

An Agent-based Model to Evaluate Interventions on Online Dating Platforms to Decrease Racial Homogamy

Stefania Ionescu
ionescu@ifi.uzh.ch
University of Zürich
Zürich, Switzerland

Anikó Hannák
hannak@ifi.uzh.ch
University of Zürich
Zürich, Switzerland

Kenneth Joseph
kjoseph@buffalo.edu
University at Buffalo
Buffalo, NY, USA

ABSTRACT

Perhaps the most controversial questions in the study of online platforms today surround the extent to which platforms can intervene to reduce the societal ills perpetrated on them. Up for debate is whether there exist *any* effective and lasting interventions a platform can adopt to address, e.g., online bullying, or if other, more far-reaching change is necessary to address such problems. Empirical work is critical to addressing such questions. But it is also challenging, because it is time-consuming, expensive, and sometimes limited to the questions companies are willing to ask. To help focus and inform this empirical work, we here propose an agent-based modeling (ABM) approach. As an application, we analyze the impact of a set of interventions on a simulated online dating platform on the lack of long-term interracial relationships in an artificial society. In the real world, a lack of interracial relationships are a critical vehicle through which inequality is maintained. Our work shows that many previously hypothesized interventions online dating platforms could take to increase the number of interracial relationships from their website have limited effects, and that the effectiveness of any intervention is subject to assumptions about sociocultural structure. Further, interventions that are effective in increasing diversity in long-term relationships are at odds with platforms' profit-oriented goals. At a general level, the present work shows the value of using an ABM approach to help understand the potential effects and side effects of different interventions that a platform could take.

CCS CONCEPTS

• **Applied computing Sociology;**

KEYWORDS

racism, social media, dating platforms, agent-based modeling

ACM Reference Format:

Stefania Ionescu, Anikó Hannák, and Kenneth Joseph. 2021. An Agent-based Model to Evaluate Interventions on Online Dating Platforms to Decrease Racial Homogamy. In *Conference on Fairness, Accountability, and Transparency (FAccT '21)*, March 3–10, 2021, Virtual Event, Canada. ACM, New York, NY, USA, 12 pages. <https://doi.org/10.1145/3442188.3445904>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

FAccT '21, March 3–10, 2021, Virtual Event, Canada

© 2021 Association for Computing Machinery.

ACM ISBN 978-1-4503-8309-7/21/03...\$15.00

<https://doi.org/10.1145/3442188.3445904>

1 INTRODUCTION

Online platforms have been criticized for their role in a profusion of societal ills, from the increased spread of political misinformation [4], to extrajudicial killings in the Philippines [50]. Platforms have sought to address these issues with various *interventions*, from attempts to filter or restrict various forms of content (e.g. hate speech) [31] to attempts to decrease discriminatory practices by users [16]. Few, if any, would argue that these interventions have fully addressed the problems at hand. Underlying agreement to this high-level assertion, however, lies a more debated question—*what are the limits of a platform's ability to intervene on society's ills?*

Given the documented problems platforms have helped give rise to, it would be wrong to assume that they have no role to play in positive change. At the same time, however, we must be careful to ascribe too heavy of a causal arrow to platforms. Doing so is destined to lead to an ill-informed aura of technological determinism [44], where we search for solutions to societal problems by endlessly tweaking the structure of our technology, ignoring other root underlying causes [20].

How might we understand the limits to a platform's ability to intervene to reduce a given societal ill? Empirical work, especially based on natural experiments [42] and using algorithmic auditing [28, 54], is critical, but is restricted to studying the platform as it currently exists. Exploration of the effects of interventions is thus limited to what platforms are willing and able to implement given their profit-oriented goals. Even beyond these profit-related restrictions, there are ethical questions about the use of experimentation without the consent of platform users [59].

Even still, scholars have pursued ethical empirical research not bound by profit-oriented goals, for example, by constructing platforms of their own [13, 45]. But doing so is time-consuming and expensive; it would be useful if we could first scope potential solutions and their potential side effects in a more rapid and expansive way. Social theory can partially fulfill this role [51], but it can be difficult to theoretically characterize, and much less operationalize, the myriad causal factors underlying sociotechnical systems.

What is needed is a methodology that exists in this middle ground between difficult empirical work on the one hand, and often abstract theory on the other. Simulation has long filled this role of bridging theory and empirical work in the social sciences [18, 25, 55], and more recently has proved useful in the study of algorithmic fairness. For example, [41], and later [17], use simulation to show concretely how the abstract notion of feedback loops in predictive policing can lead to settings where police unfairly target one population over another, despite equal crime rates. Simulation has the added benefit of being applied to hypothetical people, allowing us to carefully and fully consider the harms that may be caused by a particular

intervention before it has any real effects. In the present work, we use simulation, and specifically, an *agent-based model (ABM)*, to make concrete a set of proposed interventions to address a given societal ill on a given online platform, and to provide new ideas about their implications and side-effects.

Online Dating. Specifically, we construct an ABM of an online dating platform and assess how interventions theorized recently by Hutson et al. in their award-winning CSCW paper [34] affect the degree of racial homogamy in a simulated society. *Racial homogamy* is defined as the *tendency of individuals to seek out and engage in long-term relationships with individuals who are the same race*.

Critically, we do not argue that racial homogamy is in and of itself a societal ill; there are many valid reasons why individuals, particularly those identifying with an oppressed racial group, may want to maintain racial homogamy in their relationships. However, racial homogamy has been well-established to contribute to the reinforcement of racial inequality by increasing racial segregation [3] and racialized income inequality [26]. Hutson et al. [34] argue that racial homogamy is hence worthy of dating-platform intervention, as it is one racializer of economic and social inequality [34].

After detailing this connection between racial homogamy and racial inequality, Hutson et al. [34] explore an array of mechanisms within online dating platforms that present opportunities for interventions to reduce racial homogamy. For example, “search, sort, and filter tools” allow users to perform initial searches for romantic partners, and thus might encourage users to make particular decisions about the race of their partner. Hutson et al. therefore take, at least implicitly, the stance that platforms could construct interventions that reduce racial homogamy. Others, however, dispute this claim. In particular, Anderson et al. [2] study racial bias in online dating from a sociological perspective, and conclude that “merely altering the structural environment [on the platform]—say, by creating more opportunities for individuals of different races to interact—would not necessarily ameliorate persistent patterns of racial homogamy in romantic relationships” (pp. 38–39). This claim puts the ideas presented in Anderson et al.’s work in contrast with those presented by Hutson et al., who explicitly argue that structural changes to platforms can reduce racial homogamy.

These two papers therefore present a difference of opinion on the extent to which interventions on online dating platforms can impact racial homogamy, and ultimately racial inequality. However, empirical analyses of this debate is difficult for the reasons listed above. We therefore employ a simulation model to bridge the middle ground between Hutson et al. [34] theorized interventions and potential future empirical work. Our work helps to operationalize the theoretical constructs and to explore their impacts and potentially unintended consequences in an *artificial*, but *empirically and theoretically grounded*, complex sociotechnical system.

Contributions. The present work makes three contributions:

- We provide a blueprint (including an open-source modeling framework)¹ for the use of agent-based modeling to help operationalize theory and explore the potential impacts of interventions of online platforms

- Specific to racial homogamy and online dating platforms, we find that various interventions theorized by Hutson et al. [34] indeed decrease racial homogamy in our simulated society according to our selected measure by as much as 10%
- However, the benefits of these interventions are dependent on specific assumptions about the sociocultural structure of that society, and are often at odds with societal outcomes and profit-oriented goals of the platform

Like many social science endeavors [64], our core findings present claims that, once stated, seem quite obvious. However, given their motivation from established differences of opinion in published literature by leading scholars, our work serves a critical role in the path forward to understanding the potential of platforms to intervene to reduce racial homogamy, and as a broader methodology to address similar questions in other domains.

2 RELATED WORK

2.1 Online Dating

Online dating plays an increasingly central role in the process of finding romantic partners, with 15% of Americans reporting to have used an online dating site [11]. Indeed, at least for heterosexual couples, online dating has become the predominant means of meeting romantic partners [53]. This change in the means for dating has opened the door to a debate on whether platforms can disrupt patterns of assortative mating, and in particular decrease racial homogamy. While Hutson et al. [34] argue that platforms have the means of decreasing racial homogamy through interventions, Anderson et al. [2] think that this would likely not be the case as homogamy is partly caused by same-race preferences which cannot be changed through opportunity alone. These preferences influence and are influenced by social norms, and their formation is based on imitation and recurrent behaviour [48] and enforced through social verification [29]. To capture this phenomena, we consider norms and their impact on preferences. In alignment with Hutson et al. [34], we acknowledge that platforms can potentially intervene to change norms on the platform, so, in our model, we differentiate between offline and online norms.

The process people go through while dating online is complex in multiple ways. First, one goes through a variety of different “phases” of dating, from seeking information about dating platforms to developing a long-term relationship [19]. Second, the decision-making process depends on distinct factors such as the perceived level of compatibility [9], how much we can validly infer about others online [22], and one’s own personal stereotypes [37].

The complexity of studying dating have naturally given rise to the use of simulations. Previous agent-based models for dating have looked at, for example, how the selection of partners is influenced by space and mobility [60], and whether assortative mating patterns could explain previously observed statistics on marriages and divorce [32]. We build upon these examples in our work, but focus on the novel problem of understanding the role of platform interventions on racial homogamy in the context of online dating.

2.2 Uses of Simulation to Study Fairness

Starting with the seminal work of Schelling [55] on housing segregation, simulation has been widely leveraged to explore how

¹Repository link: <https://github.com/StefaniaL/ABM-for-Online-Dating.git>.

inequality and unfairness can arise and be combated in social systems. These studies have provided useful fodder for the more recent application of simulation to questions of *algorithmic* fairness. For example, Martin Jr. et al. [43] use a type of simulation called System Dynamics Modeling [36] to show how simulation can help us to think through the complex impacts of different interventions in the context of credit lending. Others have leveraged different kinds of simulation models to study lending in similar ways, for example, within a reinforcement learning paradigm [14]. Still others have used simulation beyond the context of lending. For example, various forms of simulation have been used to study fairness and feedback loops in the context of predictive policing (as noted above [17, 41]) the effects of recommender systems [6], and income inequality for drivers on rideshare platforms [5].

Like Bokányi and Hannák [5], we use an agent-based model here. Agent-based modeling involves the construction of virtual worlds in which artificial agents interact according to a set of pre-specified rules [18]. While behavior is thus predictable at the individual level, ABMs are useful for understanding how these pre-specified micro-level behaviors result in emergent properties, e.g., social biases across social groups [35] or opinion polarization [56].

Our work compliments existing literature using simulation modeling to study questions of fairness and equality in three ways. First, we study a concrete, contemporary, and contested question about the role that online dating platforms can play in reducing racial homogamy. Second, rather than focus on algorithmic fairness specifically, we emphasize other ways in which online platforms can have impacts on society. Finally, and at the same time, we develop an agent-based model that can be easily extended to study algorithmic fairness, complementing existing System Dynamics and reinforcement learning platforms provided by prior work [14, 43].

3 MODEL

Our model contains a variety of parameters, listed in Table 1 alongside references, where applicable, to were used to determine their values. Below, we first provide an overview of the model, and then further details on each major element of the model.

3.1 Model Overview

Agents in our model are people that have at some point used our simulated dating website. Agents are assumed to carry out a series of activities on the site in pursuit of a romantic partner. These activities include creating a profile (Figure 1a), searching for a partner (b), browsing the list of potential partners returned by that search (c), and messaging (d). Agents in our model perceive other agents according to their *attributes*. As in most ABMs [24], attributes are just a vector of 0s and 1s, but are assumed to represent real-world traits that might be important in the context of dating, e.g. partisanship. All agents also have a “protected attribute,” that is, an attribute relevant to questions about discriminatory dating practices. We will focus on race as the protected attribute here. Moreover, we will focus on a binarized model of race, that is, agents can be only one of two races in the model. In the context of prior work we discuss above, the two races of interest here are White and Black. Section 3.2 provides full details on how agents are modeled.



Figure 1: The actions agents can take on the platform

As shown in Figure 2, relationships between agents evolve over a series of *phases*, from finding (or being found by) a potential partner, to interacting with each other on the platform via messages (online relationship), and finally to an offline relationship [19]. At any phase, a relationship might end, at which point the agents will resume their search for another romantic partner, or exit the platform. If a relationship continues for long enough, it becomes a *long-term* relationship. Our primary outcome measures focus on racial homogamy in these long-term relationships. and Section 3.3 provides complete details on the different relationships phases.

The simulation is carried out via a sequence of iterations, or *turns*, with one turn being loosely the equivalent of one day. On the first turn of the simulation, 300 agents are initialized and become the population of users on the online dating site. On each subsequent turn, we add four new agents, modeling a constant flow of new users. After new agents are added, all agents get a turn to act. During their turn, agents can 1) conduct a search, potentially using a set of search *filters* provided by the site, 2) send a message, 3) start a new relationship, 4) continue or end an existing relationship, or 5) leave the platform, either because they have found a long-term partner, or because they have become frustrated with the site.

As shown in Figure 2, which of these things agents decide to do is based on which part on the relationship phase they are in, and on the *decision model* we construct for agents. The decisions made by an agent are determined by 1) their attributes, 2) the attributes of their potential (or current) partner, 3) the *stereotypes* they hold of how different attributes are related, and 4) their *preferences* for others with certain attributes. These preferences are themselves informed by *social norms* that exist both on and off the platform. Section 3.4 provides details on how agents make decisions.

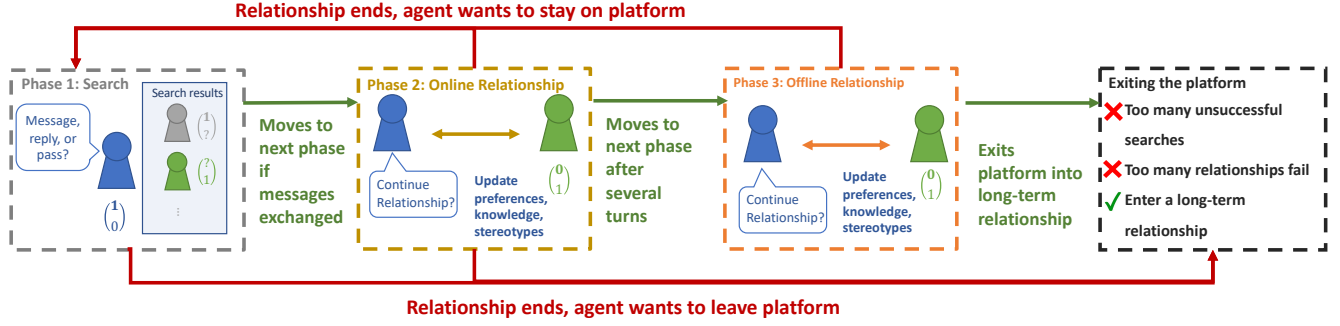


Figure 2: Overview of relationship phases and transitions in each turn of an iteration, explained through the blue agent’s choices and actions. Quote boxes represent decisions the blue agent makes, and the 0/1 vectors represent the agent’s attributes, or the blue agent’s knowledge of potential partner’s attributes.

Table 1: Tabular description of model parameters (Left) and the values taken in the virtual experiment (Right). For parameters with a red-colored value, we simulate each value for all simulated interventions but keep the values of the other parameters to red values (see Section 4).

Parameters	Values Taken
Varied in Virtual Experiment	
Tolerance to un-successful search turns	20/ 25 / 30
Out-platform norms for protected attributes	0/0.1/ 0.2 /0.3
Strength of updates: norms → pref.s	0/ 1% / 5%
Initial degree of ethnocentrism	0/ 0.25 / 0.5/ 1
Correlations in attribute generation: β, γ	0/ 0.4 /0.6/0.8, 0.2 /0.6
Fixed in Virtual Experiment	
Strength of updates: interaction → pref.	2%
Strength of updates: average pref. → norms	5%
Relative importance of out-platform norms when generating initial preferences	20%
Failure tolerance to unsuccessful relationships	20
# initial acquaintances in-group & out-group	300, 100 [35, 46, 61]
Weight of searchable vs experiential attributes	1 - 3 [22]
Pr. searchable attribute specified on profile	50% [38]
Pr. learning searchable & experiential attributes	50% [19]
# rounds until offline & long-term	7, 37 [10, 52]
Pr. to meet while offline	14%[10, 47]
Pr. to consider a message (while in search)	50%[19]
Maximum # agents considered in a search turn	30 [19, 22]
# iterations	2000
Initial population size	300
# agents added per iteration	4

Agent’s attributes are static in our model. However, their stereotypes and preferences *evolve* as they interact with others and move through relationship phases. The social norms governing all agents also evolve throughout the simulation. Section 3.5 details when and how these updates to stereotypes, preferences, and norms occur.

3.2 Agents

As in all ABMs, our agents have state and can take actions. Agents’ states encompass their *type*, *attributes*, *stereotypes*, *preferences*, and

knowledge of others. The actions agents take, based on these states, are a set of *decisions*, which we describe in Section 3.4.

3.2.1 Type. Agents can be of two types, and are only attracted to agents of the opposite type. We thus loosely mimic a hetero- and mono-normative online dating website, as was studied by Anderson et al. [2]. We return to this limitation in our conclusion. Agents are assigned to the two modeled types with equal probability.

3.2.2 Attributes. Each agent i has a vector of binary attributes, $a^{(i)}$. We distinguish between race, $a_0^{(i)}$, and K other attributes. Following theory in the online dating literature [9, 19], we make two additional distinctions among attributes. First, we distinguish between *matching* and *competing* attributes. For *matching* attributes, agents try to find partners with similar values, while for the *competing* attributes, they try to find partners with the highest possible values. We assume that the protected attribute is matching.

Second, we distinguish between attributes visible only offline as *experiential* attributes, and ones that are visible both online and offline as *searchable* attributes. Examples of searchable attributes include personal interests and location. Sense of humor is an example of experiential attribute [19]. Since people do not share everything about themselves in their profiles [38], agents have a probability of 50% to specify each searchable attribute on the platform. In the example from Figure 1, the agent chose to specify their music taste, but not their race. We describe the complete process of generating attributes for new agents in the Supplementary Materials.

3.2.3 Stereotypes. Agents’ *stereotypes* encompass their beliefs of *which attribute combinations are more likely to be encountered* [37]. We model the stereotypes of agent i as a matrix $G^{(i)}$. This matrix is initialized via a theoretically grounded model that is also based on prior work. Specifically, when generating a new agent, we first generate a sample of 300 agents having the same race, and 100 agents with a different race. The imbalance in sample sizes is motivated by homophily: people are, in general, more likely to meet and know people similar to themselves [46]. In addition to homophily, we also model in-group favoritism, i.e. that people tend to look favorably on people like them [30, 61]. To capture this in our model, we follow prior agent-based models [35] and assume that agents also have an initial degree of ethnocentrism. For example, as in

[35], if the degree of ethnocentrism is 0.3, then the agent will, on average, expect 30% of the attributes of agents in the other-race sample to have values that the agent does not like.

The Supplementary Materials contain full details on stereotype initialization, and Section 3.5 on how stereotypes evolve over time.

3.2.4 Preferences. An agent’s *preferences* define which attributes they believe are most important in a potential partner. We model the preferences of agent i as a weight vector, $w^{(i)} \in \Delta^K$. This vector of preferences is in the K th simplex, i.e. its entries sum to one, thus capturing the *relative* importance of different attributes in the eyes of the agent. For example, a weight vector with $w_j^{(i)} = 1$ and all other entries of 0 corresponds to an agent who only considers the j th attribute to be important in a potential partner.

The preferences of agents influence and are influenced by two different types of social norms, i.e. sociocultural regulations on one’s preferences [39].² First, while interacting online, an agent impacts and is impacted by *platform norms*. Second, when in an offline relationship, the agent preferences inform and are informed by off-platform norms. Critically, then, in our model as in the real world [39], agent’s preferences thus vary in part based on social norms, which change as they move between the context of the platform and their offline social life. At the same time, norms are aggregations of the preferences of individuals, except, as we will discuss, when explicit efforts are made by the platform to fix norms.

Agent preferences are initialized based on a procedure common in the decision-making literature [12]. We note here that two parameters govern important aspects of this process. First, a parameter β is used to capture the level of correlation between race and other attributes. Second, γ captures the level of correlation between the non-protected attributes. Varying these two parameters changes the predictability of one attribute, given another. In our virtual experiment, both β and γ are changed. Most importantly, β is varied to reflect different degrees to which race can be inferred by agents based on other available attributes when the agent chooses not to display race in their online profile.

3.2.5 Knowledge of other agents. As they interact with another agent, agents can become aware of the values of new attributes of that agent. This knowledge is used to make decisions about, e.g., whether to continue relationships. How agents acquire this knowledge is discussed in Section 3.3.

3.3 Relationship Phases

Agents can be in one of three relationship phases: 1) searching, 2) in an online relationship, or 3) in an offline relationship.³

3.3.1 Searching phase. When first entering the platform, an agent is in the searching phase. An agent can search in two ways. First, the agent can perform a query for potential partners using a search interface. Second, the agent can look through messages in their inbox that they have received from other agents. In one turn of the simulation, agents can perform up to 30 search actions, mimicking the typical length of a user’s session on an online dating website [19, 22]. A search action represents looking at either a single search

result or a single message in the agent’s inbox. The agent randomly decides which of these two actions to take. If the agent decides to view a profile from their search result list, they look at the profile of the next search result and decide whether or not they should send that person a message. If the agent decides to view a message in their inbox, they look at the oldest un-read message and consider whether or not they should reply.

Search results are a randomly ordered subset of other agents currently on the platform. Importantly, however, the agent *decides which other agents will appear in their search results*. They do so by *filtering* for potential dating partners. The agent filters on a single attribute, the one that, on this turn, is most important to them (i.e. for agent i , the attribute with the highest value of $w^{(i)}$). In Figure 1, for example, interest in music is the most important factor for the blue agent, and they hence use this as a filter. The blue agent is then only shown profiles of other agents having a passion for music.

If the agent decides to reply to a message in their inbox, their search ends, and the two agents involved proceed to an online relationship phase. However, the agent can send as many initial messages to agents in their search results as they would like. An *unsuccessful search* is considered to be a turn in which the agent is not interested in answering any of the messages they have looked at, and in which they did not find at least half of the profiles they’ve looked at interesting enough to send an initial message. Too many unsuccessful search turns would make the agent exit the platform.

3.3.2 Online relationship phase. After answering or getting a reply to an initial message, the agent moves to an Online Relationship Phase with another agent. While in this phase, in each iteration the two agents within the relationship have an online interaction. During this interaction, three things happen. First, agents may update their knowledge about the attributes of the other. Going back to the example in Figures 1 and 2, when entering the relationship, the blue agent only knows the value of the green agent’s interest in music. After one interaction, they have also learned the self-identified race of the green agent. Second, agents update their stereotypes based on what they learn about the other; see Section 3.5. Finally, agents decide whether or not to continue the relationship. If both decide to continue, the interaction is *positive*, and the relationship will continue with a new interaction in the next iteration. Otherwise, the interaction is *negative*, and the relationship ends.

After deciding on whether or not to continue the relationship, the preferences of agents are updated based on whether the interaction was positive or negative (see Section 3.5) and the turn of the agents ends. If the agents have decided to end the relationship, their *tolerance to failed relationships decreases*. Agents will exit the platform if this tolerance reaches 0, and otherwise they return to searching. If the agents decide to continue the relationship, they either continue the online relationship, or, if the relationship has lasted long enough, move to the Offline Relationship Phase.

3.3.3 Offline relationship phase. The turn of an agent in an Offline Relationship Phase progresses similarly to the Online one, with two key differences. First, if the interaction was positive, the agents either continue the relationship (as above); however, if the relationship lasts for 30 iterations, the agents then exit the platform into a *long-term* relationship. Second, while offline, in each iteration there

²We discuss initialization of preferences and norms in the supplementary material.

³We make the simplifying assumption that an agent can only be in one of the before-mentioned phases at one time, e.g. one cannot be both searching and dating offline.

is a fixed $1/7$ chance of interacting. This means that, in expectation, people in offline relationships meet once per week [10, 47].

3.4 Agent Decision-Making

In the Searching phase, agents answer the questions (posed to themselves), "should I reply to/send an initial message to this agent?". When in a relationship, the agent answers questions of the type "should I continue being in a relationship with this agent?". Both of these are yes/no questions, and agents use the same decision function, described below, to answer them.

Suppose that agent i must make a decision about sending a message or continuing a relationship with agent j . We assume agent i may know only an incomplete vector of agent j 's attributes, which we call $\hat{a}^{(j)}$. We use $\hat{a}^{(j)} \subseteq a$ if a extends the incomplete vector of attributes $\hat{a}^{(j)}$, i.e. a is a complete vector that has the same entries as $\hat{a}^{(j)}$ on the known attributes. Lastly, for the agents i, j there is some time value, t , capturing the amount of time spent in that specific phase with one another (for somebody in the searching phase or who just started an online or an offline relationship, $t = 0$).

In making decisions, agents consider a) the time spent so far with the other agent and b) how well the attributes of the other agent align with their own attributes and preferences. Time is an important predictor for breakups, as agents that have stayed longer in a relationship are less likely to interrupt their relationship. For example, people tend to be especially selective in their first offline date together [40]. Consequently, using the above introduced notation, the probability that agent i answers yes to a decision question regarding agent j is given by the following sigmoid function:

$$\mathbb{P}(\text{yes}) = \frac{1}{1 + \exp(-\mathbb{E}[w^{(i)} \cdot s^{(i)}(a) | \hat{a}^{(j)} \subseteq a] - t)}.$$

Two aspects of the above equation require further explanation. First, the expected value is taken over all possible attribute combinations extending the incomplete information of i over the attributes of j . The probability of each combination is given by the current stereotypes of i . Thus, i 's stereotypes impact what they *expect* the unknown attributes of j to be. Second, the function $s^{(i)}(a)$ is an attribute scoring function that for a given agent and complete attribute vector returns a vector with entries 1, 0, and -1 depending on whether each attribute is viewed as good, neutral, or bad. Agents see a matching attribute value as being *good* if it is the same as their own and *bad* if it is different. Similarly, they see a competing attribute value as *good* if it is larger than their own, *neutral* if it is equal, and *bad* if it is lower. The Supplementary Material provides a formalization of this scoring function.

3.5 Belief, Preferences, and Norm Updates

Agents' stereotypes, preferences, and the norms of the platform change on each iteration.

3.5.1 Agent Stereotype Updates. Recall that when generating an agent, we construct $G^{(i)}$, a sample of attribute combinations they have encountered so far. After each interaction, an agent observes a new combination of attributes. If that observation is a complete list of attributes, they add it to the sample. Otherwise, they find the probabilities of each possible complete combination of attributes that could extend the observed one and adds a proportional fraction

to each of these categories. For example, consider somebody that observes the combination $(1, 0, \perp)$ after an interaction, and their sample was 25 observations for $(1, 0, 0)$ and 75 for $(1, 0, 1)$. Post-interaction, their sample changes to 25.25 and 75.75.

3.5.2 Norm and Preference Updates. Preferences and on-platform norms both change during each iteration. First, agent preferences change according to their interactions based on a) whether the interaction was positive or negative and b) the attributes of agents. For example, a positive interaction with somebody of the same race will change the preference of the agent by increasing the preference of an agent to date someone of the same race. In contrast, a negative interaction with someone of the same race will make the agent *less* likely to seek out a same-race partner in the future.

At the end of each simulation turn, norms and preferences are updated to reflect their convergence [39]. Specifically, the on-platform norm gets closer to the average preference of agents on the platform, and the preferences of agents are updated by getting closer to the respective norm. Each update is carried out by taking convex combinations between the original value of the norms (preferences) and the values of the influencing preferences (norms). The extent to which norms determine preferences, and vice versa, is set via model parameters. See the supplementary information for more details on how these updates are made, as well as details on initialization.

4 VIRTUAL EXPERIMENT

We now describe how we use our model to explore three different classes of *interventions* outlined by Hutson et al. [34]. In Section 4.1, we operationalize these interventions, which are also listed in Table 2. We then discuss how we vary model parameters to evaluate the effectiveness of these interventions under different assumptions about the underlying society. Finally, we detail the outcome measures we use to understand the impacts of these interventions.

4.1 Interventions

Hutson et al. [34] suggest three places where platforms might intervene to reduce bias and discrimination on intimate platforms: 1) at the search, sort, and filter level, 2) at the matching algorithm level, and 3) at the community policy and messaging level. Here, we focus on the first and the last, and leave out the second for two reasons. First, not all platforms use matching algorithms, and even when they do, the algorithm differs substantially. Second, the literature on the possible algorithm to implement is vast [1, 8, 23, 33, 58, 63, 65]. Consequently, there are many different alternatives to substitute the existing matching algorithm with. A comparison between these various procedures is beyond the scope of this paper.

4.1.1 Filtering Interventions. The ease of filtering out people based on their attributes when searching for potential partners in online systems may exacerbate racial homogeneity in two ways: 1) it reduces the diversity in search results and 2) it makes race a legitimate basis for decisions [7, 34, 38]. The first suggested solution to mitigate this problem by Hutson et al. [34] is to purposefully introduce diversity into search results, e.g. by removing the ability to filter by certain attributes entirely. We refer to any intervention along these lines as a *filter intervention*. We consider four different filter interventions, comparing them to a baseline of traditional platform behaviors.

Table 2: The three types of Interventions, the baselines for each, and the different intervention conditions we test.

Intervention	Description
Filtering Interventions	
<i>Strong</i>	<i>Baseline. Only see others with the preferred attribute value on the most important attribute.</i>
Strong non-race	Strong filtering, if not on race. Otherwise, Strong filter on second most important.
Weak	Also shows profiles with unknown values on most important attribute
Weak non-race	Weak filtering, but not on Race.
Off	No filtering allowed.
Attribute Intervention	
5 attributes	Each agent has the protected attribute, plus one of each kind of attribute
9 attributes	Each agent has the protected attribute, plus two of each kind of attribute
Norm Intervention	
Off	<i>Baseline. On-platform norms are an average of agent preferences</i>
On	On-platform norms are fixed to 0 on race, encouraging agents not to consider it

As explained above, without any intervention, agents will filter their searches for romantic partners by the attribute they consider to be most important. We refer to this baseline option as *Strong filtering*. From this baseline, Hutson et al. [34] note that interventions to reduce racial homogamy can be constructed along two dimensions. First, we can restrict the possibility of filtering on race entirely. To signal such an intervention we add a "non-race" label to the filter name. Second, we can introduce diversity by showing profiles of agents who did not provide the given attribute in their profile. We label filtering strategies that only include the desired setting of the given attribute as "Strong," and those that also include potential partners who simply do not fill out a value for this attribute as "Weak". Combining interventions on these two dimensions gives three more filtering options, namely Strong non-race, Weak, and Weak non-race. We also add a fourth intervention where no filtering is allowed at all, on any attribute, i.e. the "Off" filtering option.

The five filtering conditions are most clearly stated with an example. Consider the agents in Figure 1. Let us assume the platform opts to use *Strong filtering*. Further, let us assume the blue agent's strongest preference is for a romantic partner that is the same race. In this case, the blue agent will search for others with the same race, and will see only other potential partners who publicly self-identified with that race on the platform. If the platform had *Weak filtering*, the agent would also see any other agents who a) have the same self-identified race or b) do not specify any race on the platform. If the platform employed *Strong non-race filtering*, the platform would not allow the agent to search by race at all. The agent would therefore use their second strongest preference (say, music), to perform their search. Weak non-race follows similarly, and in the Off condition, agents would simply not be able to filter at all. Finally, note that we have described the process with a matching attribute; for competing attributes, agents are always assumed to search for other agents who hold the desired attribute.

4.1.2 Attribute Intervention. Removing filtering options can make it harder for agents to find the partners they desire. A suggested solution by Hutson et al. [34] to mitigate this problem is to add or give additional attention to new non-race attributes on which agents can filter (e.g. political view or smoking preferences). We refer to this intervention as an *Attribute Intervention*. In our model, this intervention is implemented by increasing the number of dating-relevant attributes from a baseline of 5 to a total of 9, i.e. from having one non-racial attribute of each type, to having two of them.⁴

4.1.3 Norm Intervention. The final intervention from Hutson et al. [34] that we model are attempts to change or outline platform norms. Hutson et al. [34] suggest that different community policies, such as introducing pledges, educating users on the effect of expressing racial preferences, and detailing the disparities faced by racial minorities, can encourage social norms to date interracially. Norms, in turn, are known to be perhaps the most powerful mechanism for changing discriminatory behaviors [62]. In our experiment, the platform norm intervention fixes the value of the on-platform norm for the race attribute to 0. This means that agents, via the link between norms and agent preferences, are encouraged not to consider race at all in their decisions. Differently from the baseline condition, when on-platform norms are updated to reflect the average of preferences, the value on the race dimension does not change and continues to be 0 (see the Supplemental Material).

4.2 Assumptions about Society

Table 1 lists the model parameters we vary or keep constant in our virtual experiment. As we will show, consideration of these parameters is critical to understanding how well claims about intervention effects generalize to different assumptions about society. At the same time, a full exploration of the parameter space is infeasible. We take a three-pronged approach to addressing this. First, as noted in the table, many of these parameter and their values are theoretically grounded and/or empirically informed by prior work. Second, we engaged in sensitivity testing, described in the appendix, across a range of values for all parameters. Finally, for our virtual experiment, we fixed parameters that did not appear to have significant impacts on model outcomes. For parameters that did appear to have an effect, we select a default value for each (coloured in red in Table 1); when varying one such parameter, we keep the others at their default values.

As shown in Table 1, we vary five model parameters in our simulation. First, tolerance to unsuccessful search turns, which allows us to see how results change if agents are more or less tolerant to seeing undesired profiles and messages while browsing. Second, we vary racial bias in the out-platform norms, capturing how much importance society puts on same-race relationships compared to the other attributes of individuals. Third, we vary the influence norms have on preferences, i.e. the percentage of individual preferences explained by social norms. Fourth, we alter the initial degree of ethnocentrism, which intuitively corresponds to the extent to which agents are likely to prefer same-race relationships. The last

⁴As a reminder, attributes differ along two dimensions matching-competing and searchable-experiential, thus giving 4 possible types. Note also that Hutson et al. [34] additionally suggest a variant of this intervention, namely one based on the minimal group paradigm in social psychology [61]. We provide results for this similar intervention in the supplemental material.

parameters we change are the levels of correlation between attributes, i.e. β and γ ; we test in turn scenarios with zero, low, mild, and high correlations between race and the other attributes (β), and low and medium correlations between the other non-race attributes (γ). Each simulation runs over 2000 iterations, which corresponds roughly to a period of over 5 years. Moreover, we use 20 distinct random seeds for each scenario, thus having a total of 3000 runs.

4.3 Outcomes

We treat racial *heterogamy*—the opposite of racial homogeneity—as the main outcome variable when studying the impact of the simulated interventions on racial homogeneity.⁵ We operationalize racial heterogamy as the percentage of long-term relationships that are interracial. However, a proportion-based measurement can of course vary based on changes in the numerator or denominator. Thus, an increase in racial heterogamy might be caused by a decrease in efficiency, i.e. the overall number of long-term relationships, or an increase in the number of interracial relationships. We therefore investigate these two quantities separately as well.

Lastly, we provide two measures of user satisfaction, which are important for two reasons. First, profit-oriented decisions by platforms depend in part on the number and, hence, also on the happiness of users. This in turn suggests platforms may be less likely to implement interventions which threaten these metrics. Second, even if a platform was somehow persuaded to implement such interventions, the free nature of the market permits unsatisfied users to easily move to a different dating platform. To capture user satisfaction, we use two measures informed by literature on online dating [22]. First, we measure the percentage of time spent in an offline relationship, relative to online dating phases. Second, we measure the number of users who exit the platform because they have exhausted their tolerance for unsuccessful online searches.

5 RESULTS

5.1 Average Effects on Racial Heterogamy

Each intervention proposed by [34] increases racial heterogamy. However, as shown in Figure 3, the magnitude of these effects vary. The largest positive effect of any intervention is the Strong non-race filtering condition, i.e. when our simulated platform removes the option of filtering searches by race but allows agents to search on all other attributes. Doing so increases racial heterogamy from 10.5% to 18.3%. In contrast, a Weak filtering option increases the percentage by 3.4% over the baseline.

Combining the Strong non-race filtering intervention with the norm and attribute interventions further increase racial heterogamy. However, the effects are non-additive. For example, the effect of the norm intervention decreases after implementing a filtering intervention. Applying the norm intervention to the baseline filtering option results in a 4% increase in interracial relationships, while if it is applied on top of a Strong non-race intervention, the impact reduces to less than 3%. Finally, adding dating-relevant attributes produces an additional and relatively stable increase of 2% to 3%. These results suggest the potential of interventions on the platform, but also that their impacts may quickly reach a saturation point.

⁵We do so since we would like to emphasize the positive nature of the proposed change.

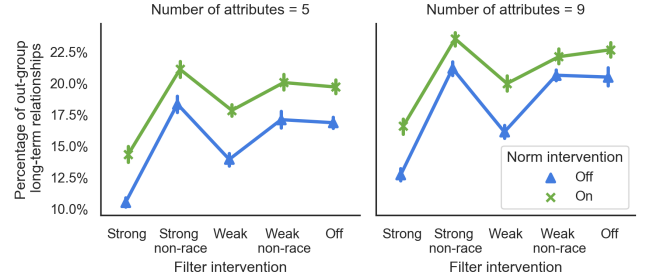


Figure 3: Racial heterogamy (y-axis) given a particular filter intervention (x-axis), and depending on whether or not the norm intervention is On (green lines, Xs as points) or Off (blue lines, triangles as points) and whether or not the attribute intervention is off (left subplot) or on (right subplot)

5.2 Effects on Total Number of Relationships

We also find that increases in racial heterogamy—the *percent* of long term interracial relationships—are not solely due to an increase in the *number* of interracial long-term relationships. Rather, as depicted in the first part of Figure 4, most interventions are also paired with a significant decrease in the overall number of long-term relationships.

With respect to the filtering interventions, removing the ability to filter by race causes an almost 20% drop in the total number of long-term relationships. This happens, in short, because agents, and people, often *want* to filter by race, whether they are conscious of that fact or not [2]. Thus, removing this filter makes it harder to find partners that agents desire. The attribute intervention causes this decrease for a separate, non-obvious reason. Specifically, including additional attributes makes the search space becomes more diverse, thus making it harder to find the “right” partner. Most obviously, with more attributes, the chance of two agents matching on all attributes decreases.

Finally, with respect to the norm intervention, influencing the preferences of agents by changing on-platform norms produces fluctuating evaluation criteria, especially when paired with racially biased offline norms. Agents enter the simulation with their own potentially racially-biased preferences. As they interact on the platform, their preferences slowly converge towards the racially unbiased on-platform norm. However, when going through the offline relationship phase, their preferences again adapt to the racially biased societal (offline) norms, thus shifting the evaluation criteria of agents again. These fluctuations means it takes longer to find suitable partners; so, if the process extends over the tolerance of agents for either bad recommendations or failed relationships, they will exit the platform without entering a long-term relationship. In our model, as in modern cultural sociological theory [49], then, the impact of the norm intervention is therefore contingent on the extent to which norms within the context of the platform permeate to other normative context by way of individual preferences.

5.3 Effects on User Satisfaction

Our simulations suggest that most interventions, on average, will decrease user satisfaction, especially filter interventions that do not

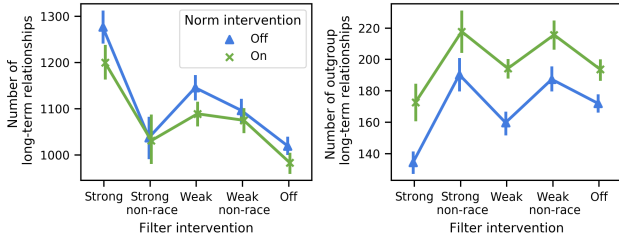


Figure 4: Number of total (left) and interracial (right) long-term relationships (y-axis) for the filter conditions (x-axis) and the norm intervention (Blue: off, Green: on). Results are only shown without the attribute intervention

allow users to filter on race. However, the magnitude, and in select cases, even the sign of these effects vary according to the settings of the other interventions, and across the two metrics considered. These findings present evidence that interventions can potentially be carried out without decreasing user satisfaction, but that these interventions typically are weaker (in terms of reducing racial homogamy), that effects on one satisfaction metric may contradict results on another, and that which combinations of interventions are most effective is not obvious a priori.

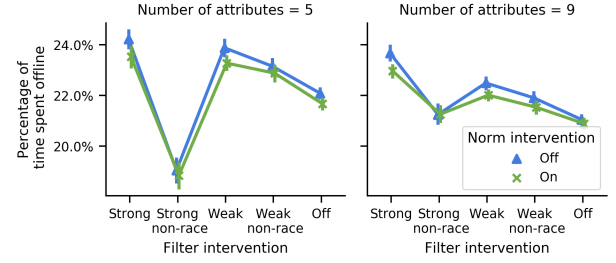
Specifically, Figure 5a shows that filtering interventions, to varying degrees, all decrease the percentage of agent’s time spent in offline relationships, and in doing so likely decrease user satisfaction. The most drastic impacts are from the Strong non-race intervention, where 19.0% of agents’ time is spent offline, as opposed to 24.2% in the baseline condition. This effect is, however, mitigated when paired with the attribute intervention. This is likely because agents become more satisfied with the non-race filtering options they have available to them. Other interventions, while less capable of reducing racial homogamy than the Strong non-race filter, have a much more limited effect on user satisfaction.

These observations for time spent offline also appear when measuring user satisfaction in term of the percentage of users who leave the platform without a partner. As shown in Figure 5, however, certain filter interventions can actually *increase* user satisfaction along this metric. Specifically, the use of Weak or no filtering can increase satisfaction on this metric, because agents often do not report all of their attributes, and thus including also the profiles with unknown attribute-values lowers the chance of excluding possibly interesting partners. Full descriptions of other patterns in Figures 5a and 5 are included in the Supplemental Material.

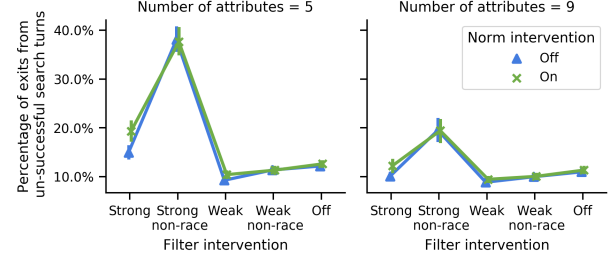
5.4 Intervention Impacts under Different Assumptions about Society

Finally, we explore how interventions fare under different assumptions about society. Broadly speaking, we find, as reflected in Figure 6, that interventions do have strong stand-alone beneficial effects, but that their impact is diluted if certain, often reasonable, assumptions about society are met.

The figure shows coefficients resulting from a linear regression with racial heterogamy as the outcome variable, and all (centered



(a) The percentage of time spent in offline relationships.



(b) The percentage of exits due to exhausting the tolerance for un-successful search turns.

Figure 5: Plots showing the level of user satisfaction with the application of different interventions.

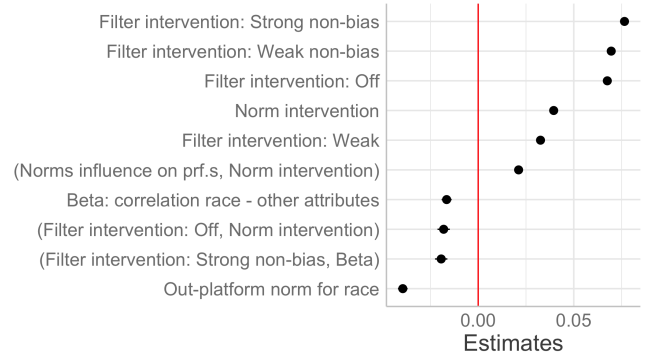


Figure 6: The first and second order marginal effects of the varied parameters on the percentage of out-group long-term relationships, for the 10 effects of strongest absolute value.

and scaled) model parameters in Table 1 and their pairwise interactions as independent variables. For readability, Figure 6 only includes the 10 coefficients with the strongest absolute effect. The regression model fits the data quite well ($R^2 = 0.83$), indicating that first- and second-order effects explain much of the dynamics of the complex system we model, but also that the model cannot be fully explained by phenomenon enumerable by theory alone.⁶

Filter interventions have the largest absolute effect on racial heterogamy. On average, across all simulated societies, removing the possibility to filter by race increases racial heterogamy by 7.65%

⁶See the Supplementary Material for full results of the first-order effects.

(CI = (7.43, 7.88))⁷. The Norm intervention also has a considerable positive impact, as it alone induces an average increase of 3.95% (CI = (3.73, 4.17)). The impact of these interventions reflects their potential to have real and possibly lasting changes on society.

Not all interventions we tested, however, increase racial homogeneity in the manner theorized by Hutson et al. [34]. Specifically, we find that only *matching* (as opposed to competing) and *searchable* (as opposed to experiential) attributes produce a meaningful increase in racial heterogeneity. Adding other types of attributes can actually have negative effects, either in decreasing overall numbers of relationships while not impacting the number of interracial relationships, or in simply leading to a decrease in racial heterogeneity.⁸ These findings underscore the importance of using simulation to think through the operationalization of potential interventions, especially by combining various theoretical models (here, of intervention and attribute types) into a single simulation model.

Other results in Figure 6 show, unsurprisingly, that different assumptions about society reflect drastically different baselines from which this improvement might come. For example, we see that racial heterogeneity is significantly *reduced* as correlations between race and the other attributes increase, and with an increase in racial bias in off-platform norms. Perhaps most importantly, however, Figure 6 reveals that intervention effects and societal conditions can *interact* to change the effectiveness of an intervention. For example, the Norm intervention is particularly effective when we make the assumption that agent preferences are strongly impacted by social norms. But, adding a non-race filtering option is less effective with increased values of the β parameter. Put another way, the effectiveness of not allowing individuals to search by race on an online platform is limited by the extent to which race is signaled by, or correlated with, other attributes that the platform *does* allow one to search on.

6 CONCLUSION

In their recent articles, Hutson et al. [34] and Anderson et al. [2] present contrasting views on the extent to which online dating platforms can intervene to impact long-term patterns of racial homogeneity in online dating. In the present work, we developed an agent-based modeling approach that synthesizes modern perspectives in (cultural) sociology, the psychology of online dating, and social computing to help us explore this debate.

Before discussing the implications of our work, we again explicitly call out two critical assumptions that our model makes. While we allow race to co-vary with other individual attributes, reflecting its multidimensional nature [57], our virtual experiment does ultimately assume that race, and gender, are both binary, i.e. that one can easily and completely specify an individual's race or gender. Doing so lends an aura of racial and gender determinism to our work that FAccT scholars are rightfully skeptical of [27]. Our experiment also assumes that these variables are independent, and thus ignores important intersectional identities. However, we emphasize that these assumptions are made only in the present virtual experiment, and are by no means hard-coded into our open

source simulation framework. We hope that we and others can therefore easily explore other constructions of race, dating, gender, and sexual preference in future work.

These limitations in mind, our findings shed critical light on the effectiveness and side effects of the interventions proposed by Hutson et al. [34] in reducing racial homogeneity, and ultimately racial inequality. Our work implies that the extent to which these interventions will be effective depends on four things. First, it depends on what we assume about society. We find that some of the proposed assumptions are effective only if we assume that people are unable to infer (perceived) race from other cultural preferences along which race can be performed online, an assumption that is potentially untenable given what is known about correlations between cultural attributes in general [15] and perceptions of race in particular [21]. And, other interventions are effective only if we assume that 1) the platform can significantly impact social norms towards interracial dating on their site and that 2) these norms have more significant effects on individuals than other “offline” normative structures (e.g. the family) that have historically encouraged racial homogeneity.

Second, it depends on what platforms, and society, are willing to accept as sacrifice for progress on racial homogeneity. Our model makes explicit these potentially non-obvious tradeoffs, quantifying how certain interventions to reduce racial homogeneity have negative impacts on user satisfaction and rates of overall long-term relationships. Third, it depends on how long one is willing to wait. We find that changing on platform norms have a limited effect; however, our model only studies short-term effects. Changes in large-scale social norms can take much longer, and thus the payoff of these normative interventions may exist only much further into the future as on-platform norms influence agent preferences, which in turn slowly reshape norms in other contexts. Finally, it depends on what one's ultimate goal is. If one's ultimate goal is to increase racial homogeneity to any degree, our results provide evidence that this is possible under a fairly wide range of assumptions. However, if one's goal is to eradicate racial homogeneity, then interventions at the platform level seem to hit a saturation point well below true equality. These technological interventions will therefore almost certainly never be enough; they must be supplemented with structural and normative social change beyond the platform.

While one could argue that these findings are post-hoc somewhat obvious, our work is motivated by a disagreement in the existing literature over the potential for platform intervention. This should serve as an important reminder that everything is obvious once it is observed [64], and we hope motivates the use of our approach to other domains. Especially when unsuitable changes could have large negative effects, constructing a model and simulating different possible outcomes could help explore potential problems to those interventions that emerge within a complex sociotechnical system.

ACKNOWLEDGMENTS

We thank Yuhao Du, David Garcia, Atri Rudra, Christoph Stadtfeld, Aleksandra Urman, and the FAccT reviewers for their valuable feedback on this work. KJ was supported by an NSF award IIS-1939579, with partial support from Amazon.

⁷Note that confidence intervals are provided here but should be cautiously interpreted, as they can be made infinitely small simply by running additional simulations

⁸See the Supplement for full results

REFERENCES

- [1] Atila Abdulkadiroğlu, Parag A Pathak, Alvin E Roth, and Tayfun Sönmez. 2005. The Boston public school match. *American Economic Review* 95, 2 (2005), 368–371.
- [2] Ashton Anderson, Sharad Goel, Gregory Huber, Neil Malhotra, and Duncan J Watts. 2014. Political ideology and racial preferences in online dating. *Sociological Science* 1 (2014), 28.
- [3] Delia Baldassarri and Peter Bearman. 2007. Dynamics of political polarization. *American sociological review* 72, 5 (2007), 784–811.
- [4] Amber Benham, Eli Edwards, Ben Fractenberg, Laura Gordon-Murnane, Cynthia Hetherington, Deborah A Liptak, Meg Smith, Craig Thompson, and Anne P Mintz. 2012. *Web of deceit: Misinformation and manipulation in the age of social media*. Information Today, Inc.
- [5] Eszter Bokányi and Anikó Hannák. 2020. Understanding inequalities in ride-hailing services through simulations. *Scientific reports* 10, 1 (2020), 1–11.
- [6] Dimitrios Bountouridis, Jaron Harambam, Mykola Makhortykh, Mónica Marrero, Nava Tintarev, and Claudia Hauff. 2019. SIREN: A Simulation Framework for Understanding the Effects of Recommender Systems in Online News Environments. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 150–159. <https://doi.org/10.1145/3287560.3287583>
- [7] Geoffrey C Bowker and Susan Leigh Star. 2000. *Sorting things out: Classification and its consequences*. MIT press.
- [8] Lukas Brozovsky and Vaclav Petricek. 2007. Recommender system for online dating service. *arXiv preprint cs/0703042* (2007).
- [9] Elizabeth E Bruch and MEJ Newman. 2018. Aspirational pursuit of mates in online dating markets. *Science Advances* 4, 8 (2018), eaap9815.
- [10] Robert J Brym and Rhonda L Lenton. 2001. Love online: A report on digital dating in Canada. *MSN. ca, February* 6 (2001).
- [11] John T. Cacioppo, Stephanie Cacioppo, Gian C. Gonzaga, Elizabeth L. Ogburn, and Tyler J. VanderWeele. 2013. Marital satisfaction and break-ups differ across on-line and off-line meeting venues. *Proceedings of the National Academy of Sciences* 110, 25 (2013), 10135–10140. <https://doi.org/10.1073/pnas.1222447110> arXiv:<https://www.pnas.org/content/110/25/10135.full.pdf>
- [12] George T Cantwell, Yanchen Liu, Benjamin F Maier, Alice C Schwarze, Carlos A Serván, Jordan Snyder, and Guillaume St-Onge. 2020. Thresholding normally distributed data creates complex networks. *Physical Review E* 101, 6 (2020), 062302.
- [13] Dongyu Chen, Xiaolin Li, and Fujun Lai. 2017. Gender discrimination in online peer-to-peer credit lending: evidence from a lending platform in China. *Electronic Commerce Research* 17, 4 (2017), 553–583.
- [14] Alexander D'Amour, Hansa Srinivasan, James Atwood, Pallavi Baljekar, D. Sculley, and Yoni Halpern. 2020. Fairness is not static: deeper understanding of long term fairness via simulation studies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 525–534. <https://doi.org/10.1145/3351095.3372878>
- [15] Daniel DellaPosta, Yongren Shi, and Michael Macy. 2015. Why Do Liberals Drink Lattes? *Amer. J. Sociology* 120, 5 (March 2015), 1473–1511. <https://doi.org/10.1086/681254>
- [16] Megan Rose Dickey. 2016. Here's Airbnb's plan to fix its racism and discrimination problem. *TechCrunch* (Oct. 2016). <https://social.techcrunch.com/2016/09/08/airbnb-plan-fix-racism-discrimination/>
- [17] Danielle Ensign, Sorelle A. Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian. 2017. Runaway Feedback Loops in Predictive Policing. *arXiv:1706.09847 [cs, stat]* (June 2017). <http://arxiv.org/abs/1706.09847> arXiv: 1706.09847.
- [18] Joshua M. Epstein. 1999. Agent-based computational models and generative social science. *Complexity* 4, 5 (1999), 41–60. <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.118.546&rep=rep1&type=pdf>
- [19] Eli J Finkel, Paul W Eastwick, Benjamin R Karney, Harry T Reis, and Susan Sprecher. 2012. Online dating: A critical analysis from the perspective of psychological science. *Psychological Science in the Public Interest* 13, 1 (2012), 3–66.
- [20] Ray Fisman and Michael Luca. 2016. Fixing discrimination in online marketplaces. (2016).
- [21] Jonathan B. Freeman and Nalini Ambady. 2011. A dynamic interactive theory of person construal. *Psychological review* 118, 2 (2011), 247.
- [22] Jeana H Frost, Zoe Chance, Michael I Norton, and Dan Ariely. 2008. People are experience goods: Improving online dating with virtual dates. *Journal of Interactive Marketing* 22, 1 (2008), 51–61.
- [23] David Gale and Lloyd S Shapley. 1962. College admissions and the stability of marriage. *The American Mathematical Monthly* 69, 1 (1962), 9–15.
- [24] Nigel Gilbert. 2007. *Agent-Based Models*. Sage Publications Inc, Thousand Oaks, CA.
- [25] N. Gilbert and K. Troitzsch. 2005. *Simulation for the social scientist*. Open university press. <http://books.google.com/books?hl=en&lr=&id=fBlauJpmNowC&oi=fnd&pg=PR1&dq=simulation+for+the+social+scientist&ots=PBj6JbJ3Wn&sig=rb6HhSIWO2HhHhOHs8IB3uaP9708>
- [26] Eric Grodsky and Devah Pager. 2001. The structure of disadvantage: Individual and occupational determinants of the black-white wage gap. *American Sociological Review* (2001), 542–567.
- [27] Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. 2020. Towards a critical race methodology in algorithmic fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 501–512.
- [28] Aniko Hannak, Piotr Sapiezynski, Arash Molavi Kakhki, Balachander Krishnamurthy, David Lazer, Alan Mislove, and Christo Wilson. 2013. Measuring personalization of web search. In *Proceedings of the 22nd international conference on World Wide Web*. 527–538.
- [29] Curtis D Hardin and E Tory Higgins. 1996. Shared reality: How social verification makes the subjective objective. (1996).
- [30] Albert H. Hastorf and Hadley Cantril. 1954. They saw a game; a case study. *The Journal of Abnormal and Social Psychology* 49, 1 (1954), 129. <http://psycnet.apa.org/journals/abn/49/1/129/>
- [31] Alex Hern. 2020. How hate speech campaigners found Facebook's weak spot. *The Guardian* (June 2020). <https://www.theguardian.com/technology/2020/jun/29/how-hate-speech-campaigners-found-facebooks-weak-spot>
- [32] Thomas Hills and Peter Todd. 2008. Population heterogeneity and individual differences in an assortative agent-based marriage and divorce model (MADAM) using search with relaxing expectations. *Journal of Artificial Societies and Social Simulation* 11, 4 (2008), 5.
- [33] Gunter J Hitsch, Ali Hortaçsu, and Dan Ariely. 2010. Matching and sorting in online dating. *American Economic Review* 100, 1 (2010), 130–63.
- [34] Jevan A Hutson, Jessie G Taft, Solon Barocas, and Karen Levy. 2018. Debiasing desire: Addressing bias & discrimination on intimate platforms. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–18.
- [35] Kenneth Joseph, Geoffrey P. Morgan, Michael K. Martin, and Kathleen M. Carley. 2014. On the Coevolution of Stereotype, Culture, and Social Relationships An Agent-Based Model. *Social Science Computer Review* 32, 3 (June 2014), 295–311. <https://doi.org/10.1177/0894439313511388>
- [36] Dean C Karnopp, Donald L Margolis, and Ronald C Rosenberg. 2012. *System dynamics: modeling, simulation, and control of mechatronic systems*. John Wiley & Sons.
- [37] Joachim Krueger. 1996. Personal beliefs and cultural stereotypes about racial characteristics. *Journal of personality and Social Psychology* 71, 3 (1996), 536.
- [38] Kevin Lewis. 2016. Preferences in the early stages of mate choice. *Social Forces* 95, 1 (2016), 283–320.
- [39] Omar Lizardo and Michael Strand. 2010. Skills, toolkits, contexts and institutions: Clarifying the relationship between different approaches to cognition in cultural sociology. *Poetics* 38, 2 (2010), 205–228.
- [40] Bridget L Long. 2010. *Scripts for online dating: A model and theory of online romantic relationship initiation*. Ph.D. Dissertation. Bowling Green State University.
- [41] Kristian Lum and William Isaac. 2016. To predict and serve? *Significance* 13, 5 (2016), 14–19. Publisher: Wiley Online Library.
- [42] Momin M Malik and Jürgen Pfeffer. 2016. Identifying Platform Effects in Social Media Data.. In *ICWSM*. 241–249.
- [43] Donald Martin Jr., Vinodkumar Prabakaran, Jill Kuhlberg, Andrew Smart, and William S. Isaac. 2020. Extending the Machine Learning Abstraction Boundary: A Complex Systems Approach to Incorporate Societal Context. *arXiv:2006.09663 [cs]* (June 2020). <http://arxiv.org/abs/2006.09663> arXiv: 2006.09663.
- [44] Merritt Roe Smith Leo Marx et al. 1994. *Does technology drive history?: The dilemma of technological determinism*. MIT Press.
- [45] Anna May, Johannes Wachs, and Anikó Hannák. 2019. Gender differences in participation and reward on Stack Overflow. *Empirical Software Engineering* 24, 4 (2019), 1997–2019.
- [46] M. McPherson, Lynn Smith-Lovin, and J. Cook. 2001. Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology* 1 (2001), 415–444. <https://doi.org/citeulike-article-id:3022278>
- [47] Marina J Muñoz-Rivas, Jose Luis Graña, D Daniel O'Leary, and M Pilar González. 2007. Aggression in adolescent dating relationships: Prevalence, justification, and health consequences. *Journal of Adolescent Health* 40, 4 (2007), 298–304.
- [48] Karl-Dieter Opp. 1982. The Evolutionary Emergence of Norms. *British Journal of Social Psychology* 21 (06 1982). <https://doi.org/10.1111/j.2044-8309.1982.tb00522.x>
- [49] Orlando Patterson. 2014. Making Sense of Culture. *Annual Review of Sociology* 0 (2014). <http://www.annualreviews.org/doi/abs/10.1146/annurev-soc-071913-043123>
- [50] Benjamin Pimentel. [n.d.]. Facebook was once a hot startup looking to conquer the world. Today, my journalist friends in the Philippines grapple with threats and lies on the social network. <https://www.businessinsider.com/facebook-dangerous-platform-philippine-journalists-duterte-2019-7>
- [51] Jason Radford and Kenneth Joseph. 2020. Theory In, Theory Out: The Uses of Social Theory in Machine Learning for Social Science. *Frontiers in Big Data* 3 (2020), 18. <https://doi.org/10.3389/fdata.2020.00018>
- [52] Artemio Ramirez, Erin M Sumner, Christina Fleuriot, and Megan Cole. 2015. When online dating partners meet offline: The effect of modality switching on relational communication between online daters. *Journal of Computer-Mediated Communication* 20, 1 (2015), 99–114.

- [53] Michael J Rosenfeld, Reuben J Thomas, and Sonia Hausen. 2019. Disintermediating your friends: How online dating in the United States displaces other ways of meeting. *Proceedings of the National Academy of Sciences* 116, 36 (2019), 17753–17758.
- [54] Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. 2014. Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and discrimination: converting critical concerns into productive inquiry* 22 (2014).
- [55] Thomas C. Schelling. 1971. Dynamic models of segregation. *The Journal of Mathematical Sociology* 1, 2 (1971), 143–186. <https://doi.org/10.1080/0022250X.1971.9989794>
- [56] Simon Schweighofer, Frank Schweitzer, and David Garcia. 2019. A Weighted Balance Model of Opinion Hyperpolarization. *56q12sds wx8daaggo3jb0ek2mxts6ndn* (2019).
- [57] Maya Sen and Omar Wasow. 2016. Race as a Bundle of Sticks: Designs that Estimate Effects of Seemingly Immutable Characteristics. *Annual Review of Political Science* 19 (2016), 499–522. <http://www.annualreviews.org/doi/abs/10.1146/annurev-polisci-032015-010015>
- [58] Lloyd Shapley and Herbert Scarf. 1974. On cores and indivisibility. *Journal of mathematical economics* 1, 1 (1974), 23–37.
- [59] Dan Siroker and Pete Koomen. 2013. *A/B testing: The most powerful way to turn clicks into customers*. John Wiley & Sons.
- [60] Paul E Smaldino and Jeffrey C Schank. 2012. Human mate choice is a complex system. *Complexity* 17, 5 (2012), 11–22.
- [61] Henri Tajfel, Michael G. Billig, Robert P. Bundy, and Claude Flament. 1971. Social categorization and intergroup behaviour. *European journal of social psychology* 1, 2 (1971), 149–178. <http://onlinelibrary.wiley.com/doi/10.1002/ejsp.2420010202/full>
- [62] Margaret E Tankard and Elizabeth Levy Paluck. 2016. Norm perception as a vehicle for social change. *Social Issues and Policy Review* 10, 1 (2016), 181–211.
- [63] Kun Tu, Bruno Ribeiro, David Jensen, Don Towsley, Benyuan Liu, Hua Jiang, and Xiaodong Wang. 2014. Online dating recommendations: matching markets and learning preferences. In *Proceedings of the 23rd international conference on world wide web*. 787–792.
- [64] Duncan J Watts. 2011. *Everything is obvious: * Once you know the answer*. Currency.
- [65] Peng Xia, Benyuan Liu, Yizhou Sun, and Cindy Chen. 2015. Reciprocal recommendation system for online dating. In *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 234–241.

An Agent-based Model to Evaluate Interventions on Online Dating Platforms to Decrease Racial Homogamy

Supplemental Material

Stefania Ionescu
ionescu@ifi.uzh.ch
University of Zürich
Zürich, Switzerland

Anikó Hannák
hannak@ifi.uzh.ch
University of Zürich
Zürich, Switzerland

Kenneth Joseph
kjoseph@buffalo.edu
University at Buffalo
Buffalo, NY, USA

ACM Reference Format:

Stefania Ionescu, Anikó Hannák, and Kenneth Joseph. 2021. An Agent-based Model to Evaluate Interventions on Online Dating Platforms to Decrease Racial Homogamy: Supplemental Material. In *Conference on Fairness, Accountability, and Transparency (FAccT '21)*, March 3–10, 2021, Virtual Event, Canada. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3442188.3445904>

A FURTHER DETAILS ON THE DECISION MODEL

Recall that the decision function is based on the following equation:

$$\mathbb{P}(\text{yes}) = \frac{1}{1 + \exp(-\mathbb{E}[\mathbf{w}^{(i)} \cdot \mathbf{s}^{(i)}(a) | \hat{a}^{(j)} \subseteq a] - t)}.$$

Formally, $\mathbb{E}[\mathbf{w}^{(i)} \cdot \mathbf{s}^{(i)}(a) | \hat{a}^{(j)} \subseteq a] = \sum_{a' \subseteq \hat{a}^{(j)}} (\mathbf{w}^{(i)} \cdot \mathbf{f}(a^{(i)}, a)) \cdot \mathbb{P}_{G^{(i)}}(A = a' | \hat{a}^{(j)} \subseteq A)$, where $\mathbb{P}_{G^{(i)}}(A = a' | \hat{a}^{(j)} \subseteq A)$ to be the probability an agent with the known attributes $\hat{a}^{(j)}$ has the vector of attributes a' in the view of agent i . This probability is obtained by dividing the entry for a' in $G^{(i)}$ over the sum of the entries corresponding to all possible extension of $\hat{a}^{(j)}$.

A.1 Attribute Scoring Function

In the main text, we informally introduced the attribute scoring function. Below, we include its formal definition:

$$\left(\mathbf{s}^{(i)}(a) \right)_k = \begin{cases} 1 & \text{,if } k \text{ matching and } a_k^{(i)} = a_k, \text{ or} \\ & k \text{ competing and } a_k^{(i)} < a_k \\ 0 & \text{,if } k \text{ competing and } a_k^{(i)} = a_k \\ -1 & \text{,if } k \text{ matching and } a_k^{(i)} \neq a_k, \text{ or} \\ & k \text{ competing and } a_k^{(i)} > a_k \end{cases}$$

A.2 Time Variable

Finally, note that the time contribution gets reset to 0 when moving from an online to an offline relationship. This is consistent with

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions.acm.org.
FAccT '21, March 3–10, 2021, Virtual Event, Canada

© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8309-7/21/03...\$15.00
<https://doi.org/10.1145/3442188.3445904>

the literature, as the first offline date is a crucial point in deciding whether or not the relationship will continue [4].

B FURTHER DETAILS ON INITIALIZATION

B.1 Agent attributes

To generate agent attributes during initialization of a new agent, we use a generative model common in both the decision-making literature [3] and to other areas of computational social science, e.g. topic modeling [1]. For each agent i , we start by generating one instance, $d^{(i)}$, of a $K + 1$ dimensional normally distributed random variable $A \sim \mathcal{N}(\mu, \Sigma)$. For interpretability, we use a zero-mean distribution, i.e. we fix $\mu = 0$. Therefore, $d^{(i)}$ will be the degree to which agent i fits with a 1 category on each attribute. For example, if $d_0^{(i)} = 0.5$ then i identifies more as having a race-value of 1, while if $d_0^{(i)} = -1$ they identify more as having a race-value of 0. Moreover, without loss of generality, we fix the variance of each component to 1, i.e. we let $\Sigma_{j,j} = 1$. To complete the values of the covariance matrix, Σ , we use two parameters: 1) a parameter β capturing the level of correlation between the protected attribute and the others, and 2) a parameter γ capturing the level of correlation between the non-protected attributes. That is, $\Sigma_{0,j} = \beta$ and $\Sigma_{j,t} = \gamma$ for every $j \neq t$ non-zero. Varying these two parameters changes the predictability of one attribute, given another. In our virtual experiment, both β and γ are changed to reflect various degrees of correlation (none, low, medium, and high) between the protected attributes and all other attributes (β), and between all non-protected attributes (low and medium) for γ - see Table ??). The binary attribute vector for each agent is obtained from the signs of the entries in $d^{(i)}$: negative values of $d_j^{(i)}$ correspond to $a_j^{(i)} = 0$, while non-negative value correspond to $a_j^{(i)} = 1$. Also generated from this distribution is the *contingency table*, G , which is used by agents to construct their *stereotypes*.

B.2 Stereotypes

Agents do not necessarily know the true contingency table G . Instead, based on their previous experience, they form a belief over this table, namely $G^{(i)}$ for agent i . We refer to this as the personal stereotypes of agents [9]. This matrix is generated together with a new agent, and, as mentioned in the main text, changes thought their time in the simulation. In this subsection we discuss its initialization.

Since this matrix is a result of the experience of the agent prior to entering the platform, we first create a sample of agents that

they encountered (either directly or indirectly) before entering the platform. Because racial homophily[11] and in-group favoritism[6, 14] influence the formation of stereotypes, this sampling is not race-balanced. We first generate, using the true contingency table G , 300 agents having the same race as agent i and 100 having a different race. Next, similar to previous work [8], we use the degree of ethnocentrism, e , and alter the true attributes of the different-race agents in the second sample. More precisely, for each agent of a different race in the sample, and each attribute that is perceived as good by agent i there is a chance e of changing it into a negatively-perceived attribute.

The (modified) sample of 400 agents induces a belief for agent i over which attribute combinations are more likely to be encountered. Formally, this belief is captured into the stereotype matrix $G^{(i)}$.

B.3 Norms and preferences

As mentioned in the main text, out-platform norms, agent preferences, and on-platform norms are initialized sequentially.

To start, the out-platform norm vector is generated inside the K th simplex. More precisely, since there is a restriction on the value on the race-component and the sum of searchable attributes versus the sum of experiential attributes, three parts are generated separately. The first part corresponds to the race attribute and is a $(K + 1)$ -dimensional vector with the first value given by the respective parameter and 0s on all the remaining positions. We refer to this as n_{race}^{off} . The second part corresponds to the searchable attributes. The sum on the components corresponding to the searchable attributes is given by accounting for the weight of the race attribute and for the relative importance of searchable vs experiential attributes, i.e. this part is given by the equation $\text{sum}_{searchable} = 1 - (\text{the norm value on the race attribute}) \times (\text{the importance of searchable attributes})$. Next we take an S -dimensional vector uniformly at random from the $(S - 1)$ -th simplex, complete it with $(K + 1 - S)$ values of 0 (one on the first position, and the remaining at the end), and multiply it by $\text{sum}_{searchable}$. This gives the searchable part of the out-platform norm, namely $n_{searchable}^{off}$. The last part, $n_{experiential}^{off}$, is obtained similarly as the searchable one. The out-platform norm is the sum of these three components, i.e. $n^{off} = n_{race}^{off} + n_{searchable}^{off} + n_{experiential}^{off}$.

Next, the individual preferences are obtained by taking the out-platform norm and adding some noise. Last, the on-platform norm vector is obtained by averaging the preferences on the searchable components. More precisely, for each agent i on the platform we consider $w^{(i,s)}$ to be the searchable entries of the preference vector of that agent. Since $w^{(i)} \in \Delta^K$, the sum of the searchable entries of that vector might be below 1, say $s_i = \sum_j w_j^{(i,s)}$. A re-scaling of $w^{(i,s)}$ will ensure that it is in the $(S - 1)$ th simplex, therefore $w^{(i,s)}$ becomes $\frac{1}{s_i} \cdot w^{(i,s)}$. We refer to this version of $w^{(i,s)}$ as the searchable vector of preferences. Their average is the initial on-platform norm, i.e. $n^{on} = \frac{1}{N} \cdot \sum_{i \leq N} w^{(i,s)}$.

When intervening on the on-platform norm, its value on the protected characteristic, i.e. its first entry, remains 0. To reflect this, after initially setting the on-platform norm as the average of

preferences, we let its first entry to be 0, and re-scale the others to keep the norm in the $(S - 1)$ th simplex. Similarly, after getting the norm closer to the average of preferences, we once again set its first entry to 0 and re-scale the other ones. With such an intervention in place, at the end of each iteration, when preferences are getting closer to norms, the agents' weight on the protected characteristic will lower.

C FURTHER DETAILS ON UPDATES

Norms and preferences

We now turn our attention to norm and preference updates, which, as mentioned in the main text, are carried out in sequence during each iteration. First, the preferences update after each interaction to reflect its outcome. Second, at the end of each iteration, the on-platform norm gets closer to the average preference of agents. Third, the preferences are updated by getting closer to norms. Each of these three updates are done by taking convex combinations between the original value of the norms/preferences and the values of the influencing vector. Since both preferences and norms are in a simplex, which is closed under convex combinations, they will remain in the simplex after the update. The remaining of this section explains the details of these updates.

First, the preferences of each agent in a relationship update based on the interaction. Using the decision function, agents decide whether or not to continue the relationship. We consider a mutual continuation decision as a sign of a *positive interaction*, and any other decision combination as a sign of a *negative interaction*. Let agent A update their preferences after an interaction of quality q with agent B . In this case, we construct an interaction vector i . The preference of A will then update to the convex combination between this interaction vector and the old preference vector.

We construct the interaction vector into stages. First, based on the quality of the interaction and how good each attribute is perceived, we determine whether the weight of each attribute should increase, decrease, or stagnate (see Table 1). For example, if there was a negative interaction with an agent having a good attribute, then perhaps that attribute is not so important, and its weight should decrease. Second, for each attribute k where the value should decrease, we put $i_k = 0$, and for each attribute k where the value should stagnate we put the previous weight value, i.e. $i_k = w_k^{(A)}$. Last, the interaction values for attributes that whose weight should increase are completed in proportion to their old-value, such that $i \in \Delta^K$. As a note, we only construct such a vector if there is some relative change, i.e. some attribute weight values should increase and some should decrease. If, for example, there is some positive interaction with somebody with all attributes perceived as being bad, then there will be no post-interaction preference update, as this indicates no relative change in preferences. Using this interaction vector, the preferences of agent A updates to a convex combination between i and the old preference, with a strength given by some parameter θ , i.e. $\theta \cdot i + (1 - \theta) \cdot w^{(A)}$.

Next, at the end of each turn, the on-platform norms are changed to get closer to the average of the searchable preferences of agents. As a reminder from Section ??, the searchable vector of preferences is obtained by taking from the weight vector of preferences only the searchable entries, and re-scaling it to be in the $(S - 1)$ th simplex.

	k good	k neutral/unkwnon	k bad
$q = \text{positive}$	↑	–	↓
$q = \text{negative}$	↓	–	↑

Table 1: Table showing whether the interaction vector should dictate an increase, decrease, or stagnation in each attribute k depending on the value of the interaction q .

The on-platform norms are updated to the convex combination between the average of agents' searchable vector of preferences and the old value of on-platform norms. Lastly, afterwards, each preference becomes the convex combination between its old value and the norm. Depending on whether the agent is online or offline, the norms in question could be the on-platform norm, which updates only the searchable attributes, or the out-platform norm.

D FURTHER DETAILS ON FIXED PARAMETER VALUES

As outlined in the table showcasing the parameters in our model, some parameters remain constant thought the experiment. In some cases, their values are chosen based on past work, while for others we did a sensitivity test to see if they have a meaningful change. Below, we discuss in turn some of the reasoning that went behind fixing each of these parameters:

- *Strength of updates: interaction \rightarrow preferences.*
In the initial experiment we also tried more extreme values, such as a 50% and an 100% influence. Increasing this parameter also increases the unpredictability of the results: the variance and the time taken to achieve convergent behaviour surge. Therefore, we settled for more reasonable values on this parameter.
- *Strength of updates: preferences \rightarrow norms.*
For this parameter, we also tested larger values, such as 10%, 20% and 50%. Using such values did not affect the results, especially when paired with low level of influences of interactions on preferences, i.e. when preferences were more stable.
- *Relative importance of out-platform norms when generating initial preferences.* This parameter controls how much alike the preferences of agents are originally, and how much they resemble the out-platform norms. Variations of these parameter influenced more the short-term results, i.e. the ones after only a couple of hundred of iterations, rather than the end-results observed after 2000 iterations
- *Number of initial acquaintances in-group and out-group.*
We have also tried having less such acquaintances, namely 150 and 50. Besides a slight increase in variability no difference was observed.
- *Weight of searchable versus experiential attributes.*
Frost et al. [5] conducted a study on how much weight do searchable attributes have when compared to experiential attributes when considering a potential partner. We choose the values of this parameter based on their result.
- *The probability a searchable attribute is specified on the profile and the probability of learning an attribute of someone else.*

Past work indicates that users do not always specify all the attributes they could on their dating profile[10] and that attributes are discovered during interactions [4]. These motivates the using of these two parameters in our model. The exact numbers influence only the amount of time the two agents act on incomplete information. Small variations to these numbers do not qualitatively change our results.

- *The number of rounds until a relationship becomes offline and long-term and the frequency of offline interactions.* The time it takes one to move to the next step of their relationship as well as how often they interact varies a lot across people. Therefore, we used values close to the averages reported in previous work to set these parameters [2, 12, 13].
- *The relative time spent looking at received messages versus new profiles during search and number of profiles and messages considered per iteration.* During their time searching, users can both look at profiles and consider messages [4]. Based on previous work, we estimated that, on average, one considers around 30 potential partners (either for sending an initial message, or for giving a reply to such a message and start an online relationship) [4, 5].
- *Time-span for observing the platform.* The simulation runs for 2000 iterations, the rough equivalent of 5 years. In our sensitivity test we also tracked 5000 iterations, but the results revealed that 2000 are enough to fully observe the effects of interventions, i.e. achieve convergence and small variance.
- *Initial population size and the number of agents added per iterations.* The sensitivity test also experimented with using less (150 initially and 2 added per iteration). This decrease did not produce qualitative changes. Due to computational limitation we did not further increase the number of agents.
- *Types of agents.* In the main text, agents can be of two types, loosely mimicking a heteronormative online dating website. However, we also tested a platform with only one type of agent, thus moving from a two-sided matching problem to a one-sided matching one. The results remind qualitatively similar.

Further analysing the impact of changing these parameter could potentially reveal other interesting phenomena. For example, we know that users usually meet within 1 – 2 weeks, but what would happen if an intervention was made such that they are encouraged to meet sooner? That being said, this is outside the scope of our paper and is thus left for future research. Even though such an analysis could reveal new dependencies between parameters as well as potentially new effective interventions, they should not change the causal implications observed in this paper.

E ADDITIONAL RESULTS AND RESULT DETAILS

E.1 Explaining Patterns in User Satisfaction Metrics

Norm interventions produce smaller and often insignificant drops in user satisfaction, as they change the underlying preference of users, so also the way they perceive potential partners. The slight drop, of only 1.06% on average, in offline-time when intervening on

on-platform norms is due to the persistence of off-platform norms: even though users change their preference while on the platform because of the intervention there, they will still go through an offline state, when their preferences are altered to reflect these different norms. When increasing the number of attributes, the negative effects of adding non-race filtering options decrease. For example, a shift to Strong non-race changes the percentage of time spent offline from 23.6% to 21.2%. The reason for this is that removing the possibility to search by one attribute affects users less if they have multiple others to search by; with more attributes there is a higher chance of having a second attribute that is almost as important as the first.

Slightly different results can be noted with respect to the dissatisfaction resulting from un-successful search turns (see Figure 5b in the main text). While the addition of non-race filtering options on top of both strong and weak filtering increase the level of un-satisfaction, the adding of variety with weak filtering and sometimes even with no filtering does not necessarily do so. Since agents often do not report all of their attributes, including also the profiles with unknown attribute-values lowers the chance of excluding possibly interesting partners. Consequently, when around 50% of users do not report their value on the search attribute, a strong filtering option excludes many potentially suitable candidates, thus explaining the gaps. Similarly to our other measures of satisfaction, disregarding of the filtering or norm intervention, the addition of attributes improves the satisfaction. When agents have more attributes to search on, excluding one of these or indirectly influencing the preferences through norms disturbs less the capabilities of users to signal and search based on this desires.

E.2 Regression Model with First-Order Effects

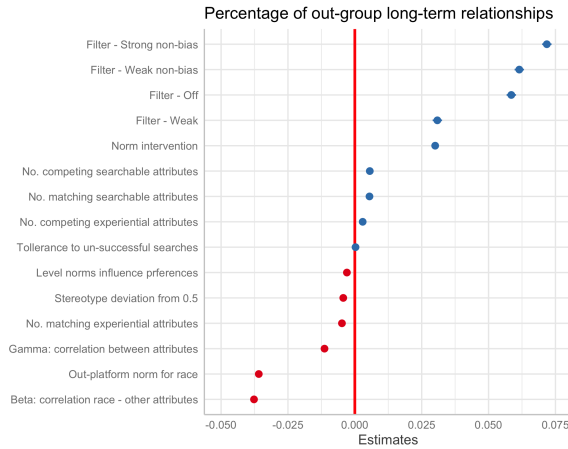


Figure 1: The marginal effect of each varied parameters on the percentage of out-group long-term relationships.

In the main text, we used a linear regression with racial heterogamy as the outcome variable and the varied model parameters and their pairwise interactions as independent variables to analyse the impact of parameters and pair of parameters on the percentage of out-group long-term relationships. That model had a high

predictability ($R^2 = 0.83$). However, the first-order terms alone also fit data quite well ($R^2 = 0.79$). In this part of the supplementary material we look at all the first order effects that were not already discussed in the main text. Figure 1 provides an overview of these effects.

Before, we have seen that racial heterogamy is significantly *reduced* as correlations between race and the other attributes increases, i.e. β , and with an increase in racial bias in off-platform norms. Figure 1 shows that the correlation between non-race attributes also negatively impacts heterogamy.

In contrast, as expected, there are also societal structures that can *increase* racial heterogamy. For example, adding more attributes usually improves the percentage of out-group relationships, with the only exception being the addition of matching experiential attributes. To understand why this is the case, one needs to remember that the effect on the percentage is a mix between the effect on the total number of long-term relationships and the effect on the number of out-group ones. Adding competing attributes decreases the total number of relationships and has no effect on the number of out-group relationships. Consequently, adding them increase the percentage. Conversely, adding matching attributes has no effect on the total number of relationships, so the effect on the percentage is given by the change in the number of out-group ones. Increasing the number of matching attributes that are observable while online result in more out-group relationships, while adding matching attributes that are only observable when offline produce a decrease. In conclusion, even though adding attributes generally increases the percentage of out-group relationships, the only attributes that produce a meaningful improvement are the matching searchable ones, as their effect is due to the increase in the total number of out-group long-term relationships rather than due to a decrease in the overall number of relationships.

The level of stereotyping is another interesting parameter, as values away from 50%, in either direction have a negative effect on the percentage of out-group relationships. Putting this another way, stereotyping that agents of a different race have half of their "good" attributes "bad" have worse effect than both perceiving the different race agent as they are and believing that all of their attributes are "bad" (where "good" and "bad" matching attributes are in relationship with the agent considering them). To give an example, the blue agent in Figure ??, has a higher chance of finding a partner within a different bias group if either they believe that people with a race of 1 have exactly the true distribution over music liking (0% stereotyping) or that they never like music (100% stereotyping) than if they think half of the ones that do actually like music and have a different race do not have a passion for it (50% stereotyping). To understand why this is the case, let us consider the case when the blue agent observes another agent with a music passion and an un-specified race. If the blue agent has a 100% level of stereotyping they will assume that the agent has the same race and possibly also associate other positive attributes with them. When they learn the true race value of the agent they will also have learned more of the true attributes of the agent, so by that time, race might not be a determining factor in decision making. On the other hand, if the blue agent has a 50% level of stereotyping, the blue agent still perceives them being of a different race as a possibility, which might result in projecting the perceived negative

attributes on the agent in question, consequently resulting in a lower probability of forming and continuing a relationship.

Two smaller effects are due to the level at which norms influence preferences and the tolerance level of agents to un-successful search turns. Although their marginal effect on the percentage of out-group relationships is small, they affect the user satisfaction. For example, an increased level of influence of the norms on preferences results in higher dissatisfaction when implementing norm interventions due to an increase in the drop rate after the first relationships, while a lowered bad recommendation tolerance decreases both the time spent offline and the percentage of exits due to too many un-successful search turns.

E.3 Minimal Group Paradigm Intervention

As a variant of the attribute intervention, Hutson et al. [7] also suggest creating attributes that help users "categorize potential partners along new axes unassociated with protected characteristics". This intervention is based on the minimal group paradigm in social psychology [14]. To implement it, we added a new attribute searchable and matching attribute that is correlated neither with race nor with any other preexisting attributes. We also made the assumption that agents value this newly introduced attribute in the same way they value the others. That is, when generating and updating preferences and norms we treat these attribute as any of the other matching searchable non-bias attribute. Figure 2 shows the impact of this intervention on racial heterogamy as well as on the total and out-group number of long-term relationships.

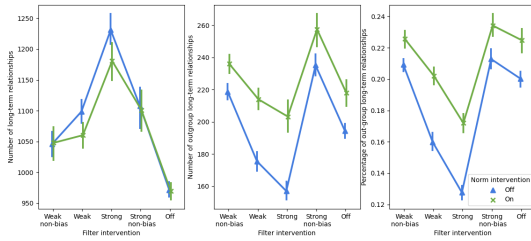


Figure 2: The percentage and number of out-group long-term relationships after the minimal group paradigm intervention.

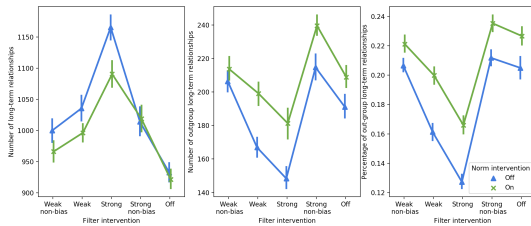


Figure 3: The percentage and number of out-group long-term relationships after the attribute intervention.

Similarly to the attribute number intervention, this intervention produces an increase in the percentage of out-group long-term relationships over the baseline (see main text for this figure). Moreover, under the assumption mentioned before, i.e. that creating this new attribute can be done in such a way that users value it as much as the other attributes, this intervention is more effective than the general attribute intervention (Figure 3). More precisely, the increase in the percentage of out-group long-term relationships is the same, but the decrease in the total number of relationships is smaller and the increase in the number of out-group ones is larger. The analysis of the first order effects in the section above explains this difference. As observed there, matching searchable attributes are the only type of attributes with meaningfully effects, i.e. an improvement in heterogamy due largely to an increase in the number of out-group long-term relationship rather than to a decrease in the total number of relationships. Thus, one would expect the minimal group paradigm intervention, which only adds matching searchable attributes, to have more positive effects than the general one, which adds one attribute of each.

REFERENCES

- [1] David M Blei, John D Lafferty, et al. 2007. A correlated topic model of science. *The Annals of Applied Statistics* 1, 1 (2007), 17–35.
- [2] Robert J Brym and Rhonda L Lenton. 2001. Love online: A report on digital dating in Canada. *MSN. ca, February* 6 (2001).
- [3] George T Cantwell, Yanchen Liu, Benjamin F Maier, Alice C Schwarze, Carlos A Serván, Jordan Snyder, and Guillaume St-Onge. 2020. Thresholding normally distributed data creates complex networks. *Physical Review E* 101, 6 (2020), 062302.
- [4] Eli J Finkel, Paul W Eastwick, Benjamin R Karney, Harry T Reis, and Susan Sprecher. 2012. Online dating: A critical analysis from the perspective of psychological science. *Psychological Science in the Public Interest* 13, 1 (2012), 3–66.
- [5] Jeana H Frost, Zoe Chance, Michael I Norton, and Dan Ariely. 2008. People are experience goods: Improving online dating with virtual dates. *Journal of Interactive Marketing* 22, 1 (2008), 51–61.
- [6] Albert H. Hastorf and Hadley Cantril. 1954. They saw a game; a case study. *The Journal of Abnormal and Social Psychology* 49, 1 (1954), 129. <http://psycnet.apa.org/journals/abn/49/1/129/>
- [7] Jevan A Hutson, Jessie G Taft, Solon Barocas, and Karen Levy. 2018. Debiasing desire: Addressing bias & discrimination on intimate platforms. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–18.
- [8] Kenneth Joseph, Geoffrey P. Morgan, Michael K. Martin, and Kathleen M. Carley. 2014. On the Coevolution of Stereotype, Culture, and Social Relationships An Agent-Based Model. *Social Science Computer Review* 32, 3 (June 2014), 295–311. <https://doi.org/10.1177/0894439313511388>
- [9] Joachim Krueger. 1996. Personal beliefs and cultural stereotypes about racial characteristics. *Journal of personality and Social Psychology* 71, 3 (1996), 536.
- [10] Kevin Lewis. 2016. Preferences in the early stages of mate choice. *Social Forces* 95, 1 (2016), 283–320.
- [11] M. McPherson, Lynn Smith-Lovin, and J. Cook. 2001. Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology* 1 (2001), 415–444. <https://doi.org/citeulike-article-id:3022278>
- [12] Marina J Muñoz-Rivas, Jose Luis Graña, K Daniel O’Leary, and M Pilar González. 2007. Aggression in adolescent dating relationships: Prevalence, justification, and health consequences. *Journal of Adolescent Health* 40, 4 (2007), 298–304.
- [13] Artemio Ramirez, Erin M Sumner, Christina Fleuriet, and Megan Cole. 2015. When online dating partners meet offline: The effect of modality switching on relational communication between online daters. *Journal of Computer-Mediated Communication* 20, 1 (2015), 99–114.
- [14] Henri Tajfel, Michael G. Billig, Robert P. Bundy, and Claude Flament. 1971. Social categorization and intergroup behaviour. *European journal of social psychology* 1, 2 (1971), 149–178. <http://onlinelibrary.wiley.com/doi/10.1002/ejsp.2420010202/full>