Joseph *et al. Applied Network Science* *(2022) 7:49*
https://doi.org/10.1007/s41109-022-00486-8

Applied Network Science

RESEARCH Open Access

A qualitative, network-centric method for modeling socio-technical systems, with applications to evaluating interventions on social media platforms to increase social equality

Kenneth Joseph[1*], Huei-Yen Winnie Chen[2], Stefania Ionescu[3], Yuhao Du[1], Pranav Sankhe[1], Aniko Hannak[3] and Atri Rudra[1]

*Correspondence: josephkena@gmail.com

[1] Computer Science and Engineering, University at Buffalo, Buffalo, NY, USA
[2] Industrial and Systems Engineering, University at Buffalo, Buffalo, NY, USA
[3] Social Computing Group, University of Zurich, Zurich, Switzerland

Abstract

We propose and extend a qualitative, complex systems methodology from cognitive engineering, known as the *abstraction hierarchy*, to model how potential interventions that could be carried out by social media platforms might impact social equality. Social media platforms have come under considerable ire for their role in perpetuating social inequality. However, there is also significant evidence that platforms can play a role in *reducing* social inequality, e.g. through the promotion of social movements. Platforms' role in producing or reducing social inequality is, moreover, not static; platforms can and often do take actions targeted at positive change. How can we develop tools to help us determine whether or not a potential platform change might actually work to increase social equality? Here, we present the abstraction hierarchy as a tool to help answer this question. Our primary contributions are two-fold. First, methodologically, we extend existing research on the abstraction hierarchy in cognitive engineering with principles from Network Science. Second, substantively, we illustrate the utility of this approach by using it to assess the potential effectiveness of a set of interventions, proposed in prior work, for how online dating websites can help mitigate social inequality.

Keywords: Abstraction hierarchy, Social media and social inequality, Online dating, Qualitative network analysis

Introduction

Recent Congressional inquiries into companies like Facebook and Twitter underscore the various ways that social media platforms are negatively impacting society (Bond 2021). From increased harassment of female journalists (Chen et al. 2020), to the (algorithmic) amplification of majority opinions (Espín-Noboa et al. 2021) and male content (Wachs et al. 2017), there is growing evidence that social media platforms have played a causal role in the creation and reinforcement of social inequality. As with many technological innovations, however, social media is not universally beneficial or detrimental. For example, there is also significant evidence that social media provides an important platform for budding social movements to push forward towards a more equal and just world (Jackson et al. 2020).

Social media platforms that at one point in time serve to exacerbate inequality also are not destined to do so forever. Instead, the role that a particular platform plays can change over time as it seeks to address existent issues. Twitter, for example, has long served as a medium for hateful speech (Davidson et al. 2017), but has recently implemented a number of reforms, including significant efforts to curb hate speech on the platform.[1]

A core question of interest to researchers of social media systems today is whether or not these changes made by the platform will actually work to increase social equality. A number of methods are available to address such "What-if?" questions. Platforms in particular have the luxury of conducting empirical research, implementing changes they are willing to make and then evaluating the effects of those interventions. For example, recent work from researchers at Twitter use internal data and privileged access to system internals to study potential partisan biases of the company's recommender systems (Huszár et al. 2022). Those outside the platform's boundaries can engage in similar empirical studies, but only within the confines allowed by the platform's current state. For instance, scholars have used ingenious pseudo-experimental designs (Malik et al. 2015) as well as algorithmic auditing approaches (Hannak et al. 2014) to interrogate the ways in which platform affordances and/or algorithms may lead to social inequality.

However, even the platform's own empirical studies are bound by the fact that what we can measure is not necessarily a direct cause of social inequality and/or marginalization. For example, increasing inter-racial contact online does not necessarily serve to address the core economic and resource disparities at the heart of racial inequality (Ionescu et al. 2021). Mapping from platform behaviors to offline inequality instead requires us to engage in extremely difficult data collections to link the online and offline worlds (Radford and Joseph 2020). Further, even the factors that we *can* measure are often only imperfect proxies of the true underlying causal quantities of interest, and thus are limited in what they can say about casual relationships of interest in an intervention setting (Wang and Blei 2021). This problem is often compounded by the simplifications required in order to study complex systems with computationally tractable algorithms (Baumer and Silberman 2011)..

In the face of these challenges, scholars have devised various ways of simulating potential interventions (Ionescu et al. 2021; Martin Jr et al. 2020; Schweitzer and Garcia 2010). Simulations are bound only by the imagination of the modeler, and thus data availability is of little concern (Epstein 1999). But simulations still require the ability to parameterize potential causes and effects of social inequality. And quantifying these causes and effects in a simulation model is difficult when the causal processes we know to exist occur at different levels of analysis. For example, in understanding what might cause Twitter to ramp up their efforts to combat hateful speech, it is important to model macro level

[1] https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy.

factors like the high-level financial motivations of platforms (Van Dijck 2013). However, it is at the same time important to model factors at the micro-level; something as minute as how profiles are displayed can change social interactions (Jaidka et al. 2021) in ways that impact willingness to spew hateful content (Munger 2017).

Here, again, scholars have made significant headway, e.g. in the modeling and analysis of multilevel networks and other hierarchical and higher dimensional representations of complex systems (Torres et al. 2021). Such models permit simulations (and quantitative network studies) to perform analysis at multiple levels of analysis. However, while incorporating multiple levels of analysis into a single simulation model is thus possible, it unlikely to produce interpretable results because it would require a vast arrangement of hyperparameters at the discretion of the researcher (Galán et al. 2009). For example, existing simulation models of social media often model only micro-level interactions, and still have dozens of parameters (Ionescu et al. 2021; Bokányi and Hannák 2020). Space therefore exists for new methods that help us to characterize these kinds of relationships without the need to immediately resort to overly complex simulation models, or to empirical methods that are bound by platform decisions and proxy measures.

In the present work, we propose and extend the *Abstraction Hierarchy (AH)* modeling framework as a tool that fills this existing methodological gap. Originating from the field of Cognitive Engineering (Wilson 2014), the AH allows us to use a formal qualitative approach to specify assumptions about the indirect, hierarchical interrelationships between components of complex sociotechnical systems (Read et al. 2015). From this qualitative specification, which seeks to provide a high-level model of the *entire* system of interest, we can then dig deeper with quantitative methods knowing what variables we can and cannot measure, and how these variables interrelate across different levels of analysis.

The first contribution of the present work is thus to bring the AH modeling framework into the study of social media platforms and social inequality.[2] While a long literature exists developing the AH in the cognitive engineering literature, including recent work to understand the relationship between online communities and their platforms (Euerby and Burns 2012), we are unaware of any work that has ported the model into current debates over the why's and how's of social media's impacts on society.

The AH modeling framework defines a procedure by which scholars can qualitatively construct a hierarchical model of a socio-technical system. More specifically, scholars in the field of cognitive engineering have for several decades been developing the AH as a means to determine how best to provide a representation of a work domain that will help human operators to navigate complex human-machine systems (Vicente and Rasmussen 1992; Salmon et al. 2019; Wurst et al. 2021). A research team can use literature reviews, interviews with experts, and/or their own knowledge to construct this model. A completed AH has a set of five hierarchical levels, each of which contains one or more nodes that define the system at different levels of intent and granularity (Bisantz and Vicente 1994). For example, the top level, called the

[2] The abstraction hierarchy modeling framework pre-defines a *process* by which an Abstraction Hierarchy *model* can be constructed. Unless explicitly stated otherwise, our use of the phrase abstraction hierarchy in this article refers to a model, not the process of constructing one.

Functional Purpose level, defines the high-level goals of the entities modeled within the system. A functional purpose of most social media platforms that can be modeled in an AH, for example, is profit maximization, and thus "Maximize Profit" might be a node in the functional purpose level of our AH. In contrast, the bottom level, entitled the *Physical Forms* level, encompasses the physical artifacts that can be assessed or measured to characterize how the system is operating. For example, retweets on Twitter might serve as physical forms, which can be measured to, e.g., assess gender-biased patterns of social attention. "Retweets" might therefore be a node as well.

Critically, nodes in an AH can be connected to nodes in the level above or the level below. At a high level, these links represent how/why relationships between nodes. For example, as detailed further below, the second level of the AH represents constraints on functional purposes. A "Maximize Profit" functional purpose node might therefore be connected to a "User demands" node at the second level, as user interests are sometimes a constraint on the profits of platforms. If we ask *why* a platform is not appropriately maximizing profit, we may look to "User demands" as a cause. Collectively, an AH is therefore a directed graph (constrained by the imposed hierarchy). The graph represented by a completed abstraction hierarchy in principle addresses the two challenges specified above: it allows us to 1) explicitly link analyses and/or conceptual models at different levels of analysis/abstraction through how/why relationships, and 2) to specify and trace pathways from components of the platform to outcomes of interest.

In cognitive engineering, these analyses of Abstraction Hierarchies largely consist of qualitative evaluation of the structure as a whole. Because the AH modeling framework produces a network, however, we can also use tools and concepts from network science to shed additional light on the graph structure. A second contribution of the present work is therefore to articulate how specific tools from network analysis can be brought to bear on models created from the AH. Specifically, we show how even the simplest and most common network analyses—clustering using the long-established Newman-Girvan algorithm (Girvan and Newman 2002), and counting the paths between pairs of nodes—can provide additional insight into what a completed AH model can tell us.

Finally, to illustrate the utility of the abstraction hierarchy in the context of social media platforms and social inequality, we present a case study. The focus of our case study is on the role that online dating platforms have to play in creating and/or reducing racial income inequality and racial segregation. The case study we select extends recent simulation (Ionescu et al. 2021) and quantitative empirical (Anderson et al. 2014) work that speak to how proposed interventions (Hutson et al. 2018) on online dating platforms are (or are not) likely to have a downstream impact on these two tightly correlated forms of inequality.

In comparison to the methods used in prior work, we show that the AH, in combination with concepts from network science, provides a unique ability to succinctly capture an interdisciplinary perspective that spans multiple levels of analysis on the potential constraints on and pathways to platform interventions successfully decreasing social inequality.

Background

Abstraction hierarchy

The abstraction hierarchy, as part of the cognitive work analysis framework (Vicente 1999), originated from efforts by cognitive engineering scholars to understand the complex links between concrete design variables and the purposes they serve, intentionally or not. Much of the early work focused on physical domains with more structure, but arguably higher risk, than those considered here, e.g. in control systems for nuclear power plants (Bisantz and Vicente 1994). However, the abstraction hierarchy has proven to be a very flexible tool. Researchers have adopted it for work domains that are much less well-defined and increasingly intentional, such as naval commands and control (Burns et al. 2005), ambulance dispatch management (Wong et al. 1998), financial systems (Achonu and Jamieson 2003) and most relevant to our work here, community practices in online social networks (Euerby and Burns 2012). In contrast to other types of hierarchies that are more process or activity oriented, the abstraction hierarchy framework is unique in the cognitive engineering space in that it is goal-oriented. Each predefined level of the AH describes the same system but using a different language (e.g., purposes vs. processes involved). Connecting the levels are the means-end links where moving up illustrates the significance of the element as it relates to the goals and purposes of the system, and moving down provides a more detailed explanation of how such goals are achieved. This explicitly goal-oriented nature makes it particularly suitable for us to study how specific mechanisms in online dating websites can play a role in mitigating social inequality. In addition, the pre-defined hierarchical levels into which nodes must be situated provide a good match for the needs of our problem setting.

In the present work, we extend analyses of Abstraction Hierarchies by using simple network analytic techniques to analyze the final model. This extension of the abstraction hierarchy parallels a long line of research that has combined qualitative analysis of text corpora to produce networked representations of the frames, or "maps," inherent to the data (Carley 1993) (Xi et al. 2021; Christensen and Kenett 2021). These maps can then be used in network analyses that, for example, identify themes (Matthes and Kohring 2008). The present work parallels this approach of extracting a network model using qualitative methods and then applying standard ideas from network science to better understand this network. The primary differences are, of course, the data and questions of interest, and the qualitative method applied—here, we use the abstraction hierarchy, rather than a content or mapping analysis.

Other related methodological approaches

The network structure of an abstraction hierarchy also can be understood to underlie the other methods common in studying results of interventions from social media platforms, namely, quantitative empirical analysis and simulation. With respect to the former, the assumptions that underlie all statistical models can, via the logic of probabilistic graphical models (Koller and Friedman 2009), be translated into a network graph. Such graphs can then be analyzed to understand the causal assumptions made and/or discovered by that statistical model. Scholars have in recent years turned towards a subset of probabilistic graphical models, called causal graphical models (Pearl 1998), to reason about the dynamics of complex social systems, including the study of how bias and discrimination emerge (Schröder et al. 2017; Stauffer and Solomon 2007). Like the AH, causal graphical models encode cause-effect relationships between variables based on assumptions made about the underlying data generation process. Causal graphical models provide a way to express conditional dependencies among variables, draw inferences when one or more variables are interviewed on, estimate causal effects, and compute counterfactuals (Shpitser and Pearl 2008). For example, (Glymour and Herington 2019) use causal graphical models to show how different measurements used to quantify algorithmic bias present conflicting pathways on the problem, and thus conflicting means of addressing the downstream effects.

Simulation models can also be understood as network-theoretic models of cause and effect in sociotechnical systems. The best example of this are system dynamics models. System dynamics models are a particular form of simulation that involves constructing a large discrete-time differential equation by creating a network of nodes (stocks) and edges (flows) that represent how some quantity of interest (e.g. money) moves through the system over time. These models can be used to evaluate how complex systems of causes and effects lead to particular outcomes. For example, Ghamizi et al. (2020) use a system dynamic simulation to keep track of the number of people in different states (e.g. exposed to Covid, infected by Covid) and model transmission probabilities between states (e.g. exposed to infected). By leveraging this simulation model, they are able to investigate the impact of policy to lift lockdowns. System dynamics models have also been used in the study of social inequality, e.g. in the analysis of the efficiency and scale of policies on foster youth (Fowler et al. 2020) and in analyses of how to reduce homelessness (Fowler et al. 2019). Specifically, Fowler et al. (2020) leverages system dynamic simulation to model the flow of child through the foster care system and test the impact that scaling and increasing efficiency of The Family Reunification Program have on reunification rate of foster care youth.

Both causal graphical models and system dynamics models are quantitative modeling techniques. However, scholars using both of these methods have begun to investigate ways of making model construction processes more collaborative and qualitative for both causal graphical models (Mayoux and Chambers 2005) and system dynamics models (Martin Jr et al. 2020). For example, the community of scholarship around community-based system dynamics modeling encourages the co-construction of system dynamics models between researchers and affected communities (Hovmand 2014). The abstraction hierarchy, statistical modeling of real-world data, and simulation models (in particular system dynamics models) all therefore share, at least implicitly, an underlying network graph that encodes assumed or identified causes and effects. Further, recent work emphasizes the qualitative nature underlying all three of these approaches.

The primary differences between system dynamics modeling, empirical work using causal graphs, and the AH are thus not as significant as they appear at first. Still, two differences are worth noting. First, in contrast to causal graphical models and system dynamics models, the AH explicitly forces us to explore and explain the system at varying levels of abstraction. This allows us, for example, to characterize the overarching causes of the system (the *foundational purposes*) in the same model as the minutia of, e.g., user profile layouts. Doing so, we find, helps to create interdisciplinary discussions that move the project forward.

Second, the AH approach does not require a final model that can be immediately instantiated empirically. As we show below, this allows for a model to emerge that captures different components of the system to be modeled that we might capture in constructing a system dynamics model or a probabilistic graphical model. As such, while this is outside the bounds of the present work, we see Abstraction Hierarchies as an important complement to causal graphical models and system dynamics simulations. Specifically, we believe that new methods in causal graphical modeling that allow for quantities that are known but cannot be observed or quantified (D'Amour 2019; Manski 2003) might have important synergies with Abstraction Hierarchies, which can qualitatively provide a joint model of both quantifiable and unquantifiable system components.

Related work on social media platforms

System-level perspectives on the interrelationships between people and technology have long been a staple of those studying sociotechnical systems (Orlikowski 2008; Ahuja and Carley 1998). More recently, literature relevant to this topic has focused heavily on how algorithms within social media platforms interact with the social aspects of these platforms. For example, Selbst et al. (2018) argue explicitly for a socio-technical frame of analysis for the relevant fair-ML problems that "recognizes... that a machine learning model is part of a socio-technical system, and that the other components of the system need to be modeled." (pg. 3). Green (2018) emphasizes that algorithms must be viewed not as isolated pieces of code, but as a part of a broader system in which algorithms are embedded within institutions that have distinct rules, norms and goals. And, amongst other critiques, Hoffmann (2019) notes that questions about algorithmic fairness too strongly emphasize the importance of discrimination against "bad actors", rather than acknowledging the existence of a "bad" (or discriminatory) system. These works emphasize the fact we cannot understand social inequality on social media platforms as being produced only by algorithms, or by people, but both by the broader system composed of these parts and by inspection of their more micro-level interconnections.

More directly relevant to our case study is the work of Hutson et al. (2018), who study how the design of intimate platforms shape interpersonal bias. They observe three classes of mechanisms that shape bias on these sites- algorithms that match individuals to each other, community politics and messaging, and tools that allow users to search, sort, and filter. In a similar vein, Levy and Barocas (2017) identify ten different categories of platform design and strategy choices that may impact the degree of discrimination on the platform. These categories include mechanisms by which platforms set policies (impacted by, for example, the composition of the community and hiring policies at the company itself), structuring of interactions (e.g. is there a rating or review system on the platform, and how does it work?), and the processes by which companies monitor and measure discrimination on their platform.

Other scholars focus on bringing known mechanisms from the "offline" world to bear in the online setting. For example, in the context of online dating, scholars have closely assessed the extent to which mechanisms that are known to have biased dating patterns before the web emerge online. For example, market structures induced by the nature of

Table 1 An overview of the steps required for AH-development with examples from our application to online dating

Steps	Example from online dating (OD) setting
1. Choose the *perspectives*	Platform and User
2. Decide on *model boundaries*	Only one online dating platform
	Exclude other contributors to racial inequality
	Exclude processes by which users join and leave the platform
	Analyze an online dating platform focused on building long-term relationships
3. Construct the *Abstraction hierarchy*	Graph where nodes exist at five different levels of analysis, with edges between select nodes in consecutive layers
	E.g., We added 'Finding a long term partner' on the functional purpose layer, 'Cultural expectations' on the layer below (abstract functions), and an edge between the two

mate finding induce biases in, e.g., the likelihood of a less attractive individual dating a much more attractive individual (Das and Kamenica 2005; Herrenbrueck et al. 2018), and homophily predicts, correctly, that same-race preferences will exist in dating (Bruch and Newman 2018).

Yet even with these widely established market and homophily-based mechanisms, much remains to be understood. For example, scholars disagree on the extent to which matching in markets, relative to homophily-based sorting, causes bias in partner selection (Bruch and Newman 2018; Lewis 2016). And homophily itself is itself a generic term for a phenomena that likely has many underlying root cognitive and structural causes (McPherson et al. 2001). More broadly, there are myriad cognitive and structural factors that are likely to produce bias in online platforms. Take, for example, the recent workshop on “Perceptual Biases on Social Media”,[3] which focused exclusively on how different classes of cognitive bias interact with social media platforms. Or the work of Van Dijck (2013), who describes how the norms and practices of one online platform are greatly impacted by economic factors and the decisions of other platforms in the same space.

In sum, then, while much is known about how social biases and social inequality might emerge from or be combated within social media, this knowledge is dispersed throughout myriad methodological approaches and exists at different levels of systemic analysis. We argue here that the abstraction hierarchy, and our extension of it, provide a capable tool to unify this existing and disparate literature.

Methods

The abstraction hierarchy

As shown in Table 1, there are three key steps involved in developing an abstraction hierarchy. The first step is to define the *perspective(s)* that are to be modeled. The perspective(s) chosen define through whose “eyes” the Abstraction Hierarchy is created, and also determines the number of different Abstraction Hierarchies that may need to be developed. For example, Burns et al. (2005) used multiple and separate AHs (frigate,

[3] https://pbsm2017.github.io.

environment, and enemy) to model the commands and control structure in the navy. In contrast, St-Maurice and Burns (2017) showed in their model of patient treatment with medical records how patient flows (patient perspective) and information flows (system perspective of medical records) were placed into separate views of the same AH. Multiple stakeholder perspectives can also be combined into a set of Abstraction Hierarchies that share one or more levels. For example, Euerby and Burns (2012) develop a shared community/online platform AH, where the community (people) and the platform are connected only at the bottom physical levels.

Once the perspectives have been determined, the next step is to define the *boundaries* of what is to be modeled. Boundaries are necessary to ensure that the AH does not try to model "everything," but rather that it maintains a specific focus that aligns with the most important relevant factors impacting the research question. While a physical system, like the nuclear power plant, has a clear system boundary, boundaries for intentional systems are often vague and require upfront considerations of the analysis' purpose, what can be monitored, and what is useful to include. Such a decision is obviously critical in the context of social media and societal impacts; attempts to model all possible social media, and/or all possible societal impacts, would be infeasible. There are no established bounds to what can be used to inform the system boundary construction, but common tools include prior knowledge, system documentation, interviews with subject matter experts, and reviews of the literature (Bisantz and Vicente 1994; Wilson 2014). In case there is insufficient prior work, the framework is effective in underlying areas where more investigations are needed to complete the model with requisite knowledge.

Finally, given a (set of) perspective(s) and system boundaries, we construct the abstraction hierarchy. The construction of an abstraction hierarchy (or multiple Abstraction Hierarchies) consists of identifying nodes and edges within a prespecified five-level hierarchy. The first (top) level, called the *Functional Purpose* level, defines the highest-level goals of the modeled entities within the system. As noted above, an example of a functional purpose for a social media company might be to "Maximize Profit". The second level, called the *Abstract Function* level, defines the constraints that are placed on, e.g., platforms as they seek to implement processes that will help them to achieve their functional purposes. As above, for example, social media platforms are likely constrained by some understanding of user preferences. The third level, called the *Generalized Function* level, incorporates the implemented processes that define how an entity attempts to solve, or accomplish, its functional purposes within the constraints specified by the abstract functions. This level generally represents the main processes of the system. For example, our abstraction hierarchy of an online dating platform described below and shown in Fig. 1 includes the Generalized Function "Find a Long-term Partner".

The fourth level, called the *Physical Function* level, defines the " concepts, objects, and actors that were needed to perform the processes modeled in the generalized functions" (St-Maurice and Burns 2017). For example, in our example application below, a Physical Function node of users of online dating systems is their "Mental Model of Online Dating". The fifth level, called the *Physical Forms* level, defines the components of the physical functions that manifest as physical or virtual objects or measurements. For example, while a recommendation algorithm might be a physical function, the data input to that algorithm, and its optimization function, are more aptly physical form nodes.

The creation of nodes at these five levels happens iteratively with the creation of directed edges between consecutive levels of the hierarchy. At a high level, links between levels have a consistent meaning, with the target of an edge representing a possible explanation to or cause of a problem at the source. For example, in our model below, a user struggling to "Find long term partner" (functional purpose) may be explained, or caused, by "Cultural Expectations" abstract functions that the user does not meet, or is not in tune with. And a problem with an "Algorithm" (Physical Function) maybe caused by inconsistencies in observations used to construct the training data (Physical form). To understand what that problem is, we would then look up the hierarchy, ultimately to the functional purpose of the system.

Applying concepts from network science to a completed abstraction hierarchy

Traditional studies that use the abstraction hierarchy typically draw only descriptive, introspective inferences from a completed model. These typical analyses of Abstraction Hierarchies (like that presented first in our results below) focus largely on outlining why nodes were developed, and the purposes of specific edges. Such analyses tend to miss the possibility that the model encodes larger components that form within the network.

Network science, of course, has many tools to help us identify such subcomponents of networks. Here, we identify two simple tools that are well known to do so. First, we apply the Newman-Girvan clustering algorithm (Girvan and Newman 2002) to the final abstraction hierarchy model, and analyze the resulting components. The Newman-Girvan algorithm is a common algorithm in the network analysis community used for identifying clusters. The algorithm is a divisive clustering algorithm. It creates a hierarchical clustering by iterating between two steps: 1) calculating the betweenness centrality of all edges in the network, 2) removing the edge in the network that has the highest edge betweenness centrality (Girvan and Newman 2002). Second, we construct an undirected network on the three most concrete levels of the AH, and use the resulting network to trace the diffuse set of paths from proposed interventions to outcomes of interest. The undirected network is constructed by simply taking the qualitatively constructed network graph and taking the induced network. A *path* is a connection between two nodes, a and b, in the induced network, that may traverse between intermittent nodes $c_0, ...c_K$. The length of the path is $K + 2$. Such analyses of all paths (within reason; here, paths shorter than a pre-specified path length of 8) show both the many, potentially unanticipated, ways that platform interventions can serve to increase equality, and also the constraints that apply to many, or all, of these paths.

Case study

Overview

As a case study in the utility of our framework, we focus on online dating and its links to racial inequality. Online dating sites are increasingly popular - the majority of romantic relationships now begin online (Finkel et al. 2012). A wide array of scholarship has used the digital trace data resulting from these sites to explore patterns in romantic relationships (Bruch and Newman 2018; Anderson et al. 2014; Lewis 2016). This data is particularly valuable because romantic partnerships are a uniquely difficult social relationship to characterize (Finkel et al. 2012), as even we ourselves are

Table 2 A high-level description (right) of potential interventions proposed by Hutson et al. (2018) to increase inter-racial relationships on online dating sites, and the nodes in our abstraction hierarchy that are associated with that intervention (left)

Relevant node	Proposed intervention
Fairness constraint	Add a mechanism to the algorithm to promote more racial diversity in search results
Community policies	Construct policies that restrict users from engaging in explicitly racist behaviors, e.g. listing racial preferences in their text profiles
Community messaging	Target social norms on the platform by encouraging racially heterogomous dating and/or anti-racist behaviors
Minimal group categories and filters	Assign users to randomly created social groups ("minimal groups" (Tajfel et al. 1971)) and encourage interaction amongst them by promoting them in user profiles and as filters for searching
Race-based categories and filters	Remove the ability to filter by racial categories, and remove race information from user profiles

not always certain why we are attracted to a given individual (Lizardo and Strand 2010).

While these data have been valuable in understanding dating patterns, others have begun to question how studies of online dating behaviors are influenced by decisions made by the dating platforms themselves. For example, although the concern was ultimately addressed (Anderson et al. 2015), Lewis (2015) challenged the findings of Anderson et al. (2014) use of online dating data to assess human preferences in romantic partners by claiming interference from the algorithm. Pushing this discussion one step further, Hutson et al. (2018) argue that there are three components of platforms— "search, sort, and filter tools," automated matchmaking, and community rules and norms— by which online dating platforms may influence the decisions made by users. The focus of their work, and ours, is to emphasize how these three mechanisms can serve as a means of exacerbating or mitigating *sexual racism*. Sexual racism is defined as biases (intentional or unintentional) in how romantic partners are selected that ultimately serve to reinforce racial inequality.

Hutson et al. (2018) argue that changes made by the platform can increase inter-racial dating on the site, and ultimately, through the construction of inter-racial relationships, create conditions that mitigate racial inequality. The most direct path, they argue, are the widely established ways in which racially homogomous relationships concentrate inter-generational wealth in White families in the United States, and the potential for inter-racial relationships to break these generational cycles. Here, we also point to the creation of inter-racial relationships by breaking down existing patterns of racial segregation.

A critical contribution of Hutson et al. is moving beyond simply stating that these elements of platforms can lead to increased sexual racism to identify and/or propose concrete interventions that platforms can make changes to help *reduce* racial bias. For example, drawing on the minimal group paradigm literature (Tajfel et al. 1971), they argue that sites could introduce new, abstract social groupings of users which reach across racial boundaries, and then create filters tools that allow users to search for others using these new found in-groups. Where possible, Hutson et al. also provide explicit examples of how to do so from existing platforms. For example, the authors describe an existing Japanese site that uses this approach. In explicating potential “fixes” to the structural impacts of the platform on shaping racial preferences, Hutson et al. suggest ways in which these sites can be forces for positive change.

In the present work, we construct an abstraction hierarchy that models the interventions proposed by Hutson et al. (2018) in the broader context of what is known about online dating. The interventions modeled are described in Table 2. We select this particular case study for three reasons. First, the interventions proposed by Hutson et al. are, like the examples posed in the introduction, indirect in their impact. Put another way, it is difficult, without an explicit model, to understand the precise pathways that may lead to, and/or prevent, these interventions from actually impacting racial income inequality, or racial segregation. Second, our research team has considerable experience in the study of online dating platforms and its connections to racial inequality. This allowed us to be more confident in our sample application without needing to resort to an extensive interview process of external subject matter experts.

Finally, the question we ask has already been studied by other researchers using other methods. This allows us to better understand how the abstraction hierarchy modeling framework is, or is not, useful in comparison to these approaches. Anderson et al. (2014) use data from an online dating platform to build a statistical model that teases apart the impacts of structural patterns versus personal preferences on the persistence of same-race romantic relationships. They conclude, “our findings imply that merely altering the structural environment-say, by creating more opportunities for individuals of different races to interact” would not necessarily ameliorate persistent patterns of racial homogamy in romantic relationships” (pg. 38-39). Ionescu et al. (2021) use an agent-based model to explore how the interventions posed by Hutson et al. (2018) might impact interracial relationships. They find that these interventions do, in their model, decrease racial homogamy, but argue that the benefits of the interventions are subject to difficult-to-validate assumptions about societal structure and human behavior and could have important side effects (e.g., decreasing the total number of relationships and the user satisfaction).

Implementation

Our abstraction hierarchy was constructed via an iterative process. First, we determine which perspectives we would seek to model. We decided on modeling the system from the perspective of (1) the user, and (2) the platform. We also decided to emphasize in our final model explicit points where the interventions described in Table 2 were modeled.

Second, all authors of the paper met to determine system boundaries. It was decided that the analysis would (1) focus only on a single online dating platform, and not how that platform connected to the broader eco-system, (2) not seek to explicitly model other societal processes beyond romantic relationships that lead to racial income inequality or segregation, (3) not seek to model the processes by which people join and leave the online dating platform, and (4) to only model online dating platforms that focus on long-term relationships, as opposed to those geared (normatively or explicitly) more towards shorter-term “hook ups” . These decisions limit

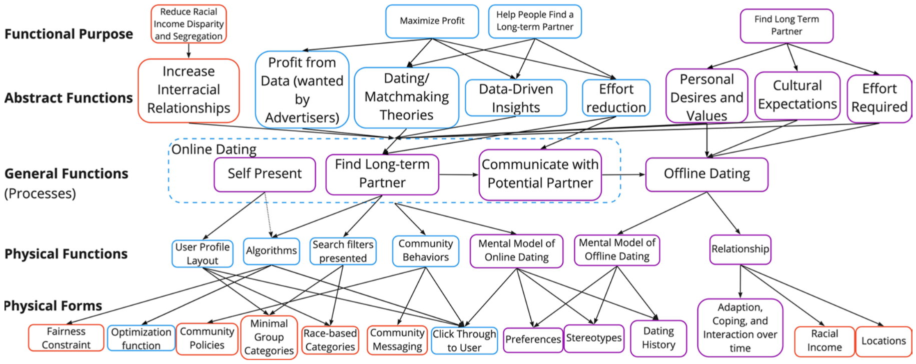

Fig. 1 Our constructed abstraction hierarchy. The five levels of the AH are listed on the left, nodes at each level are boxed phrase, with links connecting boxes between and within levels. Links to the dotted box entitled "Online Dating" connect Abstract Functions to all three of the enclosed General Functions. Nodes are colored to represent the perspective they are relevant to: the Platform (blue), or the User (purple). Additionally, nodes that are directly relevant to the proposed interventions or the primary outcomes are colored red

what can be said with the model developed, but were necessary to maintain a feasible model that could be presented as a sample application.

Having determined the boundaries of the system, two authors of the paper, one with significant experience in the abstraction hierarchy framework, and the other with expertise in the link between social media and social inequality, each created their own Abstraction Hierarchies. Both scholars made use of a recent book-length review of the online dating literature (Finkel et al. 2012) and the abstraction hierarchy model of healthcare described above (St-Maurice and Burns 2017) to do so. The models from these two authors were then combined into a single model, which was reviewed and modified by a third team member with expertise in the study of online dating platforms. These three scholars iterated to construct a final version of the Abstraction Hierarchy, which was then presented to two other team members with expertise in machine learning. This allowed the team to evaluate how easily the model could be interpreted by scholars outside the domain. Final modifications were then made based on this feedback.

Results

Abstraction hierarchy

Figure 1 displays the platform-users joint abstraction hierarchy that was developed to illustrate the intertwined nature of the online dating platform and its users (who are assumed to be seeking long-term relationships) that we modeled. Our earlier efforts in developing the AH had modeled the platform and the user independently to separate the two different perspectives; in the end, the two perspectives were connected to illustrate the influence of platform design on users and conversely how collective user behaviors can shape the system–which includes both the platform and its users–as a whole.

Functional purpose and abstract functions

The resulting platform-users AH showed that the platform and the users shared the functional purpose of finding a long-term partner. According to the model, an individual user's purpose of finding a long-term partner is fulfilled by examining and calibrating their personal values and desires, adhering to cultural expectations they feel the need to meet (e.g., whether one feels pressured to get married and to be married to someone of similar background and status), and the balancing of their required efforts (time and money invested) against resources available and opportunity costs (i.e., what they could accomplish/gain if efforts are directed elsewhere). On the other hand, from the platform perspective, the functional purpose of helping people find a long-term partner is primarily achieved by implementing (potentially pseudo-scientific) theories about dating/matchmaking, data-driven insights (e.g., what the platform has learned about user behaviors), and platform design that aims to minimize effort while maximizing returns for the users (simply referred to as effort reduction in Fig. 1). An online dating platform not only aims to help people find partners, but as a business also has the purpose of maximizing profit. Profit, in turn, is made largely from the rich data of user profiles and behaviors that third-party advertisers would pay to access. Hence we include profit from data as another abstract function of this system.

Disconnected from the platform's other general functions and abstract functions is a presumed desire to reduce racial income disparity and segregation. The explicit definition of this functional purpose was introduced because we could not find a plausible way to explain the proposed interventions from Hutson et al. (2018) without such an explicit motivator. This emphasizes a useful function of the AH; by asking us to ensure that the entire hierarchical network is coherent, we are forced to accept that certain changes we may hope social media platforms make cannot occur unless there is a foundational desire to genuinely address social inequality.

General functions

At the General Functions level, the platform and user components shared the online dating functions–self present, find long-term partner, and communicate with potential partners. These generalized terms capture a number of the activities and exchange of information that happen in online dating platforms, as defined by the extensive review article presented by Finkel et al. (2012). Specifically, they capture (1) the deliberate creation of a profile to present oneself in a certain light, (2) the search and browsing of potential matches, and (3) the one-way or two-way communication with other users. This is the level where all abstract functions intersect in guiding how the online dating process works. Conversely, the offline dating process–initial dates and the on-going maintenance and evaluation of a relationship–continues to be governed by only the individual's priorities and values. As our network analysis below further illuminates, this separation between online and offline becomes a potential limiting factor in the impacts that interventions online can have on racial inequality.

Physical functions

Physical Functions of the AH contain the fundamental components of the system. With respect to the Platform perspective, this amounts to (1) user actions (Community Behaviors), (2) the algorithms used and (3) interface design components, namely how user profiles are laid out and which filters are presented for people to use to find romantic partners.

Matchmaking algorithms are essential in the search function. It is at this point well-known that recommendation algorithms like those used in online dating websites can exacerbate various forms of bias and inequality. However, to fully understand this point, it is important to note how such recommendation algorithms are commonly constructed and trained; our discussion here follows the generic structure of algorithms used to provide recommendations on social media platforms (Bobadilla et al. 2013). Recommender algorithms for social media are typically developed using at least one of two ideas: *collaborative filtering* or *content-based recommendation*. In collaborative filtering, recommendations are based on (often implicit) feedback gained from user interactions with other "items"; in this case, user's interactions with the dating profiles of other users. In *content-based filtering*, recommendations are based on shared properties of the content; here, shared properties of the dating profiles. Regardless of the approach, such algorithms are thus trained on data from users interacting on the platform. The algorithms, in turn, impact user behavior (Wu et al. 2020). This kind of feedback loop can create situations in which the algorithm creates selection biases in the data it is trained on Mansoury et al. (2020).

Beyond this issue of feedback loops, the harms perpetrated by recommendation algorithms vary based on the types of information that is both gathered and used. For example, Anderson et al. (2014) show that biases are minimized when users are allowed to self-present their racial preferences rather than having them inferred from their site behavior (see dotted line in Fig. 1). They can, as has been noted in many other contexts (Mehrabi et al. 2021), be modified to create more "fair" results by constraining the data and/or optimization function. (Hutson et al. 2018) note that for dating sites, this "fairness constraint" intervention might take the form of purposefully presenting a more diverse set of potential partners.

Interface designs of the dating site can also be intervened upon in ways that might reduce inequality. User profile layouts and available search filters, while essential in facilitating the browsing and searching activities, ultimately dictate what is known about a user and whose profiles a user can see. During the profile creation stage,"users are often encouraged (and in some cases required) to categorize themselves according to a number of characteristics."(Hutson et al. 2018) [pg. 7] and these categories "(and the concomitant exclusion of others)" are reinforced throughout the online dating process (e.g., in searching, sorting, and filtering), which "legitimizes such categories as socially reasonable bases for including or excluding potential partners" (Hutson et al. 2018) [pg. 7]. Interventions that remove race as filters, and/or add other filters uncorrelated with race, may thus drive users to other forms of social categorization.

From the user perspective, analogous to the platform recommendation algorithm is the user's mental model about dating. An individual relies on their mental model to process the relevant information and make decisions throughout the dating process. The online dating mental model helps a user navigate the dating site, evaluate the various profiles they come across, and strategically present themselves through their profiles and communication with other users. Moving into offline dating, the user will use a slightly different mental model that is more focused on maintaining their real-life interactions and evaluating their long-term compatibility. The final physical function in the AH is the (longer-term) relationship between a couple matched by the site. The relationship is necessary to model, because it is the physical function that directly maps to the expected way in which online dating platforms can impact racial inequality: by helping to create interracial relationships that help to breach historical racial inequalities in income and place.

Physical forms

At the bottom level of the AH, we emphasize attributes of the physical functions that are most relevant to our objective of examining the role of online dating platforms in racial income inequality and segregation. We explicitly include fairness constraints alongside the optimization function to illustrate how algorithms can, as envisioned by Hutson et al. (2018) and denoted in Table 2, or cannot, be designed with equity in mind. Technically, there is an active area of research that surrounds how best to implement these efforts to make recommendation algorithms more fair. A review of this literature is out of scope for the present work, but a number of relevant reviews exist elsewhere for the interested reader (Li et al. 2021; Milano et al. 2020).

User clicks on the dating site are the most direct observations for examining users' choices and decisions, which can tell us where biases may exist, whether they are innate to the user (their mental models of dating) or the result of site designs. These observations collectively reflect the online dating community behaviors and often become the training data for commonly used recommendation algorithms on social media websites. It is, on the other hand, much harder to observe and track individuals' offline behaviors, i.e., the *adaption, coping and interaction over time* that occurs within all long term relationships (Finkel et al. 2012). An individual's preferences and stereotypes may be inferred to some extent from their online profiles and their clicks on other profiles, but a portion of these attributes will remain private to the individual or exist, as shown by Anderson et al., subconsciously. The dating history of a person can reveal some of their innate beliefs and values, and these personal experiences will also in turn influence the person's value system and thus their dating preferences. For example, one way for site-specific interface design to overcome inter-racial barriers in dating is to present more matches from other races that are compatible with the individual in various aspects. Such mechanisms will allow a user to reassess their stereotypes (a feature of their mental model) about compatibility issues associated with inter-racial relationships. With each relationship formed, initiated via the dating site, the make up of the couple (e.g., their racial income differences, their respective locations or distance between their locations) can contribute to societal efforts in reducing racial income disparities and segregation.

The platform-user abstraction hierarchy described above provides a way to explain how elements of an online dating site and its users are tied to the overall purposes of the

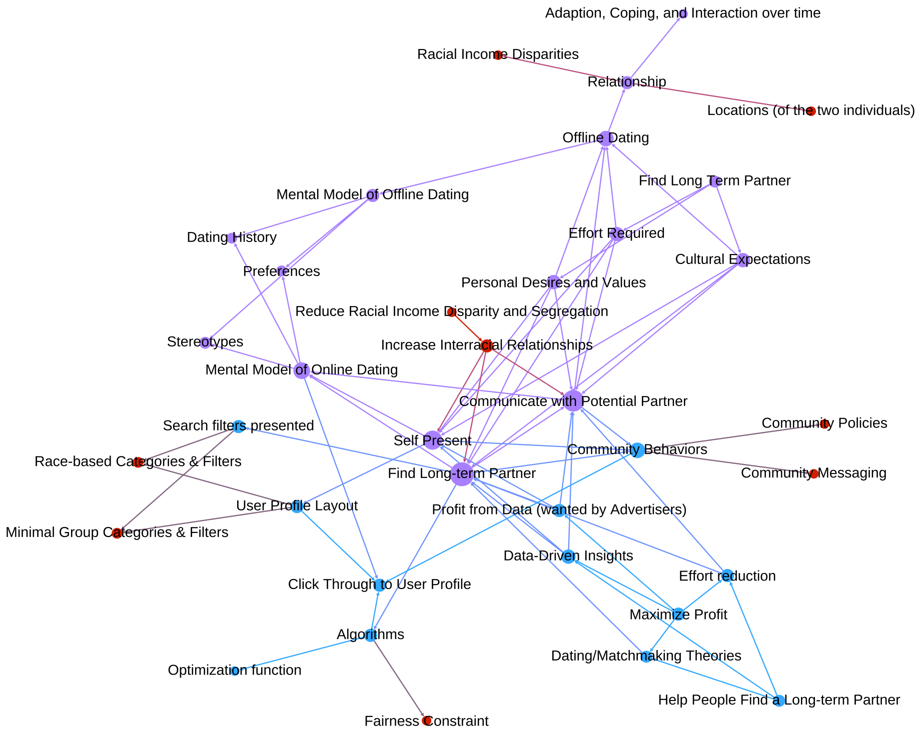
Adaption, Coping, and Interaction over time
Racial Income Disparities
Relationship
Locations (of the two individuals)
Offline Dating
Find Long Term Partner
Mental Model of Offline Dating
Dating History
Effort Required
Preferences
Cultural Expectations
Personal Desires and Values
Reduce Racial Income Disparity and Segregation
Stereotypes
Increase Interracial Relationships
Mental Model of Online Dating
Communicate with Potential Partner
Search filters presented
Community Policies
Self Present
Race-based Categories & Filters
Community Behaviors
Find Long-term Partner
Community Messaging
User Profile Layout
Profit from Data (wanted by Advertisers)
Minimal Group Categories & Filters
Data-Driven Insights
Effort reduction
Click Through to User Profile
Maximize Profit
Algorithms
Dating/Matchmaking Theories
Optimization function
Help People Find a Long-term Partner
Fairness Constraint

Fig. 2 A network visualization of the Abstraction Hierarchy. Nodes are colored using the same approach described in Fig. 1. Nodes are sized by their eigenvector centrality, computed on the undirected network

Table 3 Results from applying the Newman-Girvan clustering algorithm to our abstraction hierarchy

Component name	Nodes
Relationship	Adaption, coping, and interaction over time, locations (of the two individuals), racial income disparities, relationship
Algorithms, policies, and norms	Algorithms, click through to user profile, community behaviors, community messaging, community policies, fairness constraint, optimization function
Categorization	minimal group categories & filters, race-based categories & filters, search filters presented, user profile layout
User cognition	dating history, mental model of offline dating, mental model of online dating, preferences, stereotypes
User goals, constraints, and actions	communicate with potential partner, cultural expectations, effort required, find long term partner, increase interracial relationships, offline dating, personal desires and values, reduce racial income disparity and segregation, self present
Platform goals and constraints	Data-driven insights, dating/matchmaking theories, effort reduction, find long-term partner, help people find a long-term partner, maximize profit, profit from data (wanted by advertisers)

The left-hand column represents a name we construct to refer to each cluster in the text; on the right, the nodes in the cluster

online dating ecology through the illustrated means-end links. The network analysis discussed below provides a closer look to these important links.

Network analysis of the abstraction hierarchy

The abstraction hierarchy we developed has 36 nodes and 70 total edges. Figure 2 presents the same information as in Fig. 1, but laid out using a standard network layout algorithm, the ForceAtlas2 algorithm (Jacomy et al. 2014). This algorithm is an efficient force-directed algorithm used in common graphing libraries in the network science field, such as the Gephi tool.[4] A number of parallels exist between the hierarchical visualization in Fig. 1 and the network visualization in Fig. 2. In both, for example, general functions are central, and the physical forms of the interventions are at the edges. We now turn to an analysis of the components within this network, and our analysis of paths from intervention to outcome.

Cluster analysis

Using the Newman-Girvan algorithm, we identified six clusters in our abstraction hierarchy. Here, we briefly discuss what we can learn from looking in detail at each cluster, and then turning to what we can understand by considering how the clusters themselves are interconnected.

The first cluster in Table 3 contains nodes relevant to the relationship as it develops offline, and how this relationship (or, in reality, relationships collectively) can serve the ultimate purpose of reducing racial income inequality and segregation. This cluster reinforces the fact that there is a strong separation between the processes that occur on the platform which serve to initiate a relationship, and the longer-term processes that encourage a lasting partnership which ultimately can lead to shared location and finances.

The second cluster in Table 3 contains the platform's matching algorithm and its physical forms, community behavior and its physical forms, and the primary action that users take that is modeled in our AH: clicking on another user's profile. This cluster emphasizes the dual purpose that user behavior serves on the platform. On the one hand, user behavior becomes training data for the algorithm that is in turn embedded into future matches presented to users. On the other, user behavior collectively defines community behavior, which in turn defines community norms. These norms, Hutson et al. emphasize, can also be targeted for intervention, by changing company messaging and/or policies to, e.g., reflect anti-racist ideals. This cluster therefore emphasizes the co-construction of site norms, user behavior, and the algorithm, and highlights why interventions that target norms through messaging and/or policy may be effective.

The third cluster in Table 3 identifies a separate set of platform functions that center around the categorization processes the site allows users to leverage in profiles and in their searches. This cluster includes two interventions proposed by Hutson et al., one to introduce minimal groups as additional forms of social categorization, and the other to remove racial categorization from filters and profiles. As Anderson et al. (2014) results imply, and Ionescu et al. (2021) point out in their simulation, the explicit removal of race information from selection criterion may not necessarily lead

[4] https://gephi.org/.

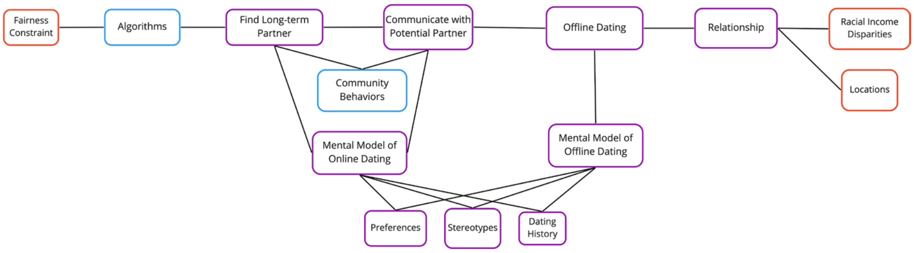

Fig. 3 All paths in the undirected network of length less than eight that move only through general functions and physical functions/forms, from Fairness Constraint to the two inequality physical form nodes

to a change in interracial relationships, because user stereotypes that link race to other forms of social categorization can, and do, exist.

As such, it is important to consider how the Categorization cluster is distinct from, but has important relationships to, the fourth "User Cognition" cluster we identify, which centers on user's mental models and the physical forms that underlie them. Linkages between categorization interventions proposed by Hutson et al. and user stereotypes have been pointed out in prior work. However, the AH, further informed by our clustering, presents a clear picture of the problematic assumption: categorizations presented by the site have no bearing on user's mental models of dating, and thus are restricted in how they can impact user decisions because users can always identify their own processes of categorization that exist due to the correlations of social dimensions (Blau 1977).

The final two clusters in Table 3 represent the functional, abstract, and general functions of users, and separately, the functional and abstract purposes of platforms. However, the clustering places the functional purpose associated with addressing racial inequality in the user-oriented cluster, rather than with the platform. This observation underscores the overarching issue we found in our AH with assumptions that platforms can drive change on matters of racial inequality. Simply put, we could not find an obvious way in which this functional purpose linked to the core functional purposes of online dating platforms: making money, and making matches. As such, the drive for racial inequality stems more from a society-wide goal, which in turn suggests a deeper connection to the collection of users that make up the society of the site. In turn, on most social media platforms today, as with society in general, individuals are left to make decisions about individual preferences versus collective goods (Rogers 2020; Schröder et al. 2017). One core area, then, where our AH calls for intervention is in creating more concrete links, e.g. through legislation tying equity-based outcomes to profit, between the standard functional purposes of social media platforms and the goal of addressing, e.g., racial inequality.

Path analysis

Our sample path analysis, displayed in Fig. 3, displays the range of pathways through which the proposed intervention on the algorithm might impact the racial inequality outcomes we care about. The diagram emphasizes two important points.

First, the obvious pathway to impact is not the only pathway to impact. In Fig. 3, the shortest and most obvious path from intervention to impact assumes that algorithmic recommendations will directly impact partner selection, which in turn will impact who the user communicates with online, begins dating offline, and ultimately forms a relationship with. This relationship, if inter-racial, would in turn (on average) lead to a decrease in racial income inequality and location.

However, one need not assume that the recommendation algorithm has a direct impact on a given user's partner selection in order to lead to an increase in social equality. We highlight two alternative paths here. First, as highlighted in the path through "Community Behaviors", if the algorithmic intervention works for some users, but not others, it may still be enough to shift community behaviors, and thus norms, in a direction that promotes racial heterogamy. As noted above, these norms further reinforce more diverse ("fair") recommendations by the algorithm, creating further, more organic, shifts towards diversity in recommendation. Second, as a corollary to Hutson et al. (2018) claim about removing race from filtering options, simply seeing more diverse search results, even if pairwise communication does not ensue, can shift user's mental model of dating by shifting their stereotypical associations of race and their preferences. In turn, mental models of online dating, which tie to the same stereotypes and preferences as mental models of offline dating, can reshape how the user makes decisions in offline settings. While the potential for negative effects of well-intentioned interventions is well known (Suresh and Guttag 2019), these observations from our AH reinforce the point that interventions to reduce inequality can also have unintended positive, diffuse consequences as well.

However, Fig. 3 also re-emphasizes that there are, for each of these potential diffuse paths towards increased equality, a network choke-point that lives largely beyond the impact of any proposed intervention. In particular, offline dating is governed, as Fig. 1 reminds us, by more static abstract functions that incorporate individual's desires and values, and perhaps more importantly here, cultural expectations placed on them by others. Our AH shows that these cultural expectations, particularly when derived from close personal ties (e.g. family), must be faced by *any* platform intervention, and pathways to change for these abstract functions require even more indirect pathways than those expressed in Fig. 3. The AH thus helps us to both understand the potential paths to addressing inequality, and reminds us of the often difficult to quantify, more abstract factors that can restrict change.

Discussion

Our work, conducted using an abstraction hierarchy methodology combined with network analytic concepts, presents a number of important insights to prior empirical (Anderson et al. 2014; Lewis 2016), simulation-based (Ionescu et al. 2021), and theoretical (Hutson et al. 2018; Finkel et al. 2012) research focusing on race, racial equality, online dating and the potential for intervention.

Existing empirical research focusing on race and online dating must focus on specific mechanisms that are testable or quantifiable. As noted above, most such work thus focuses on how personal preferences at specific stages of the dating process (Anderson et al. 2014; Lewis 2016), or overarching patterns of popularity and assortativity (Bruch and Newman 2018), are associated with race. The AH approach taken here complements these efforts by providing, through a rigorous qualitative research process, a model

(Fig. 1) of how these various identified mechanism fit together into a broader sociotechnical ecosystem of racialized romantic relationships. What emerges from Fig. 1, and our use of network science concepts to help analyze it, is the existence of myriad intertwined, long-term pathways that both mediate and may subtly enhance the impact of potential interventions. With respect to the latter, for example, we note above that these paths challenge Anderson et al.'s assumption that structural changes on the platform cannot impact patterns of interracial dating, because they show how those changes can, over time, impact preferences.

Theoretical work in online dating has made some of the same arguments. In particular, our results rely in part on the work of Finkel et al. (2012), who outline a process model of the general functions of online dating, and of Hutson et al. (2018), who outline physical functions of online dating platforms and provide a perspective on potential interventions. The AH approach taken here compliments these efforts in two ways. First, it introduces a formal qualitative methodology for modeling socio-technical systems. This model helped, for example, to outline potential difficulties in some of the interventions proposed theoretically by Hutson et al. (2018), in particular the fact that the boundaries of online dating sites bleed into the offline dating world, which may restrict progress made on the platform.

Second, the AH approach here complements the more individualistic perspective of Finkel et al. (2012) in asking us to start at the functional purpose of the platform and its users. Doing so forces our analysis to focus not only on individual decisions, but also on the broad macro-social and macro-economic conditions in which social media platforms and their links to social (in)equality play out. Doing so has two additional benefits that move beyond our current understanding of interventions in online dating platforms. First, it emphasizes that in the absence of a goal and/or incentive for lasting social change, it is difficult to conceive of how platforms will produce interventions that lead to genuine shifts in social equality. More explicitly, our AH surfaces the fact that attempts to address platform issues that are motivated solely by profit-oriented responses to public outcry are bound to fail, because there is no sustainable functional purpose to maintain momentum. This is consistent with the ways in which proposed interventions by existing platforms, most notably Facebook, have played out since the 2016 U.S. election, and the reason why governments are seeking policy solutions (Hao 2021) . Second, and related, is that it encouraged our team to rethink the bounds of interventions. More specifically, it emphasizes that while interventions to, e.g. reduce hate speech are important, equally necessary are the social movements that serve as 'interventions" which drive social media platforms to take on social justice and/or social equality-oriented perspectives.

Thus a core benefit of the modeling approach taken in this work relative to other qualitative research in online dating and racial equality is that the AH is more formal in how it aims to characterize the inner-workings of a sociotechnical system. Others have argued that simulation has similar benefits, in that it provides a concrete and computational explanation of a theoretical model. In our own prior work (Ionescu et al. 2021), we make a similar argument in this same context of online dating and racial equality. As noted above, the AH approach complements a simulation approach in that it does not require us to pre-define or experiment with a vast space of hyperparameters which might influence results.

However, we also found as we carried out the work presented here that the AH was also an important compliment to simulation modeling because our interdisciplinary team found it easiest to engage each other in discussion through the abstraction hierarchy than through simulation results. Specifically, we were able to compare and discuss clearly labeled nodes and their links, rather than diagrams of computer code and equations specifying the functions of specific parameters. This ability to use the abstraction hierarchy to talk across disciplinary boundaries helped us to rectify the fact that systemic thinking requires a need to bridge often artificial disciplinary boundaries. Future work is needed, however, to take systemic thinking to the forefront in an effort to move beyond disciplinary thinking all together.

In sum, then, we believe that our work shows how the Abstraction Hierarchy modeling framework can serve as a bridge between more quantitiative modeling approaches, which are necessarily narrow in scope, and theoretical efforts aiming to bridge multiple levels of analysis in search of a more coherent overarching conceptual model. Finally, our extensions to the AH with even the most basic network methods proved to be useful. While findings from both the cluster and path-based analyses were potentially possible without network methods, clustering rapidly surfaced the core components of the AH that were modeled upon completion, and path analysis emphasized the myriad and indirect paths through which on platform interventions may address social inequality.

Conclusion

The present work introduces a tool from cognitive engineering, namely the abstraction hierarchy, as a qualitative, network-centric method that helps us to answer the question: how do we know if changes made by social media platforms to address social inequality will actually work? We then extended the typical abstraction hierarchy analysis to leverage basic network analytic tools. These allowed us to surface additional insights from the model. Finally, we applied this tool to a case study of how potential interventions for online dating websites proposed in prior work (Hutson et al. 2018) might actually work to reduce racial inequality.

Our use of the AH comes in large part from our frustration with what is possible with existing quantitative methods, in particular with respect to how they can be practically used in interdisciplinary teams aimed at interventions in sociotechnical systems to increase social inequality. We have argued, and sought to show here in our sample application, that the abstraction hierarchy provides two important mechanisms— an easily understandable graph and a hierarchical modeling framework—that can help to break down these boundaries.

A core point of emphasis in our work is that in order to truly understand the effects of interventions on social media platforms to increase social equality, it is critical to use network formalisms. We depart from the existing literature on this point in two important ways. First, scholars of networked inequality have placed significant emphasis on how social network structure creates and inscribed social inequality. Instead, here, we emphasize how social inequality is inscribed and created within network of interconnected causes and effects that underlie these sociotechnical systems. That is, rather than think only about how social relationships shape and are shaped by these technologies, we can use network approaches to characterize the interrelationships between, e.g., the beliefs, algorithms, and profit motives of people and organizations, and how these all combine to enhance and/or restrict the impact of possible interventions.

The abstraction hierarchy is, as noted, not novel in its emphasis on the network structures of causes and effects. Most obviously, probabilistic graphical models and system dynamic models explicitly encode cause/effect structures in a graphical format. Still, the abstraction hierarchy, and our networked conceptualization of it, brings a new perspective here as well. In contrast to a probabilistic graphical modelling approach, we were not bound to discussions of the correct statistical model (e.g. linear versus additive regression), or limitations of the data, and could instead focus more widely on highlighting misunderstandings, different perspectives, and difficult choices about what the boundaries of the system of interest were. This allowed us to analyze, qualitatively, potential causal pathways between the interventions proposed by Hutson et al. (2018) and a potential reduction in racial inequality. These pathways would have been difficult to operationalize quantitatively. This, in turn, helped us to highlight both "obvious" facts—in particular, that all interventions occurring on the platform are subject to the myriad factors that arise when a relationship moves primarily offline— and less obvious facts—such as the various indirect cognitive, normative, and value-oriented changes that are possible via interventions on social media platforms that may lead to long-term change. Some of these less obvious pathways are noted by Ionescu et al. (2021), but the AH here expresses them in an easy-to-understand visual format that promotes further interdisciplinary discussion.

Our work therefore places the abstraction hierarchy as an important future tool for understanding social media systems and social inequality, and methodologically extends existing approaches to the abstraction hierarchy with ideas from network science. However, these benefits are not without limitations. First and foremost, as argued in other contexts (Bisantz and Vicente 1994), it can be difficult to construct an abstraction hierarchy without the guidance of a methodological expert. Further work is hence needed to make the AH modeling framework more accessible to non-experts. Second, the method also currently requires significant domain expertise; developing an AH for a new socio-technical system, rather than for a well-known one, will thus also require additional work. Finally, the method as proposed here (although not in general St-Maurice and Burns 2017) is heavily influenced by the researchers' *beliefs* of the user and platform perspectives. While our beliefs are based on decades of combined work in relevant areas, our AH would nonetheless benefit from additional data points drawn explicitly from the perspectives of the involved stakeholders. Similar concerns have arisen in the use of system dynamics models, pushing scholars to adopt more community-based approaches to modeling that may also be useful here (Hovmand 2014; Martin Jr et al. 2020).

Limitations aside, there are a number of other areas in the study of socio-technical systems and social inequality that may benefit from the abstraction hierarchy approach. For example, the literature on algorithmic bias is inundated with statements emphasizing that "human bias" and "structural factors" can potentially cause algorithmic bias, and in turn inequality. While true, these claims are not specific enough to begin to think about interventions, or what the limits of those interventions might be. The AH gives us a tool to incorporate both these general, widely understood causes, as well as a more detailed look at these causal processes at the physical level. For example, a number of papers have shown that machine learning models of language in the field of Natural Language Processing (NLP) have been shown to harbor biases in the alignment of gender to different occupations (Bolukbasi et al. 2016). These gender biases emerge in *other* NLP tools that use the output from these gender-biased models to, for example, caption images (Zhao et al. 2017). In turn, these image captions cause *representational harm* (Barocas et al. 2017), reinforcing gendered stereotypes of occupations. Finally, these gendered stereotypes can influence individuals to self-select out of higher-status occupations at a young age (Bian et al. 2017). It is only in this last step, then, that the biases of the initial NLP algorithm manifest in social inequality. Without a systematic view of the pathway between NLP models and biases, any debiasing intervention might work like whack-a-mole (e.g., removing the biases from the word embedding algorithm without doing the same for the image caption algorithm would likely still elicit representational harm). The AH, combined with path analysis, can thus complement the higher level insights by helping us better identify and understand these pathways. As such, we hope that our work can serve as a jumping off point for scholars in a wide range of social computing fields to adopt the abstraction hierarchy method for their own research problems.

Abbreviations

AH	Abstraction hierarchy
NLP	Natural language processing

Acknowledgements
Not applicable.

Author contributions
KJ, SI, and HC conducted the initial analyses, YD, PS, AH, AR supported efforts. All authors read and approved the final manuscript.

Funding
K.J, Y.D., P.S., and A.R. were supported by NSF IIS-1939579.

Availability of data and materials
Not applicable.

Declarations

Competing interests
The authors declare that they have no competing interests.

Received: 15 December 2021 Accepted: 19 June 2022
Published online: 06 July 2022

References

Achonu J, Jamieson G (2003) Work domain analysis of a financial system: an abstraction hierarchy for portfolio management. In: Proceedings of the 22nd European annual conference on human decision making and control, vol 1, pp 103–109

Ahuja MK, Carley KM (1998) Network structure in virtual organizations. J Comput-Med Commun 3(4)

Anderson A, Goel S, Huber G, Malhotra N, Watts DJ (2014) Political ideology and racial preferences in online dating. Soc Sci 1:28–40

Anderson A, Goel S, Huber G, Malhotra N, Watts DJ (2015) Rejoinder to Lewis. Sociol Sci 2:32–35

Barocas S, Crawford K, Shapiro A, Wallach H (2017) The problem with bias: From allocative to representational harms in machine learning. Information and Society (SIGCIS), Special Interest Group for Computing

Baumer EP, Silberman MS (2011) When the implication is not to design (technology). In: Proceedings of the SIGCHI conference on human factors in computing systems, pp 2271–2274

Bian L, Leslie S-J, Cimpian A (2017) Gender stereotypes about intellectual ability emerge early and influence children's interests. Science 355(6323):389–391

Bisantz AM, Vicente KJ (1994) Making the abstraction hierarchy concrete. Int J Hum-Comput Stud 40(1):83–117

Blau PM (1977) A macrosociological theory of social structure. Am J Sociol:26–54

Bobadilla J, Ortega F, Hernando A, Gutiérrez A (2013) Recommender systems survey. Knowl-Based Syst 46:109–132

Bokányi E, Hannák A (2020) Understanding inequalities in ride-hailing services through simulations. Sci Rep 10(1):1–11

Bolukbasi T, Chang K-W, Zou JY, Saligrama V, Kalai AT (2016) Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In: Advances in neural information processing systems, pp 4349–4357

Bond S (2021) Facebook, Twitter, Google CEOs Testify Before Congress: 4 Things To Know. NPR . Chap. Law

Bruch EE, Newman MEJ (2018) Aspirational pursuit of mates in online dating markets. Sci Adv 4(8):9815

Burns CM, Bryant DJ, Chalmers BA (2005) Boundary, purpose, and values in work-domain models: Models of naval command and control. IEEE Trans Syst Man Cybern-Part A Syst Hum 35(5):603–616

Carley K (1993) Coding choices for textual analysis: a comparison of content analysis and map analysis. Sociol Methodol:75–126

Chen GM, Pain P, Chen VY, Mekelburg M, Springer N, Troger F (2020) 'You really have to have a thick skin': a cross-cultural perspective on how online harassment influences female journalists. Journalism 21(7):877–895

Christensen AP, Kenett YN (2021) Semantic network analysis (semna): a tutorial on preprocessing, estimating, and analyzing semantic networks. Psychol Methods

D'Amour A (2019) On multi-cause causal inference with unobserved confounding: counterexamples, impossibility, and alternatives. arXiv:1902.10286

Das S, Kamenica E (2005) Two-sided bandits and the dating market. In: IJCAI, vol 5, p 19

Davidson T, Warmsley D, Macy M, Weber I (2017) Automated hate speech detection and the problem of offensive language. In: Proceedings of the international AAAI conference on web and social media, vol 11

Epstein JM (1999) Agent-based computational models and generative social science. Complexity 4(5):41–60

Espín-Noboa L, Wagner C, Strohmaier M, Karimi F (2021) Inequality and inequity in network-based ranking and recommendation algorithms. Preprint arXiv:2110.00072

Euerby A, Burns CM (2012) Designing for social engagement in online social networks using communities-of-practice theory and cognitive work analysis: A case study. J Cogn Eng Decis Mak 6(2):194–213

Finkel EJ, Eastwick PW, Karney BR, Reis HT, Sprecher S (2012) Online dating: a critical analysis from the perspective of psychological science. Psychol Sci Public Interest 13(1):3–66

Fowler PJ, Hovmand PS, Marcal KE, Das S (2019) Solving homelessness from a complex systems perspective: insights for prevention responses. Annu Rev Public Health 40:465–486

Fowler PJ, Marcal KE, Chung S, Brown DS, Jonson-Reid M, Hovmand PS (2020) Scaling up housing services within the child welfare system: policy insights from simulation modeling. Child Maltreatm 25(1):51–60

Galán JM, Izquierdo LR, Izquierdo SS, Santos JI, Del Olmo R, López-Paredes A, Edmonds B (2009) Errors and artefacts in agent-based modelling. J Artif Soc Soc Simul 12(1):1

Ghamizi S, Rwemalika R, Veiber L, Cordy M, Bissyandé TF, Papadakis M, Klein J, Traon YL (2020) Data-driven simulation and optimization for covid-19 exit strategies. In: Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining

Girvan M, Newman ME (2002) Community structure in social and biological networks. Proc Natl Acad Sci 99(12):7821–7826

Glymour B, Herington J (2019) Measuring the biases that matter: the ethical and casual foundations for measures of fairness in algorithms. In: Proceedings of the conference on fairness, accountability, and transparency. FAT* '19. ACM, New York, NY, USA, pp 269–278

Green B (2018) "Fair" risk assessments: a precarious approach for criminal justice reform. In: 5th workshop on fairness, accountability, and transparency in machine learning

Hannak A, Soeller G, Lazer D, Mislove A, Wilson C (2014) Measuring price discrimination and steering on e-commerce web sites. In: Proceedings of the 2014 conference on internet measurement conference, pp 305–318. ACM, New York, NY, USA

Hao K (2021) How Facebook got addicted to spreading misinformation. https://www.technologyreview.com/2021/03/11/1020600/facebook-responsible-ai-misinformation/

Herrenbrueck L, Xia X, Eastwick P, Hui CM (2018) Smart-dating in speed-dating: How a simple Search model can explain matching decisions. Eur Econ Rev 106:54–76

Hoffmann AL (2019) Where fairness fails: on data, algorithms, and the limits of antidiscrimination discourse. Commun Soc Inf

Hovmand PS (2014) Group model building and community-based system dynamics process. In: Community based system dynamics. Springer, New York, NY, USA, pp 17–30

Huszár F, Ktena SI, O'Brien C, Belli L, Schlaikjer A, Hardt M (2022) Algorithmic amplification of politics on twitter. Proc Natl Acad Sci 119(1)

Hutson JA, Taft JG, Barocas S, Levy K (2018) Debiasing desire: addressing bias & discrimination on intimate platforms. Proc ACM Hum-Comput Interact 2(CSCW):73–17318

Ionescu S, Hannák A, Joseph K (2021) An agent-based model to evaluate interventions on online dating platforms to decrease racial homogamy. In: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, pp 412–423

Jackson SJ, Bailey M, Welles BF (2020) # HashtagActivism: networks of race and gender justice. MIT Press, Boston

Jacomy M, Venturini T, Heymann S, Bastian M (2014) Forceatlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software. PloS one 9(6):98679

Jaidka K, Zhou A, Lelkes Y, Egelhofer J, Lecheler S (2021) Beyond anonymity: network affordances, under deindividuation, improve social media discussion quality. J Comput-Med Commun

Koller D, Friedman N (2009) Probabilistic graphical models: principles and techniques. MIT press, Boston

Levy K, Barocas S (2017) Designing against discrimination in online markets. Berkeley Tech LJ 32:1183

Lewis K (2015) Studying online behavior: comment on Anderson et al. 2014. Sociol Sci 2

Lewis K (2016) Preferences in the early stages of mate choice. Soc Forces 95(1):283–320

Li Y, Ge Y, Zhang Y (2021) Tutorial on fairness of machine learning in recommender systems. In: Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval, pp 2654–2657

Lizardo O, Strand M (2010) Skills, toolkits, contexts and institutions: clarifying the relationship between different approaches to cognition in cultural sociology. Poetics 38(2):205–228

Malik MM, Lamba H, Nakos C, Pfeffer J (2015) Population bias in geotagged tweets. In: Ninth international AAAI conference on web and social media

Manski C (2003) Partial identification of probability distributions: springer series in statistics. Springer, New York

Mansoury M, Abdollahpouri H, Pechenizkiy M, Mobasher B, Burke R (2020) Feedback loop and bias amplification in recommender systems. In: Proceedings of the 29th ACM international conference on information & knowledge management, pp 2145–2148

Martin Jr D, Prabhakaran V, Kuhlberg J, Smart A, Isaac WS (2020) Participatory problem formulation for fairer machine learning through community based system dynamics. Preprint arXiv:2005.07572

Matthes J, Kohring M (2008) The content analysis of media frames: toward improving reliability and validity. J Commun 58(2):258–279

Mayoux L, Chambers R (2005) Reversing the paradigm: quantification, participatory methods and pro-poor impact assessment. J Int Dev 17(2):271–298

McPherson M, Smith-Lovin L, Cook J (2001) Birds of a feather: homophily in social networks. Annu Rev Sociol 1:415–444

Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A (2021) A survey on bias and fairness in machine learning. ACM Comput Surv (CSUR) 54(6):1–35

Milano S, Taddeo M, Floridi L (2020) Recommender systems and their ethical challenges. Ai Soc 35(4):957–967

Munger K (2017) Tweetment effects on the tweeted: experimentally reducing racist harassment. Polit Behav 39(3):629–649

Orlikowski WJ (2008) Using technology and constituting structures: a practice lens for studying technology in organizations. Resources. Co-Evolution and Artifacts. Springer, New York, NY, USA, pp 255–305

Pearl J (1998) Graphical models for probabilistic and causal reasoning. Quant Rep Uncertain Imprecis:367–389

Radford J, Joseph K (2020) Theory in, theory out: the uses of social theory in machine learning for social science. Front Big Data 3:18

Read GJ, Salmon PM, Lenné MG, Stanton NA (2015) Designing sociotechnical systems with cognitive work analysis: putting theory back into practice. Ergonomics 58(5):822–851

Rogers KB (2020) The problem of order: understanding how culture predicts social action. Sociol Compass 14(7):12800

Salmon PM, Read GJ, Stevens N, Walker GH, Beanland V, McClure R, Hughes B, Johnston IR, Stanton NA (2019) Using the abstraction hierarchy to identify how the purpose and structure of road transport systems contributes to road trauma. Transp Res Interdiscip Persp 3:100067

Schröder T, Hoey J, Rogers KB (2017) Modeling dynamic identities and uncertainty in social interactions: Bayesian affect control theory. Am Sociol Rev

Schweitzer F, Garcia D (2010) An agent-based model of collective emotions in online communities. Eur Phys J B 77(4):533–545

Selbst AD, Boyd D, Friedler S, Venkatasubramanian S, Vertesi J (2018) Fairness and abstraction in sociotechnical systems. SSRN Scholarly Paper ID 3265913, Social Science Research Network, Rochester, NY (August)

Shpitser I, Pearl J (2008) Complete identification methods for the causal hierarchy. J Mach Learn Res 9(64):1941–1979

Stauffer D, Solomon S (2007) Ising, Schelling and self-organising segregation. Eur Phys JB-Condens Matter Complex Syst 57(4):473–479

St-Maurice JD, Burns CM (2017) Modeling patient treatment with medical records: an abstraction hierarchy to understand user competencies and needs. JMIR Hum Factors 4(3):6857

Suresh H, Guttag JV (2019) A framework for understanding unintended consequences of machine learning. Preprint arXiv:1901.10002

Tajfel H, Billig MG, Bundy RP, Flament C (1971) Social categorization and intergroup behaviour. Eur J Soc Psychol 1(2):149–178

Torres L, Blevins AS, Bassett D, Eliassi-Rad T (2021) The why, how, and when of representations for complex systems. SIAM Rev 63(3):435–485

Van Dijck J (2013) The culture of connectivity: a critical history of social media. Oxford University Press, Oxford

Vicente KJ (1999) Cognitive work analysis: toward safe, productive, and healthy computer-based work. CRC Press, Boca Raton

Vicente KJ, Rasmussen J (1992) Ecological interface design: theoretical foundations. IEEE Trans Syst Man Cybern 22(4):589–606. https://doi.org/10.1109/21.156574

Wachs J, Hannák A, Vörös A, Daróczy B (2017) Why do men get more attention? exploring factors behind success in an online design community. In: Proceedings of the international AAAI conference on web and social media, vol 11

Wang Y, Blei D (2021) A proxy variable view of shared confounding. In: International conference on machine learning. PMLR, pp 10697–10707

Wilson JR (2014) Fundamentals of systems ergonomics/human factors. Appl Ergon 45(1):5–13

Wong WB, Sallis PJ, O'Hare D (1998) The ecological approach to interface design: Applying the abstraction hierarchy to intentional domains. In: Proceedings 1998 Australasian computer human interaction conference. OzCHI'98. IEEE, pp 144–151

Wurst C, Chen H-YW, Joseph K (2021) Formative modeling of foster care work: A cognitive work analysis approach. In: Proceedings of the human factors and ergonomics society annual meeting. SAGE Publications Sage CA: Los Angeles, CA, vol 65, pp 933–937

Wu S, Sun F, Zhang W, Xie X, Cui B (2020) Graph neural networks in recommender systems: a survey. ACM Comput Surv (CSUR)

Xi Y, Chen A, Zhang W (2021) The expression of cultural identities in hong kong's anti-extradition law amendment bill movement: a semantic network analysis of tweets. Soc Sci Comput Rev:08944393211012267

Zhao J, Wang T, Yatskar M, Ordonez V, Chang K-W (2017) Men also like shopping: reducing gender bias amplification using corpus-level constraints. arXiv:1707.09457 [cs, stat]

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.