The Map/Territory Relationship in Game-Theoretic Modeling of Cultural Evolution

Tim Elmo Feiten

University of Cincinnati, Department of Philosophy

elmo.feiten@gmail.com

Abstract:

The cultural red king effect occurs when discriminatory bargaining practices emerge because of a disparity in learning speed between members of a minority and a majority. This effect has been shown to occur in some Nash Demand Game models and has been proposed as a tool for shedding light on the origins of sexist and racist discrimination in academic collaborations. This paper argues that none of the three main strategies used in the literature to support the epistemic value of these models—structural similarity, empirical confirmation, and how-possibly explanations—provides strong support for this modeling practice in its present form.

Acknowledgements: I am very grateful for important feedback and comments to two anonymous reviewers, to Melissa Jacquart and Rob Skipper, to all participants of two discussion groups—one organized by Colin Allen and one by Angela Potochnik and Tony Chemero—to Julia Bursten, Jennifer Jhun, and Philip Sink, to the audience at my PSA talk for the invaluable Q&A, especially Cailin O'Connor and Hannah Rubin, and to everyone I failed to mention here.

1. Introduction

Across a series of recent papers, a phenomenon observed within the evolution of Nash demand game (NDG) models and dubbed the “cultural red king effect” (Bruner 2017, 413) is being used to shed light on discrimination in social interactions, specifically in the context of academic collaboration (e.g. O’Connor 2017, O’Connor and Bruner 2017, Rubin and O’Connor 2018, Mohseni, O'Connor and Rubin 2019). Cailin O’Connor and Justin Bruner start their (2017) “Dynamics and Diversity in Epistemic Communities” by summarizing this cultural red king effect: “Bruner (2017) shows that in cultural interactions, members of minority groups learn to interact with members of majority groups more quickly—minorities tend to meet majorities much more often as a brute fact of their respective numbers—and, as a result, may come to be disadvantaged in situations where they divide resources.” (101). This is a great summary of the observed effect, but a potentially misleading account of what exactly Bruner (2017) shows.

The problem is that Bruner (2017) shows that an effect occurs in certain NDG models, and the paper is accordingly titled “Minority (dis)advantage in population games”. What O’Connor and Bruner are claiming in their summary is that Bruner (2017) shows that a specific effect arises from the “cultural interactions” of “epistemic communities” containing “minority groups” and “majority groups”. These two claims are not the same, and what separates them is a central aspect of all scientific modeling: the difference between the model and the target system, or the map and the territory. In moving from Bruner (2017)’s title about “population games” to O’Connor and Bruner’s (2017) “epistemic communities”, the language used to describe the cultural red king effect has tacitly slipped from one side of this distinction to the other. Bruner (2017) very clearly shows an effect occurring within the map, but O’Connor and Bruner (2017) incorrectly suggest that this effect has been shown to occur in the territory.

To move responsibly from the first claim to the second, we would have to draw an inference from the map to the territory: Because we see the effect in the model, we infer that it occurs in the real-world target system, too. We might say that the inference holds if the model represents its target accurately, or—with less technical commitment—that the map gives us reliable information about what we will find when we explore the territory. In this paper, I argue that the failure to differentiate properly between claims about the map and claims about the territory is not a singular mistake, but an overarching problem with the series of papers linking the cultural red king effect to discrimination in academic collaborations.

In the following section, I will give a brief overview of the cultural Red King Effect and how it is linked to discrimination in academic collaboration practices, focusing on Rubin and O'Connor (2018). Next, I will consider and problematize different strategies for showing the epistemic value of these NDG models which are pursued in the literature. used, first assuming that they are meant to represent real-world target systems, then also in relation to alternative ways of thinking about the epistemic value of models.

2. The cultural red king effect and discrimination in academic collaboration

Bruner's (2017) "Minority (dis)advantage in population games" first introduces the cultural red king effect: "[W]e show that when members of two groups must routinely interact so as to divide resources amongst themselves, it is exceptionally likely that an adaptive process such as the replicator dynamic or social learning will lead to a minority disadvantage equilibrium in which members of the larger group demand the bulk of a contested resource when interacting with members of the minority, who in turn acquiesce to the demands of the majority group" (413).

Bruner believes that his “results demonstrate how minorities could possibly come to be systematically disadvantaged in virtue of their minority status alone”, without any differences between the members of the two groups or the presence of prejudice against groups (414). He considers this kind of argument to be following in the footsteps of Schelling’s (1978) famous chessboard model of the emergence of racial segregation in neighborhoods. At first glance, the description of the cultural red king effect quoted above may look as if it is describing interactions between human beings, but it should not be taken as such. Terms like “minority” are used in an equivocal way to refer either to a group of “agents” within the simulation or to a group of human beings out in the world we are striving to understand (414). Although the language is ambiguous, Bruner takes care to clarify that the above statement refers to the map rather than the territory and that he does “not argue that the mechanism uncovered in this paper is in fact responsible for most real-world instances of minority disadvantage” (414).

Bruner’s work has been taken up by a series of papers produced in various constellations by a group of collaborators. These papers propose that the bargaining situation in a NDG can be used to shed light on how workload and authorship are negotiated within academic collaborations, and that the cultural red king effect can help us understand how discrimination and segregation may arise from these negotiations. Several of these papers explore the robustness of the effect under different modeling assumptions and discuss its potential links to the epistemic role of diversity (e.g. Bruner 2017, O’Connor 2017, Rubin and O’Connor 2018). Some papers include reflections on how different policies aimed at combatting discrimination might fare in light of the cultural red king effect (Rubin and O’Connor 2018, but especially Schneider, Rubin and O'Connor 2017). Mohseni, O'Connor and Rubin (2019) reports the results of an empirical study that replicates the cultural red king effect using human subjects playing the NDG.

Rubin and O’Connor (2018) will serve as a helpful example to illustrate key aspects of this literature. Where Bruner (2017) investigates the development of strategies in NDG models under the replicator dynamic, Rubin and O’Connor instead create agent-based models of a population engaging in the NDG. The agents represent academics and the bargaining events "represent division of labor and credit between academic collaborators" (Rubin and O'Connor 2018, 382). The NDG is played between two players, each of whom makes a low, medium, or high demand. If both demand high, they each get nothing, otherwise each gets what they asked for. The population consists of a minority and a majority group and each agent has two different strategies, one for bargaining with their in-group and one for bargaining with their out-group. Based on the payoff of each interaction, the agents have a chance to adjust their strategies or to change who they will play with in the future. Rubin and O'Connor ran three kinds of models to investigate the development of discrimination and segregation that can result from these interactions.

Part A models the change of strategies on fixed networks, where the pairings between agents are determined randomly at the outset and only the strategies change over time. In these models, discrimination against minorities emerges as the result of their minority status, even when the starting strategies do not reflect homophily (in-group preference). Part B uses fixed strategies and investigates how the network structure changes over time. Starting with discriminatory strategies that award the minority lower payoffs, the agents can choose to dissolve a pairing and enter into a new pairing that gives a higher payoff. In these models, minority agents break their out-group pairings in favor of in-group pairings, leaving the majority with only other majority agents willing to pair. The network moves towards complete segregation, with only in-group pairings. In Part C, agents have the chance to change either their strategy or a pairing, leading to the co-emergence of discrimination and segregation.

It is worth noting that, by their very nature, these NDG models can only account for discriminatory effects that emerge within the set of bargaining interactions, without any causal connection to existing systemic discrimination—if indeed such a purely emergent effect exists in the real world. Additionally, the minority groups in question and their social context are not characterized beyond the sheer numeric fact that they have fewer members than the majority. In contrast, existing research on discrimination highlights the "unique features of discrimination against women in academia" and the "systemic origins of racism in academic institutions" (Acorn 2000: 359, Dupree and Boykin 2021: 12).

In light of this contrast, we should ask: How much can NDG models tell us about discrimination and segregation in real-world academia? The next sections discuss a range of stances on the epistemic status of these NDG models found in the literature.

3. Assessing epistemic value

The literature on the cultural red king effect contains a range of subtly different claims and arguments about the epistemic status of the NDG models employed. It is often difficult to identify the exact strength of the claims advanced, and some papers or even passages vary in tone between careful hedging and confident assertions. In the paper that first identified the effect, Bruner (2017) is very careful to limit the scope of his claims. He highlights that he does "not argue that the mechanism uncovered in this paper is in fact responsible for most real-world instances of minority disadvantage", but only that it is "*possible* for minorities to be systematically disadvantaged" as a result of this effect (414). In contrast, Rubin and O'Connor (2018) argue that an NDG model is a "good representation" of academic collaboration practices, while O'Connor and Bruner (2017) emphasize that the effect occurs robustly across different modeling assumptions and that "[r]obust results are arguably more likely to be genuinely explanatory of real world phenomena" (15). In order to get clear on the different claims that are being made and how well they are supported, I will separate the arguments made for the likelihood of the cultural red king effect to occur in the real-world academia into three groups dealing with structural similarity, empirical confirmation, and how-possibly explanations.

3.1 Structural similarity between map and territory

One way or arguing for the epistemic value of NDG models is to claim that "the aspects of the models here that lead to disadvantage are features of the real world" (O'Connor and Bruner 2017, 115). According to this view, the cultural red king effect offers a "how-potentially" story for the emergence of discrimination in academia, meaning that it is an "effect that has the potential to really occur" (O'Connor and Bruner 2017, 116). Bruner (2017) advances such a view, albeit in a carefully hedged formulation: "We have shown how it is possible for divisional norms disadvantaging minority groups to frequently take root in a community whenever individuals (i) can condition their behavior on group membership and (ii) learn to act in a way which promotes their self-interest in strategic contexts. These conditions both seem likely to hold, suggesting the cultural red king effect may crop up in many real-world situations." (Bruner 2017, 426)

At the heart of this argument is the claim that there exists an important structural similarity between the map and the territory. It is useful to compare this claim to Michael Weisberg's (2012) discussion of similarities between models and their target systems. Weisberg differentiates between attributes and mechanisms of each, and argues that "models are similar to their targets when they share many, and do not fail to share too many, features that are thought to be salient" (794). His primary example is Schelling's (1978) chessboard model of neighborhood segregation and he argues that "Schelling's model would be informative if it shared important features with the real population dynamics of Philadelphia" (794). He also links this similarity between model and target system to the difference between a how-possibly explanation and an account that is part of the actual explanation of a target phenomenon (786). In the case of NDG models, an example of an attribute would be the occurrence of discrimination, and examples of mechanisms would be the structure of a bargaining-interaction, the updating of a strategy, or the dissolution of a pairing. There seem to be good reasons to doubt that fundamental mechanisms of the model are shared by the target and, conversely, seemingly essential mechanisms of the target are absent from the model.

The first question is whether the way that workload and authorship are negotiated in academic collaboration is the same mechanism as NDG bargaining. In NDG bargaining, each participant makes a hidden demand, then they are revealed and if the sum of demands is too high, the collaboration fails. There is no chance for reacting to the demands of the other by adjusting one's own demands, so this form of bargaining arguably does not meet the minimum criteria for how we define a negotiation. Furthermore, the way actual workloads and authorships are decided can take a variety of forms. Sometimes the most powerful collaborator just decides the order of authorship, or it is determined by existing, but potentially implicit, social norms. Sometimes coauthors make a good-faith attempt to determine who needs publication credit the most and who is best suited for each part of the work. It is not a priori obvious that actual academic collaborations are generally organized via utility-maximizing bargaining at all.

Assuming that individual utility-maximization is the driving force behind this interaction, there seem to be strong constraints on real-world academics that render successful strategies in the model completely unviable. If the only risk were the failure of this particular project, as it is in the model, individuals might learn a strategy for optimizing the work/credit ratio. But in reality, the effects of demanding something your coauthors are unwilling to grant you go far beyond the paper in question. The crucial difference is that agents in the model only have two strategies, one for in-group and one for out-group interactions. In stark contrast, human beings remember past interactions with specific individuals and adjust their behavior accordingly. In the model, an agent can just make high demands to see what happens and learn from that, with no consequences beyond this instance of negotiation. In reality, making high demands that end up wrecking a potential collaboration will reduce your chances for future collaborations, thus making a whole category of learning-trajectories that can occur in the model unfeasible. Because the utility of any single collaboration is not just individually relatively unknown to actual academics, but the opportunities for future collaborations and the effects that these future collaborations will have on one's career both depend in part on the outcome of the present project, it seems unrealistic to assume that academics have accurate knowledge of the utility of successful collaborations before they embark on the next one.

Neither is it obvious without further investigation that the kind of learning that happens in NDG models is at all possible within the structure of professional academia. In his discussion of related game theoretic claims about the evolution of morality, Arnon Levy (2011) distinguishes between two kinds of learning, success learning and source learning. In success learning, "individuals adopt new behaviors in a payoff-driven fashion", where payoff is understood as "broadly akin to utility" (179). This notion of success learning encompasses a cultural interpretation of the replicator dynamic, but also the kind of learning used in the agent-based NDG models where agents either adopt the strategy that would have given them the highest payoff last round, or adopt the strategy of another more successful agent. This is contrasted with source learning, where "one follows in the footsteps of salient individuals", e.g. "parents, teachers, superiors, celebrities" (180). Levy argues that source learning has been unduly neglected in the modeling of the emergence of morality citing arguments from psychology and anthropology that emphasize the importance of source learning (181-182). Since Levy focuses on the same kinds of models that produce the cultural red king effect—discussed as "divide the cake" games, which is a synonym for NDG—his points may be relevant for our discussion, too (172). Perhaps a far stronger claim can even be made in the case of academic collaboration: source learning sounds plausible, whereas success learning sounds implausible. In Rubin and O'Connor's (2018) models, the "[s]imulations were run for 1,000 rounds" (386).

In real academia, the opportunities for learning how to navigate collaborations successfully is several orders of magnitude lower. Even if junior academics had complete knowledge of their payoffs in any potential collaboration, the need to get it right very quickly makes it more plausible that learning is largely and especially at first a matter of source learning.

At first glance, Bruner's (2017) description seems to apply reasonably well to both map and territory: "individuals (i) can condition their behavior on group membership and (ii) learn to act in a way which promotes their self-interest in strategic contexts" (426). However, instead of shared features, a closer look reveals very different mechanics at work in the NDG models and our (pre-theoretic) understanding of academic interactions. We trust that the San Francisco Bay Model informs us reliably about the effect of potential infrastructural interventions because the model is "three-dimensional and proportional" and filled with actual water (Potochnik, Colombo and Wright 2019, 89). Because these two elements and the way in which they are combined all appear very similar to the real San Francisco Bay, we trust that it is a reliable map of the territory. In contrast, none of the fundamental aspects of NDG models discussed above individually appear to be very similar to their counterparts in the target phenomenon.

3.2 Empirical support

A further strategy for supporting the epistemic value of the cultural red king effect focuses on empirical confirmation. Mohseni, O'Connor and Rubin (2019) take this route and motivate their approach like this: "Of course, as noted, highly simplified models of social interaction cannot usually be taken at face value as explaining real social phenomena. One important epistemic role they can play is directing attention to processes that might be occurring in the real world, and which merit further empirical investigation. For this reason, we study the cultural Red King effect in the laboratory" (2).

This sounds highly promising, as it seeks to answer the question "can the cultural Red King really occur in human groups" (2)? However, the import of this experiment for our concerns is diminished by the fact that it consisted in human subjects playing the NDG. The "results show that minority groups do end up at payoff-inferior outcomes with significantly greater frequency", so the cultural red king effect occurs not just in computer simulations of the NDG but also when real humans play the NDG (16). We now know that NDG models can reliably inform us about the dynamics of real-world NDG, but that is not really what we were interested in. Our main question is still open: Which human practices besides NDG themselves can be usefully modeled by NDG, and is the process of distributing the workload and authorship of academic collaborations one of them?

Empirical investigation of this main question seems both far more important and far more difficult. Indeed, O'Connor and Bruner (2017) argue that part of the epistemic role of such models is that they "can direct our attention to an effect that has the potential to really occur, and prompt further investigation, both theoretical and empirical, into this effect" (116). Similarly, Rubin and O'Connor take their results to "generate a theoretically well-grounded hypothesis for why we see such patterns" in epistemic communities (400). Although it is not made explicit in the paper, it seems plausible that the hypothesis they have in mind would go something like "the cultural red king effect causes some of the discrimination we observe in academic collaboration practices". An important task for cultural-red-king-research would then be to devise ways of rendering this hypothesis empirically testable. Rubin and O'Connor note that "[i]t is notoriously difficult to generate empirical data testing cultural evolutionary pathways", which is why they "instead employ game theoretic models" (382). While an "empirical turn" in the literature on formal models within philosophy of science—as called for by Edouard Machery (2022)—might improve their epistemic credentials, the amount of additional work this would require seems to negate a large part of what makes these models so attractive in the first place: their extreme simplicity and tractability.

3.3 How-possibly explanations

One common phrasing in the literature on the cultural red king effect casts it as a how-possibly explanation for discrimination occurring in human social interactions. Bruner takes it to "demonstrate how minorities could possibly come to be systematically disadvantaged" (414). O'Connor and Bruner (2017) describe it as "a version of 'how-possibly' modeling—it specifies under what conditions a surprising effect can possibly arise" (116). They also state that they "use evolutionary game theoretic methods to show that minority groups can end up disadvantaged in academic interactions like bargaining and collaboration as a result of this effect" (101). What kind of possibility do these phrases express? And do they all express the same kind?

In order to clarify this subtle and vexing question, we can import a distinction between causal possibility and factual possibility made by Ylikoski and Aydinonat (2014). They are contributing to a longstanding discussion in the philosophy of economic models on "the epistemic import of highly abstract and simplified theoretical models" (19). Their main example is Schelling's (1978) chessboard model and many of the moves made in this debate are identical to those in the debates about the cultural red king effect, making their work very useful for our purposes. In order to distinguish between different kinds of how-possibly explanations, Ylikoski and Aydinonat introduce the notions of causal and factual possibility. "First, there is the question: if the scenario had taken place, would it have produced the effect to be explained? Let us call this causal possibility" (26). Causal possibility is determined based on our "general causal knowledge" (26). In contrast, factual possibility also requires "compatibility with facts known about that particular causal history" (26). A scenario can only be considered factually possibly if it "is consistent with the facts we know concerning the particular explanandum phenomenon at hand" (26). Only if a scenario is both causally and factually possible can it be "considered a possible explanation of the particular fact at hand" (26).

If we interpret part A of Rubin and O'Connor (2018) as a scenario that offers a how-possibly explanation for a particular occurrence of discrimination in academic collaboration, we could judge that it is causally possibly: if the scenario had taken place, the cultural red king effect would have occurred. However, we could also judge that it is not factually possible: part A starts with a population of individuals who behave in completely non-discriminatory ways, and we know with reasonable certainty that no present-day academic collaboration network developed from a past stage in which academia was free from discrimination based on race and gender. Our take-away from this is that not all how-possible explanations are also factually possible. It is not enough to outline an explanation that is causally possibly for it to be "considered a possible explanation of the particular fact at hand" (Ylikoski and Aydinonat, 26). It also has to be consistent with the known facts about the particular situation and its causal history.

I have argued that the scenario in part A does not meet Ylikoski and Aydinonat's criteria for a "possible explanation of the particular fact at hand" (26). But it might be the case that this is a more general limitation, which includes all models in Rubin & O'Connor (2018) and all game theoretic models of comparable complexity. For example, all interactions between two "agents" in Rubin & O'Connor (2018) are both fully private and anonymous. Both participants can use the outcome to adjust one of their own two strategies, but neither of them can take in information about the other's behavior as tied to a specific individual, nor can anyone else in the network. This is fundamentally and importantly never the case in real academic interactions. Of course, much more complex models could be built, but changing them dramatically enough that they meet Ylikoski and Aydinonat's criterium of "factual possibility" would arguably make them a different kind of model altogether than the NDG.

4. Conclusion

Within just a few years, the discovery of the cultural red king effect in NDG models has kicked off a wealth of work on the emergence of discriminatory norms in epistemic communities. Part of this research promises to enhance our understanding of discrimination in academic collaboration practices in particular. However, the relationship between these models and the real-world phenomena to which they are being linked is still not entirely clear. Whether or not this work actually improves our understanding depends crucially on this map/territory relationship.

Attempts to secure the epistemic value of this kind of modeling work have taken various strategies. It is claimed that the essential structural features that produce the cultural red king effect are shared between map and territory. No systematic effort to substantiate this claim exists and it seems prima facie implausible. Empirical support for the effect exists in the form of a successful experimental study. Importantly, this study only examines whether the cultural red king effect can occur in NDG played by human subjects instead of computer simulations, not whether it can occur in human interactions outside of the activity of playing the NDG. This latter issue is the kind of empirical support that could actually help establish a link between the map and the territory, but there is no proposal as of yet for how it might be attained. Finally, interpreting the relationship between the NDG models and real-world epistemic communities as a how-possibly explanation is used in the literature as a strategy to avoid stronger epistemic claims which cannot be justified. By using a distinction between causal possibility and factual possibility, we can show that the kind of possibility that is supported by the existing modeling work has no direct connection to real world scenarios.

In writing philosophical papers, we make careful technical points, but we also convey an overall story through the style and structure of our narrative. I am worried that the enthusiasm about a fascinating effect in a formal model leads some authors to tell a story that runs the risk of misleading its readers. NDG models are highly abstract and have been used to represent a vast range of human interactions. We have no solid reason, theoretical or empirical, to think that the cultural red king effect tells us anything about academic discrimination in the real world and until we are certain that we have found one, we should be very careful not to imply that we have. Otherwise, we risk a distorted view of both the phenomenon we are trying to understand and of the method we are applying to that end.

References

Acorn, Annalise E. 2000. “Discrimination in Academia and the Cultural Production of Intellectual.” *UCLA Women's Law Journal,* 10(2), 359-371.

Bruner, Justin P. 2017 (published online). “Minority (dis)advantage in population games.” *Synthese* (2019) 196: 413-427.

Dupree, Cydney H. and C. Malik Boykin. 2021. “Racial Inequality in Academia: Systemic Origins, Modern Challenges, and Policy Recommendations” *Policy Insights from the Behavioral and Brain Sciences,* Vol. 8(1) 11-18.

Grüne-Yanoff, Till and Verreault-Julien, Philippe. 2021. “How-possibly explanations in economics: anything goes?” *Journal of Economic Methodology* 28(1), 114-123.

Levy, Arnon. 2011. “Game theory, indirect modeling, and the origin of morality.” *The Journal of Philosophy* 108(4), 171-187.

Machery, Edouard. 2022. “Formal Modeling in Philosophy of Science – Let’s be realistic!” talk delivered at the Center for Philosophy of Science, Pittsburgh, on Jan 18 2022.

Mohseni, Aydin, Cailin O'Connor and Hannah Rubin. 2019. “On the Emergence of Minority Disadvantage: Testing the Cultural Red King Hypothesis.” *Synthese.*

O’Connor, Cailin. 2017. “The cultural Red King effect.” *The Journal of Mathematical Sociology* 41(3), 155-171.

O’Connor, Cailin and Justin P. Bruner. 2017 (published online). “Dynamics and Diversity in Epistemic Communities.” *Erkenntnis* 84(1), 101-119 (2019).

Potochnik, Angela, Matteo Colombo and Cory Wright. 2019. *Recipes for Science: An Introduction to Scientific Methods and Reasoning.* New York: Routledge.

Rosenstock, Sarita, Justin Bruner and Cailin O'Connor. 2017. “In Epistemic Networks, Is Less Really More?” *Philosophy of Science* 84(2), 234-252.

Rubin, Hannah and Cailin O’Connor. 2018. “Discrimination and Collaboration in Science.” *Philosophy of Science* 85(3), 380-402.

Schelling, Thomas C. 1978. *Micromotives and macrobehavior.* London: W. W. Norton.

Schneider, Mike D, Hannah Rubin and Cailin O'Connor. 2017. “Promoting Diverse Collaborations.” working paper.

Weisberg, Michael. 2012. “Getting Serious about Similarity.” *Philosophy of Science* 79, 785-794.

Ylikoski, Petri and N. Emrah Aydinonat. 2014. “Understanding with theoretical models.” *Journal of Economic Methodology* 21(1), 19-36.

Zucker, Julian, Daniel Rassaby, Aja Watkins, and Rory Smead. 2019. "Bargaining and Intersectional Disadvantage: Reply to O'Connor, Bright, and Bruner." Social Epistemology Review and Reply Collective 8(7): 1-8.