USING VIRTUAL CONFEDERATES TO RESEARCH INTERGROUP BIAS AND CONFLICT

CELSO M. DE MELO
USC Marshall School of Business,
Los Angeles, CA 90089-0808

PETER J. CARNEVALE
USC Marshall School of Business

JONATHAN GRATCH
USC Institute for Creative Technologies

INTRODUCTION

Existing research in intergroup bias and conflict has focused on three kinds of experimental designs: face-to-face, paper-and-pencil and computer-mediated interaction. In the first case participants engage in open-ended face-to-face interaction with other participants. This naturalistic technique benefits from ecological validity but, lacks in experimental control. To address this concern, experimenters may introduce confederates, i.e., humans that pose as participants but, unbeknownst to the real participants, are actually part of the experimental manipulation. Human confederates, nonetheless, can introduce inadvertent noise across participants due to subtle changes in their nonverbal or verbal behavior. Paper-and-pencil designs address these limitations by having participants supposedly interact with other participants through exchange of written offers but, in reality, they always receive a scripted pattern of counteroffers. Computer-mediated interaction is the modern version of paper-and-pencil designs where participants interact with each other via a computer. These techniques, despite having increased experimental control, remove much of the richness that exists in face-to-face settings from the interaction. Virtual confederates are a promising middle-ground that captures the advantages of all these techniques.

Virtual Confederates as a Research Tool

Virtual confederates are digital representations of humans (Figure 1). They have three-dimensional bodies and can communicate, like humans, using the face, voice or gesture. Virtual confederates have recently been gaining attention for their potential as a research tool (Blascovich et al., 2002). First, they allow precise definition of the manipulation (e.g., physical appearance and nonverbal behavior) while maintaining other factors constant (e.g., subtle, random or systematic, biases introduced by human confederates). Second, since virtual confederates can be made to look and act like real humans, this added experimental control can be achieved without compromising mundane realism (and, thus, the generalizability of the results). Third, they facilitate replication since everything is recorded in the program that defines how confederates look and act, which can then be shared with other researchers. Fourth, because they can run in online environments, it becomes easier to recruit a broader sample than what is available through local student pools. Fifth, they are low cost, since they work for free and don't require sleep. Finally, they allow for easy manipulation of physical attributes including age,

gender or race. For these and other reasons, we believe virtual confederates can be invaluable for conducting research in decision making and intergroup dynamics.

Figure 1 about here

Virtual confederates can be distinguished according to whether they are controlled by humans, in which case we refer to them as *avatars*, or by computer algorithms, in which case we refer to them as *agents*. This distinction is important because research shows that the mere belief about whether virtual confederates are agents or avatars can influence people's behavior. Blascovich and colleagues (2002) argue this occurs because social influence is greater the higher the perceived "agency" of the confederate. Agency refers to people's theories of mind regarding these virtual entities, i.e., the perceived sentience (e.g., attributions of consciousness, free will). This view is also in line with general findings that people attribute more mind to humans than to computers (Waytz, Gray, Epley, & Wegner, 2010). Whereas these findings might suggest it is always better to have participants believe they are engaging with avatars (independently of whether this is true), we believe there are several reasons to use and study agents. First, there is great value in having standardized negotiation counterparts. Because agents are computers, the algorithm that describes their behavior can be precisely described to participants. This kind of experimental control is harder to achieve when participants believe they are engaging with humans. Second, sometimes researchers may wish to avoid deception; in this case, unless other participants are actually controlling avatars, it is preferable to just use agents. Third, agents can be used as training tools for negotiators or conflict mediators. Finally, there has been growing interest in the development of artificial negotiators that make decisions on behalf of people and, therefore, it is important we understand how people interact and decide with such agents.

Social Categorization and Intergroup Behavior

Underlying all intergroup relations is a basic cognitive process of social categorization (Crisp & Hewstone, 2007): people categorize others into groups while associating, or self-identifying, more with some (the in-groups) than others (the out-groups). Because of this categorization, people will conform more to the values and norms of the group, and tend to favor the in-group to the out-group–a phenomenon referred to as in-group bias. One consequence of this bias is that people trust and cooperate more with in-group than out-group members. This in-group favoritism can, subsequently, escalate into out-group aggression, especially when the in-group's standing is defined in relation to the out-group and this relationship is perceived to be a zero-sum game. Social identities, however, are complex and multifaceted. In many situations, more than one social category (e.g., gender, age, ethnicity) may be relevant. On the one hand, context can prime one category to become more dominant (or salient) and effectively exclude the influence of the others. On the other hand, social categories can be simultaneously salient and have an additive effect on people's behavior (Crisp & Hewstone, 2007). These mechanisms based on multiple categories have been proposed as the basis for reducing intergroup bias. In the common in-group identity model, a common superordinate category is emphasized, thus facilitating a more inclusive definition of "us". In the crossed categorization model, a second category that is shared among two groups is made simultaneously salient, thus effectively reducing the difference between the groups.

Research Questions

To demonstrate the plausibility of virtual confederates for the study of intergroup behavior, we posit that it is, first and foremost, necessary to replicate with virtual confederates key findings from the human-human interaction literature. Secondly, it is important to understand whether people behave differently with computers that are perceived to be controlled by humans (avatars), when compared to computers that are perceived to be controlled by algorithms (agents). Specifically, we focused on the following issues: (1) Do people apply social categories to virtual confederates? (2) Can multiple social categories be applied to virtual confederates? (3) If so, can multiple categories be used to reduce bias? (4) Do people behave differently with (perceived) agents when compared to (perceived) avatars? To answer these questions we ran three experiments where participants engaged in decision making tasks with agents and avatars.

EXPERIMENT 1: ETHNICITY

To understand whether people apply social categories to virtual confederates, in Experiment 1, we had participants engage in a simple decision making task–the dictator game–with virtual confederates that were either of the same or different ethnicity. Although racial discrimination is on the decline, people still tend to make automatic distinctions based on race, which can produce subtle forms of racial discrimination (Gaertner & Dovidio, 2005). Therefore, when engaging with virtual confederates that were proxies for other participants (i.e., avatars), we expected participants to show a bias, in terms of money offered, in favor of same-ethnicity confederates. Regarding the case where virtual confederates are controlled by algorithms (i.e., agents), studies in human-computer interaction have shown that people can behave, in social contexts, in a fundamentally social manner with computers (Reeves & Nass, 1996). Moreover, this research demonstrates that people apply human stereotypes to computers: In one experiment, people perceived computers with a virtual face of the same ethnicity as being more trustworthy and giving better advice than a computer with a face of a different ethnicity. Thus, our expectation was that people would be biased in favor of same-ethnicity agents, when compared to different-ethnicity agents.

The experiment followed a 2 × 2 between-participants factorial design: Other (Agent vs. Avatar) × Ethnicity (Same vs. Different). We took special care to ensure participants believed avatars were being controlled by other participants (e.g., we ran several participants at the same time in each experimental session). Regarding ethnicity, in the ‘different ethnicity’ condition, participants were randomly matched with one of the other ethnicities. We recruited 184 participants at the USC Marshall School of Business for this experiment.

The results showed two main effects: (1) people offered more money to counterparts of the same ethnicity than counterparts of a different ethnicity; (2) people offered more money to avatars than to agents. Moreover, there was no main statistical interaction, which suggests that the effects of Ethnicity and Other were independent and additive. The results, thus, confirmed that people apply social categories to virtual confederates, agents or avatars, and accordingly showed an in-group bias towards confederates that shared the same ethnicity. However, the results also demonstrate an important distinction between agents and avatars, with offers tending

to favor avatars. This basic distinction seems to reflect participants' in-group favoritism towards confederates that belong to the human social category.

EXPERIMENT 2: NESTED SOCIAL DILEMMA

In contrast to Experiment 1, which created group membership by manipulating a (visual) characteristic of the confederates, Experiment 2 manipulated group membership by creating payoff interdependence among players. To achieve this we used the nested social dilemma, which splits players into groups and bids group interests against collective interests. This is a 6-player task where the participant is randomly allocated to position A, B, C, D, E or F and accordingly assigned to group ABC or DEF. The participant is given 30 tickets (for a lottery of $50) that can be invested in three accounts: the *private*, *in-group* and *all* accounts. Tickets invested to the private account are multiplied by 1.0 and returned to the participant; tickets invested to the in-group account are multiplied by 2.5 and split equally among all group members; tickets invested to the all account are multiplied by 4.0 and split equally by all six players. These payoff characteristics create interdependence among group members and preserve the defining properties of a social dilemma: irrespective of others' allocations, shifting points from a higher to a lower level account always increased one's individual final payoff; however, if everyone is selfish and invests in a lower account, then everyone is worse off than if they had invested in a higher account.

Participants engaged, in a between-participants factorial design, with in-group members that were (perceived to be) either agents or avatars, crossed with out-group members that were either agents or avatars. In line with earlier work on the in-group bias, we expected people to favor in-group avatars to out-group avatars. Since a previous study had already shown that people can favor a computer that belongs to the team when compared to a non-team computer (Reeves & Nass, 1996), we also expected people to favor in-group agents to out-group agents. When engaging with in-group avatars and out-group agents, we expected people to strongly favor the in-group not only because they belonged to the interdependent group but also to the human social category. The last case is more interesting: when engaging with in-group agents and out-group avatars, interdependence favors the agents but people also identify with the human social category of the out-group. Following the results in the previous experiment we expected these two influences to cancel each other out, which would result in no preference between the in- and out-groups. We recruited 116 participants at the USC Marshall School of Business.

As expected people invested more in the private than the other accounts; however, to test our hypotheses we focused on a measure for the in-group bias, which we operationalize as the difference between allocations to the in-group and the all accounts. For each condition, we tested whether the in-group bias was statistically significant from zero. The results revealed that people were indeed showing an in-group bias, except when the in-group was composed of agents and the out-group of avatars. These results, thus, confirm that it is possible to create group membership–and corresponding in-group bias–with virtual confederates by manipulating the payoff structure. Moreover, the results showed that the human social category had, once again, an additive effect with other social categories, so that their combined effects cancelled when the in-group was composed of agents and the out-group of avatars.

EXPERIMENT 3: NESTED SOCIAL DILEMMA & ETHNICITY

In the last experiment, we wanted to understand whether it was possible to create a context in which people would favor agents to avatars. Following evidence in the previous experiments that characteristic-based and structure-based social categories can combine in additive fashion, in Experiment 3 we had participants engage in the nested social dilemma with an in-group that was always composed of agents of the same ethnicity as the participant but, with an out-group that was composed of avatars of either the same or a different ethnicity. For the case where both the in-group agents and out-group avatars had the same ethnicity, we expected to replicate the result in Experiment 2, i.e., no preference between the in- and out-groups. For the case where the out-group was composed of avatars of a different ethnicity, we expected people to favor the in-group agents. The rationale is that in this case two categories (ethnicity and payoff-defined group membership) favored the agents and only one favored the avatars (human category). We recruited 47 participants on Amazon Mechanical Turk for this experiment. The results confirmed our expectations, thus showing that by associating more positive social categories with agents than avatars, it is possible to overcome people's bias in favor of avatars.

GENERAL DISCUSSION

Virtual confederates are a useful research tool to study intergroup bias and conflict. Similarly to face-to-face interaction (Crisp & Hewstone, 2007), people applied stereotypes (Experiments 1 and 3) to confederates that were (perceived to be) controlled by other humans (i.e., avatars) and, accordingly, showed a bias in favor of the in-group in terms of money offers. Even when the confederates were controlled by computers (i.e., agents), people could not help themselves to categorize the confederates and show an in-group bias (cf. Reeves & Nass, 1996). Our results from Experiment 2 also demonstrate it is possible to create artificial social categories based on interdependence through shared payoffs. Finally, Experiment 3 demonstrated that people can combine, in additive fashion, the effects of multiple social categories with computers in the same manner as with humans (Crisp & Hewstone, 2007).

Aside from being able to replicate (and eventually extend) findings from the intergroup behavior literature, virtual confederates bring with them several other advantages: experimental control, mundane realism (since confederates look and act like humans), ease of replication, facilitated access to broader samples, low cost, and easy manipulation of physical properties. However, it is also important to point out some of the challenges with using this technology. Unlike fully immersive virtual reality (Blascovich et al., 2002), virtual confederate technology is not expensive; nevertheless, considerable programming effort is still required. In this sense, researchers could benefit by having someone with appropriate computer science expertise on their teams. These issues, however, are likely to become less relevant with time as commercial or open-source frameworks become available. Another issue is that virtual confederate technology is relatively recent and, therefore, still the object of much research.

Our results also show that it is important to distinguish between virtual confederates that are perceived to be controlled by humans (avatars) from confederates that are perceived to be controlled by computer algorithms (agents). In all three experiments participants demonstrated a basic bias that favored avatars to agents. We argue this distinction is occurring because people categorize avatars as belonging to the human social category, whereas agents are not. In line with findings that people attribute more mind to humans than computers (Blascovich et al., 2002; Waytz et al., 2010), we argue that this human category captures the default expectation people have that computers possess less mental abilities than humans. The literature on dehumanization

shows that people tend to discriminate others that are perceived to have less mental abilities; similarly, people discriminate agents, by offering less money when compared to avatars in the exact same situation. An interesting line of future work, thus, is to test the prediction that proper simulation of affective and mental abilities suffices to make people treat agents in the same manner as avatars, at least in the context of decision-making tasks with clear financial incentives.

We have argued that virtual confederates are a promising tool to research intergroup behavior and we demonstrated, with distinct decision tasks and different kinds of populations, that people can apply social categories to them and show corresponding bias in their behavior. Future work should further explore more decision contexts, more social categories (e.g., age, gender, culture), more roles (e.g., receiver), more operationalizations of bias and conflict, and determine the sufficient conditions agents need to possess in order to be treated in the same manner as avatars.

REFERENCES

Blascovich, J., Loomis, J., Beall, A., Swinth, K., Hoyt, C., & Bailenson, J. 2002. Immersive virtual environment technology as a methodological tool for social psychology. **Psychological Inquiry**, 13(2), 103-124.

Crisp, R., & Hewstone, M. (2007). Multiple social categorization. **Advances in Experimental Social Psychology**, 39,163-254.

Gaertner, S., & Dovidio, J. 2005. Understanding and addressing contemporary racism: Aversive racism to the common intergroup identity model. **Journal of Social Issues**, 61(3), 615-639.

Reeves, B., & Nass, C. 1996. **The media equation: How people treat computers, television, and new media like real people and places**. New York, NY: Cambridge University Press.

Waytz, A., Gray, K., Epley, N, & Wegner, D. 2010. Causes and consequences of mind perception. **Trends in Cognitive Science**, 14, 383-388.

FIGURES

Figure 1. The ethnicities and some of the virtual confederates used in Experiments 1 and 3.