observe complicated macrobehaviors, sophisticated micro-behaviors are not necessary. Instead, modeling heuristic-based simple agents with fundamental motives can be enough to reveal remarkable results. Crowley (2001) shows that the stability of a model can be explained better when agents have memories, which is more challenging to do in equation-based models. We create an agent-based model to analyze the emergence of discrimination using a similar framework to Amadae and Watts (2022).

In this section, we will create an agent-based model to show how discrimination might emerge without any prior difference assumptions. We hypothesize that when people use categorization in interactions, a difference between observable traits will emerge over time.

We assume that all agents have two observable traits: a relevant trait $\mathbf{R}$ and an irrelevant trait $\mathbf{I}$. Each trait has two types, a positive and a negative one. A positive relevant trait $\mathbf{R}^+$ indicates that the agent is actually stronger in conflicts on average against $\mathbf{R}^-$ opponents. One might suggest that the agent has previous education on the topic of discussion, which might help in a conflict. On the other hand, we impose no difference between $\mathbf{I}^+$ and $\mathbf{I}^-$. It is possible to imagine this irrelevant trait as a characteristic that is open to being discriminated against, such as by race or gender.

Then, we assume that agents might have one of the three categorization strategies. Firstly, they can be coarse-grained categorizers, not considering either the relevant trait $\mathbf{R}$ or the irrelevant trait $\mathbf{I}$ as a factor in conflicts. Secondly, they can be regular categorizers, believing that the relevant trait $\mathbf{R}$ plays a role in power struggles, but the irrelevant trait $\mathbf{I}$ does not. Finally, they can opt to be fine-grained categorizers,

considering the relevant trait and the irrelevant trait as important factors. We should also note that we do not impose a prior bias towards $\mathbf{R}^+$ either. Even though a regular or a fine-categorizer agent will expect a difference between R^+ and R^- agents, they will not know who is stronger in the beginning.

Following Azrieli (2009), we also model agents who do not categorize themselves because this might be interpreted as a homophily-driven emergence of discrimination (Bianchi and Squazzoni, 2015). Moreover, the categorization is to estimate the probabilities of winning against an opponent, and an agent does not play against herself. Hence, she does not have any reason to categorize herself.

For example, a regular categorizer who has fought against opponents with $\mathbf{R}^+$ four times and won three encounters will believe that their winning probability against that group to be 75%. They will consider themselves the stronger player if their expected winning probability is higher than 50%. Using Equation 2, this agent will only play hawk if $\frac{0.75}{1-0.75} > \frac{C}{V} \geq \frac{1-0.75}{0.75}$.

On the other hand, a coarse-grained categorizer only considers their overall encounters ignoring specific categories. Agents update their memories after each round, only retaining information on the outcome of hawk-hawk escalations against a particular category. This assumption is because agents try to find their possibility of winning against categories in our model, and they are not interested in what those categories would play on average.

Table III illustrates two agents with differing categorization strategies facing the same opponent and getting the same outcome. Initially, each agent is endowed with self-confidence, which we model as a prior that is biased toward winning. Thus, the first row consists of ones, indicating that the estimated probability of winning against a category is always set at one at the start. As a result, after their first loss, they update their belief to $(1+0)/2 = 0.5$. After their second match, which is a win, they update their belief to $(0.5 * 2 + 1)/3 = 0.66$. If any players play dove, the agent does not update their prior, as no conflict occurs. Notice that the fine-categorizer only updates one column, while the regular categorizer updates two. In the end, despite facing the same opponent and getting the same outcome, both agents have different beliefs against the same traits.

	A regular categorizer				A fine categorizer				Result
Turn	$\mathbf{R}^-\mathbf{I}^-$	$\mathbf{R}^-\mathbf{I}^+$	$\mathbf{R}^+\mathbf{I}^-$	$\mathbf{R}^+\mathbf{I}^+$	$\mathbf{R}^-\mathbf{I}^-$	$\mathbf{R}^-\mathbf{I}^+$	$\mathbf{R}^+\mathbf{I}^-$	$\mathbf{R}^+\mathbf{I}^+$	
0	1	1	1	1	1	1	1	1	Beginning with full confidence
1	0.5	0.5	1	1	1	0.5	1	1	Lost against a $\mathbf{R}^-\mathbf{I}^+$
2	0.66	0.66	1	1	1	0.5	1	1	Won against a $\mathbf{R}^-\mathbf{I}^-$
3	0.66	0.66	1	1	1	0.5	1	1	$\mathbf{R}^+\mathbf{I}^+$ opponent played dove
4	0.66	0.66	0.5	0.5	1	0.5	1	0.5	Lost against a $\mathbf{R}^+\mathbf{I}^+$
5	0.66	0.66	0.33	0.33	1	0.5	0.5	0.5	Lost against a $\mathbf{R}^+\mathbf{I}^-$

Table III: An example update process for a regular categorizer on the left and a fine categorizer on the right playing against the same type of opponents. Values indicate an agent's estimate of winning probability against an opponent in case a hawk-hawk scenario occurs. Notice that despite having the same opponents, they ended up with different beliefs against different categories.

4.1 Evolutionary algorithm

We implement an evolutionary algorithm to determine the performance of different strategies in different circumstances. Our approach involves creating a co-evolutionary model where the relevant trait, the irrelevant trait, and categorization strategies evolve dynamically. We separate the evolution of categorization strategies from the evolution of traits. We consider that categorization is a mental simplification of a complex environment, while traits are assigned to agents. We assume that an agent can observe and imitate other agents' categorization strategies but not their traits. For the categorization strategies, we implement

a Moran birth and death process based on Moran (1958). At each turn, some completely random workers leave the organization due to typical turnover. Then, new workers will be hired based on how well traits perform in power struggles. One can also cause retiring agents to be based on their performance; however, this is equivalent to doubling the turnover rate.

Our co-evolutionary model uses the imitative logit protocol created by Björnerstedt and Weibull (1994) to assess the imitation process. In this protocol, agents can evaluate the categorization strategies of others and decide whether to adopt them based on their performance relative to their own. With this simple approach, we can analyze the advantages of different categorization strategies. We tested our model with the unconditional imitation protocol developed by Roca *et al.* (2009). Our results showed that the imitative logit protocol creates smoother transitions than the unconditional imitation protocol, which causes cascading categorization strategies. Furthermore, both protocols produce similar results regarding categorization ratios and discrimination. Equation 3 expresses the probability of agent i imitating agent j's categorization strategy.

$$P_i(I) = \frac{e^{\pi_i}}{e^{\pi_i} + e^{\pi_j}} \tag{3}$$

For the turnover process, we apply a similar method. We assume a recruiter who studies the average payoffs of each subcategory, i.e., agents with $\mathbf{R}^+\mathbf{I}^-$ traits are one category, with independence from either $\mathbf{R}^-\mathbf{I}^-$ or $\mathbf{R}^+\mathbf{I}^+$. Then, we select a new worker with a subcategory proportional to the performance of that category. For example, the probability of hiring an $\mathbf{R}^-\mathbf{I}^+$ agent ($P_i(\mathbf{R}^-\mathbf{I}^+)$) is expressed in Equation 4, where $\pi_{\mathbf{R}^-\mathbf{I}^+}$ represents the average payoff of $\mathbf{R}^-\mathbf{I}^+$ agents.

$$P_i(\mathbf{R}^-\mathbf{I}^+) = \frac{\exp(\pi_{\mathbf{R}^-\mathbf{I}^+})}{\exp(\pi_{\mathbf{R}^-\mathbf{I}^-}) + \exp(\pi_{\mathbf{R}^-\mathbf{I}^+}) + \exp(\pi_{\mathbf{R}^+\mathbf{I}^-}) + \exp(\pi_{\mathbf{R}^+\mathbf{I}^+})} \tag{4}$$

In summary, the co-evolutionary algorithm has the following key elements:

- Agents with varying traits and categorization strategies.
- An imitation protocol for agents to evaluate their categorization strategies.
- A turnover process to observe the evolution of traits.

Equations 3 and 4 provide well-defined functions for our co-evolutionary algorithm to choose traits, ensuring robust results. We also checked the robustness of our evolutionary algorithm by using another turnover process where the manager evaluated traits separately but found no significant differences.

4.2 Simulation settings

In agent-based modeling, running the simulation multiple times and taking the average to ensure robustness against stochasticity is common practice. Running the simulation 1000 times ensures the results are robust to initial stochasticity. Each run consists of 1000 consecutive turns, where we realize that the results do not change afterward. In Algorithm 1, we present a pseudo-code about how each turn works.

Since our evolutionary model does not permit complete domination (in the meaning of Traulsen and Hauert (2009), where one strategy completely dominates the population), we analyze the average population per run instead of counting the categories that dominated the simulation. Our evolutionary algorithm does not allow complete domination due to the complexity of the environment we define. As Wilensky and Reisman (2006) show, environmental complexity usually provides co-existence and prevents complete domination. We start the simulation with every trait equally distributed in the population. We set the value of the prize to 10,

Algorithm 1 A description of the simulation

```
for Every time step do
    Match agents randomly
    for Every pair of agents do
        Players estimate their probability of winning via their memories
        Players choose their best responses via Eq. 2
        if Both players play hawk then
            Winners calculated via true powers
            Players update their histories
        else
            No update happens
        end if
    end for
    for Every agent do
        Look at a random opponent and imitate via Eq. 3
    end for
    A proportion of agents leave the organization.
    New employees are selected according to how traits perform via Eq. 4
end for
```

and we only change the cost of fighting between runs, following Amadae and Watts (2022). We calibrated reliability parameter ζ to 10 to ensure a degree of stochasticity. In comparison, a smaller value will cause the game to be redundant. Our main result about discrimination is robust to the changes in the reliability parameter. However, we observe that a higher reliability parameter helps fine categorization while a lower reliability parameter helps coarse categorization strategies.

Parameter	Value range	Description
T	5000	Number of turns per run
R	1000	Number of runs per parameter combination
N	240	Number of agents per run
V	10	Value of reward
C	{20, 80}	Cost of losing a fight
ζ	10	Reliability parameter
α, β	(4, 12)	Beta distribution parameters
M	0.01	Turnover rate
ϵ	0.001	Noise

Table IV: Parameters of the simulation.

Another significant factor that we have considered is the rate of personnel turnover. Our findings indicate that a higher turnover rate hinders the formation of organizational culture. In comparison, a lower rate may not allow for sufficient observation of the evolution of trait-based behavioral differences. However, we have evaluated the robustness of our main results under varying turnover rates. While substantial changes in turnover rate disrupt the ability to observe any impact on the model or lead to excessively lengthy simulations, minor modifications do not affect our main conclusions.

Additionally, we have incorporated a minimal probability of noise in communication, where an agent will play hawk irrespective of her struggling power. The inclusion of this noise is based on the premise that human decision-making is often prone to error, as noted by Kahneman *et al.* (2021). Noise introduces a level of randomness that can trigger conflicts even if not initially desired by any of the agents. Furthermore, a mutation rate makes the model more resilient to homogeneous states (Fudenberg and Imhof, 2006).

5 Results

We explore two research questions. Firstly, we assess the efficacy of various categorization strategies. Secondly, we examine the emergence of any differences based on the irrelevant trait. Despite giving agents with the relevant trait a head start, we do not anticipate a gap as both types of the irrelevant trait start on equal footing. Any observed disparity between subgroups would therefore be considered an emergent behavior.

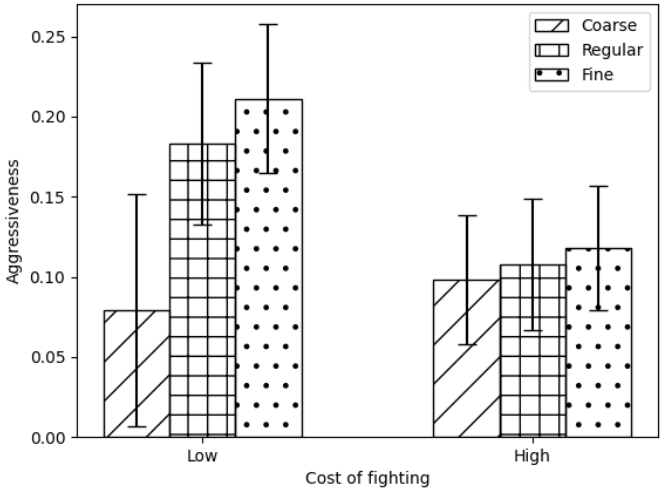


Figure 2: The difference in aggressiveness between categorization strategies. A low cost of fighting indicates that $\frac{C}{V} = 2$, while a high cost of fighting indicates $\frac{C}{V} = 8$. Error bars indicate standard deviations.

In Figure 2, each bar indicates the level of aggressiveness reached by the average of a categorization strategy. For example, the leftmost bar shows that coarse categorizers played hawk 20% of the time on average when the cost of fighting is 20. Our results reveal that finer-grained categorization strategies lead to increased aggressiveness among agents when the cost of fighting is low. This result is statistically significant, with a $p < 0.01$ value for consecutive bars. It is important to keep in mind that the error bars represent standard deviations. The reason for this trend can be easily explained using a simple example. Given a population of agents with varying levels of the $\mathbf{R}$, the overall power of the population will be in between the two groups, as shown in Figure 3. Table III highlights that fine categorization leads to exploring against a greater number of opponents, which becomes advantageous when the cost of conflict is low. This finding aligns with the study by Martignoni *et al.* (2016) that over-specification can lead to exploratory effects.

The aggressiveness difference between regular and coarse categorization is tricky. A coarse categorizer eventually evaluates herself against the average of the whole population. In contrast, a regular categorizer evaluates her power against two averages: $\mathbf{R}^-$ opponents or $\mathbf{R}^+$. Since the prior is usually lower, a regular categorizer might use this fact to her advantage. Figure 3 visualizes this example.

5.1 Cost of fighting

In our simulations, we observed a correlation between the cost of fighting and the prevalence of different categorization strategies. When the cost of fighting was low, we saw a rise in the use of fine categorization, while when the cost of fighting was high, we observed an increase in coarse categorization. Since the game is an asymmetric hawk and dove game, the cost of fighting must always be greater than the prize. To explore this relationship, we conducted simulations with the cost of fighting starting from 11 and found that values above 100 produced similar results. For our analysis, we defined low cost as 20 and high cost as 80.

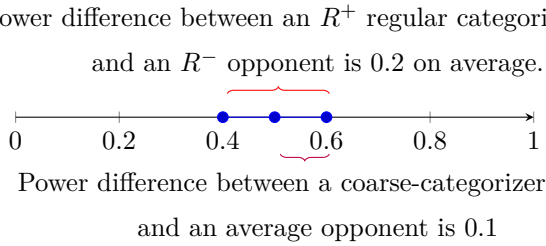
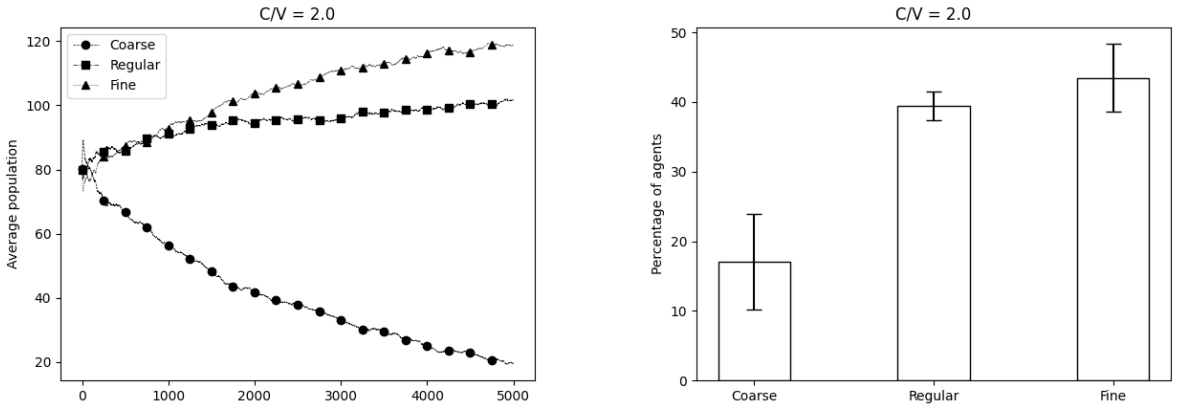


Figure 3: Coarse-categorizer agents tend to weigh their struggling power against the average population, as they do not distinguish between categories.



(a) Average time plot.

(b) Error bars indicate standard deviations.

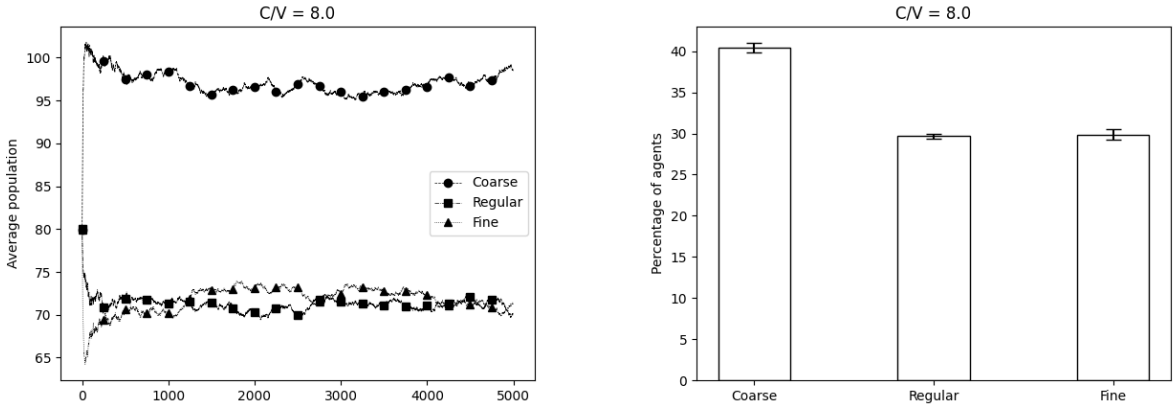
Figure 4: $C/V = 2$. Fine categorization provides a small advantage.

Figure 4 depicts the average of a thousand runs where the cost of fighting is relatively low ($C/V = 2$). We observe that fine categorization enjoys a very sharp advantage against other categorization strategies. The fact that fine-categorizer agents prevail against regular agents when the cost of fighting is low is intriguing. Even though we know that different types of **I** have the same struggling power, they may not have the same aggressiveness level throughout the simulation. The reasoning is that, initially, fine-categorizer agents may increase in one type while they decrease in the other because of stochasticity. Since coarse-categorizer agents tend to see the average population, they estimate that the struggling power difference between their opponent and them is relatively small, as explained in Figure 3. Therefore, they tend to employ a more pacifist strategy. Fine-categorizer agents can become more aggressive even against stronger opponents due to not being punished. This behavior is the emergence of paradoxical equilibrium explained in Section 3. Suppose fine-categorizer agents realize that agents with negative irrelevant positive relevant traits are coarse-categorizers on average and play dove as coarse-categorizer agents more frequently do. In that case, fine-categorizer negative relevant traited agents may exploit this fact via paradoxical equilibrium. This behavior causes the negative irrelevant positive relevant agents to leave the organization over time, causing only positive relevant positive irrelevant traits to stay in the workplace. Even though we ensure that there are no differences between subgroups initially, cultural differences emerge.

Furthermore, we observe that both relevant trait agents follow a similar categorization strategy. However, agents with positive relevant agents perform better if they choose to be coarse-categorizers than agents with negative relevant traits because they have more struggling power on average. Also, this is why we did not focus on the population difference between relevant trait types, as any difference in this regard is solely

a result of our assumptions and not an emergent behavior from the model. Nevertheless, categorization choices by agents of different relevant trait types remain intriguing. Most importantly, these results are the average of a thousand runs. Hence, this pattern is consistent through different parameter settings or initial conditions.

When we increase the cost of fighting in Figure 5, we observe that coarse-categorizing agents obtain an evolutionary advantage in the simulation more than regular or fine-categorizer agents when the cost of fighting is high because they calculate the average of the population; therefore, coarse-categorizer agents tend to reach a peaceful conclusion much faster, providing an evolutionary advantage. On the other hand, regular and fine-categorizer agents may try being aggressive for longer since they will not be punished against peaceful opponents, as explained in Table III. However, when they are punished, the cost of being punished can be too high to slow them down in the evolutionary race. More interestingly, agents of a positive relevant trait now mainly select to be coarse-categorizers.



(a) Average time plot.

(b) Error bars indicate standard errors

Figure 5: Coarse categorization proves useful when $C/V = 8$.

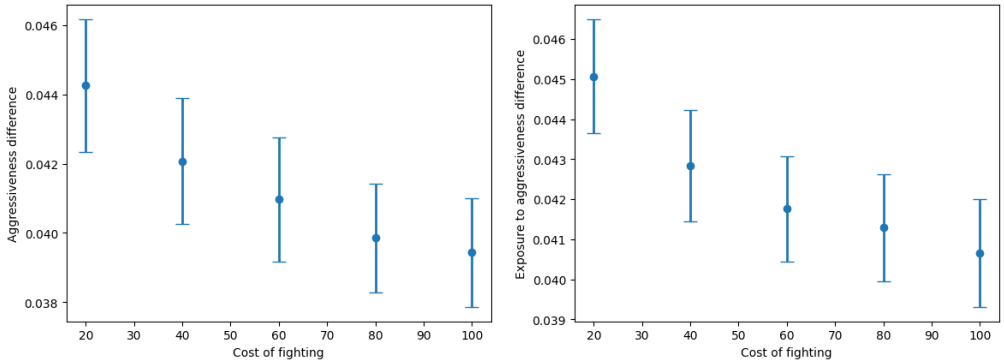
5.2 Discrimination

In our second analysis, we examine a potential discrimination based on an irrelevant trait. We model both types of the trait similarly with identical initial conditions and conflict strength; therefore, observing any emergence of discrimination is not a trivial task. Any variation we observe is solely the result of stochasticity. Hence, if a single simulation favored one of the subgroups, it would likely favor the other subgroup in the next run. Therefore, averaging multiple simulations is not an appropriate method; both subgroups are likely to win 50% of the time. To address this issue, we adopt an identification strategy where we study the absolute differences between subgroups as a meaningful metric. For example, in half of the simulations we expect $\mathbf{I}^+$ to win over $\mathbf{I}^-$, while in the other half we expect $\mathbf{I}^-$ to win. However, we show that the difference between a winning subgroup and a losing subgroup is consistent throughout different runs.

In Figure 6a, we observe the emergence of discrimination for different costs of fighting. Aggressiveness here represents the percentage of the "hawk" strategy being played by the subgroup. Furthermore, the direction of the relationship stays the same in various parameter combinations. This result is especially striking. No difference between the subgroups of the irrelevant trait is implemented; however, a cultural difference emerges when fighting costs are low. This difference eventually might lead to stereotypes, as explained by Taylor *et al.* (1978). Furthermore, our results indicate that a newcomer positive irrelevant trait agent eventually adopts a more aggressive strategy due to the turnover process.

We follow the framework of Borgonovo *et al.* (2022) for robustness analysis. Following this framework, we change imitation strategies, parameters, and an update mechanism one by one and observe the correlation between the cost of fighting and discrimination from Figure 6a. Different imitation protocols change the parameters where categorization strategies prevail; however, the correlation between the difference between subgroups in aggressiveness and the cost of fighting remains in the same direction. Furthermore, different parameters might change the cooperation rate or the winning categorization strategy, yet discrimination remains negatively related to the cost of fighting.

Figure 6b depicts the difference in being subject to aggressiveness between subgroups, revealing a clear case of discrimination. This result goes beyond what was indicated in Figure 6a, which merely pointed to cultural differences and stereotypes. Figure 6b demonstrates that one subgroup is more subject to aggressiveness compared to the other. The difference in aggressiveness and being subject to aggressiveness represent two distinct issues. The former suggests that agents' behavior will change based on their label, while the latter highlights that an agent will be treated differently simply due to their label. The reason for this discriminatory effect lies in the nature of the hawk and dove game, where the cost of fighting is always higher than the reward. As a result, a constantly fighting agent will eventually fail unless they battle opponents who always play dove. Thus, a frequently targeted category must play dove; otherwise, they will lose more than they gain overall. This result aligns with the concept of paradoxical equilibrium.



(a) Aggressiveness **(b)** Being subject to aggressiveness

Figure 6: These figures show the differences between irrelevant categories. Error bars indicate standard deviations. Each consecutive $p < 0.01$, except from 60 to 80.

6 Discussion

The findings of this study provide valuable insights into the emergence of discrimination in organizations through the interplay of categorization strategies and power struggles. Our agent-based model demonstrates that even in the absence of prior differences between groups, discriminatory behaviors can arise solely due to the presence of agents who label categories based on observable traits.

The simulation results reveal that the cost of fighting plays a crucial role in determining the prevalence of different categorization strategies. When the cost of fighting is low, fine-grained categorization provides an advantage, as it allows agents to explore a greater number of opponents and exploit the paradoxical equilibrium. Conversely, when the cost of fighting is high, coarse-grained categorization becomes more advantageous, as it promotes a more peaceful approach and faster convergence to a stable state.

Notably, our model sheds light on the emergence of discrimination without any initial assumptions of differences between groups. The results indicate that when the cost of fighting is low, a cultural difference in aggressiveness emerges, which can eventually lead to the formation of stereotypes. Furthermore, the model

reveals a clear case of discrimination, as one subgroup is exposed to higher levels of aggressiveness simply due to their label. Our findings demonstrate that any bias towards a particular category might be solely a result of categorization. These findings correlate with empirical evidence from Sarsons (2017), where a self-fulfilling prophecy occurs.

These findings have significant implications for understanding the complex dynamics of discrimination in organizations. The model suggests that categorization strategies employed by individuals can contribute to the perpetuation of discriminatory behaviors, even in the absence of pre-existing biases or differences between groups. This highlights the importance of addressing not only overt forms of discrimination but also the subtle cognitive processes that can lead to the emergence and reinforcement of discriminatory outcomes.

The study also underscores the role of organizational factors, such as the cost of conflicts and turnover rates, in shaping the dynamics of discrimination. Higher costs of conflict may discourage discriminatory behaviors by promoting more peaceful interactions, while lower costs can foster the emergence of cultural differences and stereotypes. Additionally, the turnover process can perpetuate discriminatory patterns as newcomers adopt the prevailing strategies and behaviors within the organization. Moreover, considering how intra-organizational conflicts might cause a novel discriminatory behavior, a manager might want to reduce the probability of power struggles. For example, a performance-based employee ranking system can cause employees to struggle for higher rankings, eventually decreasing the overall performance of teams (Ewenstein *et al.*, 2016). Therefore, it is possible for a manager to cause harm while trying to increase performance through competition.

While providing valuable insights, the agent-based model is a simplified representation of complex social interactions. Future research could extend the model by allowing for many traits, with possible interdependencies amongst them. It is essential to understand how agents would react to complex environments when more parameters are involved. Furthermore, one could model misspecified agents where the miscategorization does not lie in how many traits are significant but which ones are important.

Despite these limitations, our study makes a significant contribution to the literature on discrimination and categorization in organizations. By demonstrating the emergence of discrimination through the interplay of categorization strategies and intra-organizational conflicts, we offer a novel perspective on the formation and persistence of discriminatory behaviors.

7 Conclusion

This paper presents an agent-based model that explores the emergence of discrimination in organizations through the interplay of categorization strategies and power struggles. The model demonstrates that discriminatory behaviors can arise even in the absence of prior differences between groups, solely due to the presence of agents who label categories based on observable traits.

The simulation results reveal the impact of the cost of fighting on the prevalence of different categorization strategies, with fine-grained categorization being advantageous when the cost is low and coarse-grained categorization being favored when the cost is high. Notably, the model shows the emergence of discrimination without any initial assumptions of differences, with cultural differences in aggressiveness and exposure to aggressiveness arising when the cost of fighting is low.

The findings of this study have important implications for understanding and addressing discrimination in organizations. They highlight the role of categorization strategies and organizational factors in shaping the dynamics of discrimination and underscore the need for proactive measures to promote inclusivity and mitigate the emergence of discriminatory behaviors.

Future research could extend the model by incorporating additional factors and conducting empirical studies

to validate the findings. Nevertheless, this study makes a contribution to the literature on discrimination and categorization in organizations, offering novel insights into the complex dynamics that can lead to the emergence and persistence of discriminatory outcomes.

In conclusion, this paper emphasizes the importance of recognizing and addressing the subtle cognitive processes and organizational factors that can contribute to the emergence of discrimination. By fostering inclusive practices and promoting awareness of the potential pitfalls of categorization strategies, organizations can work towards creating more equitable and diverse work environments.

References

Adami, C. *et al.* (2016), "Evolutionary game theory using agent-based methods", *Physics of life reviews*, Vol. 19, pp. 1–26.

Alger, I. and Weibull, J. W. (2016), "Evolution and Kantian morality", *Games and Economic Behavior*, Vol. 98, pp. 56–67.

Amadae, S. and Watts, C. J. (2022), "Red Queen and Red King Effects in cultural agent-based modeling: Hawk Dove Binary and Systemic Discrimination", *The Journal of Mathematical Sociology*, pp. 1–28.

Arrow, K. J. (1998), "What has economics to say about racial discrimination?", *Journal of economic perspectives*, Vol. 12 No. 2, pp. 91–100.

Azrieli, Y. (2009), "Categorizing others in a large game", *Games and Economic Behavior*, Vol. 67 No. 2, pp. 351–362.

Becker, G. S. (2010), *The economics of discrimination*, University of Chicago press.

Bertrand, M. and Duflo, E. (2017), "Field experiments on discrimination", Banerjee, A. V. and Duflo, E., (Ed.s), *Handbook of economic field experiments*, Vol. 1, Elsevier, pp. 309–393.

Bianchi, F. and Squazzoni, F. (2015), "Agent-based models in sociology", *Wiley Interdisciplinary Reviews: Computational Statistics*, Vol. 7 No. 4, pp. 284–306.

Björnerstedt, J. and Weibull, J. W. (1994), "Nash equilibrium and evolution by imitation", working paper, IUI Working Paper No: 407, pp. 52–94.

Borgonovo, E. *et al.* (2022), "Sensitivity analysis of agent-based models: a new protocol", *Computational and Mathematical Organization Theory*, Vol. 28 No. 1, pp. 52–94.

Brodmann, J. *et al.* (2022), "Chief executive officer power and board gender diversity", *Finance Research Letters*, Vol. 44.

Bruner, J. P. (2019), "Minority (dis) advantage in population games", *Synthese*, Vol. 196, pp. 413–427.

Carley, K. M. (2002), "Computational organization science: A new frontier", *Proceedings of the National Academy of Sciences*, Vol. 99 No. 3, pp. 7257–7262.

Caselli, F. (2006), "Power struggles and the natural resource curse", working paper, Centre for Economic Performance Economics, London School of Economics.

Chi, M. T. (2009), "Three types of conceptual change: Belief revision, mental model transformation, and categorical shift", Vosniadou, S., (Ed.), *International handbook of research on conceptual change*, Routledge, pp. 89–110.

Crowley, P. H. (2001), "Dangerous games and the emergence of social structure: evolving memory-based strategies for the generalized hawk-dove game", *Behavioral Ecology*, Vol. 12 No. 6, pp. 753–760.

De Cara, M. *et al.* (2008), "A model for the evolution of assortative mating", *The American Naturalist*, Vol. 171 No. 5, pp. 580–596.

De Dreu, C. K. and Weingart, L. R. (2003), "Task versus relationship conflict, team performance, and team member satisfaction: a meta-analysis.", *Journal of applied Psychology*, Vol. 88 No. 4, pp. 741–749.

Doi, K. and Nakamaru, M. (2018), "The coevolution of transitive inference and memory capacity in the hawk–dove game", *Journal of Theoretical Biology*, Vol. 456, pp. 91–107.

Erickson, M. A. and Kruschke, J. K. (1998), "Rules and exemplars in category learning.", *Journal of Experimental Psychology: General*, Vol. 127 No. 2, pp. 107–140.

Ewenstein, B. *et al.* (2016), "Ahead of the curve: The future of performance management", *The McKinsey Quarterly*, pp. 1–10.
Fiske, S. T. (1993), "Controlling other people: The impact of power on stereotyping.", *American psychologist*, Vol. 48 No. 6, pp. 621–628.
Flache, A. and Mäs, M. (2008), "How to get the timing right. A computational model of the effects of the timing of contacts on team cohesion in demographically diverse teams", *Computational and Mathematical Organization Theory*, Vol. 14 No. 1, pp. 23–51.
Fryer, R. and Jackson, M. O. (2008), "A categorical model of cognition and biased decision making", *The BE Journal of Theoretical Economics*, Vol. 8 No. 1.
Fudenberg, D. and Imhof, L. A. (2006), "Imitation processes with small mutations", *Journal of Economic Theory*, Vol. 131 No. 1, pp. 251–262.
Fujiwara-Greve, T. *et al.* (2012), "Voluntarily separable repeated Prisoner's Dilemma with reference letters", *Games and Economic Behavior*, Vol. 74 No. 2, pp. 504–516.
Gibbons, R. *et al.* (2021), "What situation is this? Shared frames and collective performance", *Strategy Science*, Vol. 6 No. 2, pp. 124–140.
Greenwald, A. G. and Banaji, M. R. (1995), "Implicit social cognition: attitudes, self-esteem, and stereotypes.", *Psychological review*, Vol. 102 No. 1, p. 4.
Greer, L. L. *et al.* (2017), "The dysfunctions of power in teams: A review and emergent conflict perspective", *Research in Organizational Behavior*, Vol. 37, pp. 103–124.
Kahneman, D. *et al.* (2021), *Noise: a flaw in human judgment*, Hachette UK.
Kallens, P. A. C. *et al.* (2018), "Cultural evolution of categorization", *Cognitive systems research*, Vol. 52, pp. 765–774.
Kang, S. M. (2022), "Internal fights over resources: The effect of power struggles on team innovation", *Frontiers in Psychology*, Vol. 13.
Koriat, A. and Sorka, H. (2017), "The construction of category membership judgments: Towards a distributed model", Cohen, H. and Lefebvre, C., (Ed.s), *Handbook of categorization in cognitive science*, Elsevier.
Magee, J. C. and Galinsky, A. D. (2008), "8 social hierarchy: The self-reinforcing nature of power and status", *The academy of management annals*, Vol. 2 No. 1, pp. 351–398.
Martell, R. F. *et al.* (2012), "From bias to exclusion: A multilevel emergent theory of gender segregation in organizations", *Research in Organizational Behavior*, Vol. 32, pp. 137–162.
Martignoni, D. *et al.* (2016), "Consequences of misspecified mental models: Contrasting effects and the role of cognitive fit", *Strategic Management Journal*, Vol. 37 No. 13, pp. 2545–2568.
Mesterton-Gibbons, M. (1994), "The Hawk—Dove game revisited: Effects of continuous variation in resource-holding potential on the frequency of escalation", *Evolutionary Ecology*, Vol. 8, pp. 230–247.
Moran, P. A. P. (1958), "Random processes in genetics", *Mathematical proceedings of the cambridge philosophical society*, Vol. 54, No. 1, Cambridge University Press, pp. 60–71.
Newton, J. (2018), "Evolutionary game theory: A renaissance", *Games*, Vol. 9 No. 2.
Newton, J. *et al.* (2019), "Watercooler chat, organizational structure and corporate culture", *Games and Economic Behavior*, Vol. 118, pp. 354–365.
O'Connor, C. (2017), "The cultural red king effect", *The Journal of Mathematical Sociology*, Vol. 41 No. 3, pp. 155–171.
Pager, D. and Shepherd, H. (2008), "The sociology of discrimination: Racial discrimination in employment, housing, credit, and consumer markets", *Annu. Rev. Sociol*, Vol. 34, pp. 181–209.
Phelps, E. S. (1972), "The statistical theory of racism and sexism", *The american economic review*, Vol. 62 No. 4, pp. 659–661.
Roca, C. P. *et al.* (2009), "Imperfect imitation can enhance cooperation", *Europhysics Letters*, Vol. 87 No. 4.
Sarsons, H. (2017), "Interpreting signals in the labor market: evidence from medical referrals", Job Market Paper, Harvard University, pp. 141–145.
Schelling, T. C. (1978), *Micromotives and macrobehavior*, WW Norton & Company.
Smith, J. M. and Parker, G. A. (1976), "The logic of asymmetric contests", *Animal behaviour*, Vol. 24 No. 1, pp. 159–175.

Stewart, A. J. and Raihani, N. (2023), "Group reciprocity and the evolution of stereotyping", *Proceedings of the Royal Society B*, Vol. 290 No. 1991.

Taylor, S. E. *et al.* (1978), "Categorical and contextual bases of person memory and stereotyping.", *Journal of personality and social psychology*, Vol. 36 No. 7.

Traulsen, A. and Hauert, C. (2009), "Stochastic evolutionary game dynamics", *Reviews of nonlinear dynamics and complexity*, Vol. 2, pp. 25–61.

Twemlow, S. W. *et al.* (2001), "Improving the social and intellectual climate in elementary schools by addressing bully-victim-bystander power struggles", Cohen, J., (Ed.), *Caring classrooms/intelligent schools: The social emotional education of young children*, Citeseer, pp. 162–181.

Wall, F. (2016), "Agent-based modeling in managerial science: an illustrative survey and study", *Review of Managerial Science*, Vol. 10 No. 1, pp. 135–193.

Wilensky, U. and Reisman, K. (2006), "Thinking like a wolf, a sheep, or a firefly: Learning biology through constructing and testing computational theories—an embodied modeling approach", *Cognition and instruction*, Vol. 24 No. 2, pp. 171–209.

Yu, N. Y. *et al.* (2022), "Does Power Difference Always Escalate to Power Struggle? Two Studies with Performance Implications", *Academy of Management Proceedings*, Academy of Management Briarcliff Manor.