Coupled infectious disease and behavior dynamics. A framework to describe, compare, and select model assumptions

Andreas Reitenbach[1,6], Fabio Sartori[1,6], Sven Banisch[1], Anastasia Golovin[4], André Calero Valdez[2], Mirjam Kretzschmar[3], Viola Priesemann[4,5], and Michael Mäs[1]

[1]Chair of Sociology and Computational Social Science, Karlsruhe Institute of Technology, Karlsruhe
[2]Human-Computer Interaction and Usable Safety Engineering, Universität zu Lübeck, Lübeck
[3]Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht
[4]Max-Planck-Institute for Dynamics and Self-Organization, Göttingen
[5]Georg-August-University, Göttingen
[6]Authors contributed equally

November 2023

Abstract

To comprehend the dynamics of infectious disease transmission, it is imperative to incorporate human protective behavior into models of disease spreading. While models exist for both infectious disease and behavior dynamics independently, the integration of these aspects has yet to yield a cohesive body of literature. Such an integration is crucial for gaining insights into phenomena like the rise of infodemics, the polarization of opinions regarding vaccines, and the dissemination of conspiracy theories during a pandemic. We make a threefold contribution. First, we introduce a framework to *describe* models coupling infectious disease and behavior dynamics, delineating four distinct update functions. Reviewing existing literature, we highlight a substantial diversity in the implementation of each update function. This variation, coupled with a dearth of model comparisons, renders the literature hardly informative for researchers seeking to develop models tailored to specific populations, infectious diseases, and forms of protection. Second, we advocate an approach to *comparing* models' assumptions about human behavior, the model aspect characterized by the strongest disagreement. Rather than representing the psychological complexity of decision-making, we show that "influence-response functions" allow one to identify which model differences generate different disease dynamics and which do not, guiding both model development and empirical research in testing model assumptions. Third, we propose recommendations for future modeling endeavors and empirical research aimed at *selecting* models of coupled infectious disease and behavior dynamics. We underscore the importance of incorporating empirical approaches from the social sciences to propel the literature forward.

1 Introduction

Human behavior can have a decisive impact on a pandemic. Take, for example, the COVID-19 pandemic, where adopting protective measures such as mask-wearing, self-testing, practicing social

distancing, and getting vaccinated emerged as potent strategies to curb disease transmission between individuals. The effectiveness of each mitigation measure, however, hinges significantly on the manner in which individuals implement them. Moreover, in models not only the mean mitigation strength but also the variation of behavior across individuals in a population determines how effectively protective behavior translates into collective mitigation of a pandemic (Fig. 1). Notably, empirical research reveals that the distribution of individuals' adherence to protective measures fluctuates over time and across different countries (Fig. 1). Consequently, models of infectious disease spread must not only account for the mean strength of mitigation but also capture the dynamic variations in behavior across individuals. This nuanced understanding is crucial for these models to serve as reliable tools in anticipating the dynamics of future pandemics [57].

The state of the research on coupled infectious disease-behavior models is a double-edged sword. On the positive side, over two decades of modeling work have yielded a diverse array of alternative models, unveiling previously unnoticed connections between protective behavior and the spread of infectious diseases. Conversely, the literature is marked by an extensive, disorganized collection of alternative models, lacking a clear delineation of disparities in model predictions and the responsible underlying assumptions. Consequently, even modelers who carefully consult the literature are left with very little guidance on what assumptions their model should be based on, what aspects require empirical validation, and what classes of emergent dynamics to expect. While a diversity of models is a great asset, progress in the field is limited when the literature only accumulates models, but does not condense insights.

This article contributes in three ways to overcoming this problematic state of the literature. First, we provide a general and simple framework to *describe* agent-based models of coupled infectious disease-behavior dynamics, extending the framework proposed by [31, 96, 130, 144]. According to our framework, coupled models consist of four basic update functions. The specific choice of these four update functions allows one to categorize and to compare competing models based on the choice of each update function. This allows to structure the existing literature, which is characterized by a huge heterogeneity, in particular in terms of the assumptions about how individuals decide whether to engage in protective behavior or not. In fact, some models are based on fundamentally different theories of individual behavior. While some models, for instance, assume perfectly rational decision-makers anticipating and comparing the costs and benefits of protective behavior, other models presume that individuals always myopically copy the dominant behavior in their social network.

Second, our objective is to bring the literature closer to a systematic *comparison* of theories of individuals' protective behavior. To this end, we endorse an approach rooted in the social sciences. We argue that the often highly complicated assumptions about individual decision-making should be reduced to so-called "influence-response functions" (IRFs), i.e., functions that quantitatively describe how behavior of others and the state of the disease dynamic translates into behavior of an individual [92, 97]. Unlike approaches that treat the psychological complexity of individual decisions as the "gold standard" [145], IRFs abstract from the actual decision-making process underlying behavior and capture only the source of emergent complexity, the interaction between the microscopic entities. To highlight the effectiveness of IRFs, we illustrate that certain models—despite being grounded in fundamentally different theories of human behavior—imply similar classes of IRFs, thereby exerting analogous impacts on disease dynamics. Conversely, behavior theories that may appear similar can imply markedly different IRFs. These IRFs can be incorporated into models of coupled infectious disease and behavior dynamics to identify conditions under which they generate distinct infectious disease dynamics. Introducing IRFs allows us to approximate the complexity of human behavior in a given setting with a versatile quantitative functional dependency.

Thirdly, we outline recommendations for the *selection* of models of coupled infectious disease and behavior dynamics. We argue that a systematic comparison of IRFs will guide the empirical research needed to select the appropriate theory of protective behavior for the given context of a modeling project. While we do criticize that there is an overabundance of alternative models, we do

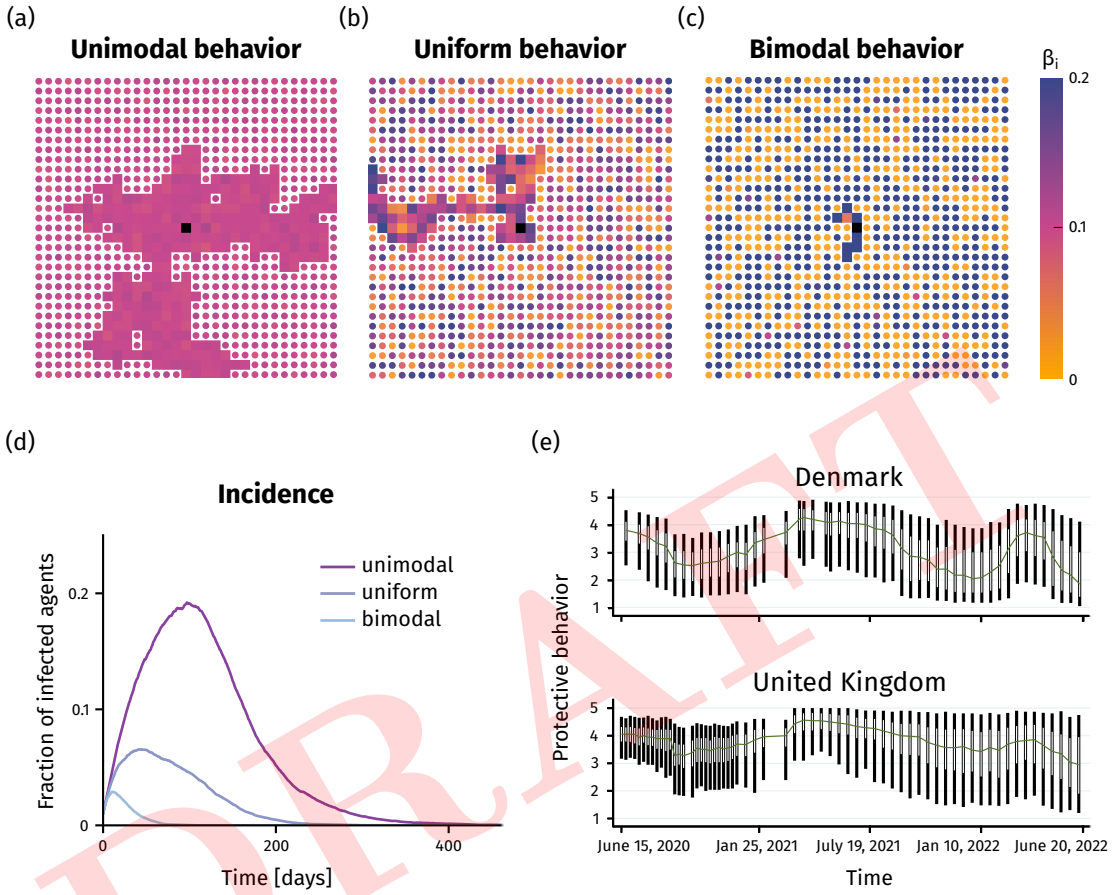


Figure (1) **Heterogeneity of individual contact behavior impacts disease spread.** (a)-(c): Representative realizations of agent-based SIR-models with the same average contact- or transmission-rate $\beta \sim 0.1$, but different variance $\mathrm{Var}(\beta)$. Nodes that were infected are shown as squares, color depicts their individual transmission rate β_i. (a) all agents have the same contact rate, i.e., $\beta_i = \beta$, (b) contact rates present a uniform distribution with $0 \leq \beta \leq 0.2$, (c) contact rates are polarized, with half of the agents having a minimal contact rate $\beta \sim 0$, the other half having no self-protecting behavior, i.e., $\beta \sim 0.2$. Details are shown in the SI in Section 1.1. (d): Development of fraction of infected agents averaged over 1,000 independent realizations, to show that (a), (b), and (c) are indeed representative. (e): During the COVID-19 pandemic, protective behavior differed across countries and time (data: [78]). For example, Denmark showed distinct collective shifts, whereas UK showed growing behavior variance. Green lines indicate the median, white bars the interquartile range. For details, see SI Section 1.2.

not propose that researchers should refrain from proposing new models. Instead, we propose a set of guidelines for the presentation and analysis of models, aiming to significantly expedite scientific progress in the realm of coupled infectious disease-behavior dynamics.

With carving complex human behavior into tractable functional forms, we see a great potential to further bridge the gap between quantitative sociology and physics of complex systems. This overview article is meant to serve as a guidance both for physicists exploring societal dynamics, as well as for sociologists with a strong interest in emergent network dynamics.

This overview article complements existing reviews that had a stronger focus on summarizing the predictions and insights derived from models of coupled infectious disease and behavior dynamics [20, 57, 139, 144, 145]. In our approach, we specifically concentrate on reviewing model inputs rather than their outputs. This emphasis arises from the recognition that a lack of clarity regarding differences in model assumptions raises questions about the consistency and potential contradiction of predictions derived from different models. Consequently, our review meticulously documents the diversity of model assumption and proposes a strategic approach to advance the current literature.

2 *Describing* coupled infectious disease-behavior models

In this section, we propose a framework to describe and categorize existing models of coupled infectious disease and behavior dynamics, building on previously introduced approaches [31, 96, 130, 144]. Specifically, we show that models can be described by their four update functions: the functions describing the disease model, the influence model, and two bridge functions coupling behavior to disease dynamics and vice versa. We show that existing models can be categorized according to which update functions are included and how each function is implemented.

2.1 Framework update functions

Formal agent-based models have extensively been studied independently for disease spread [41, 70, 80, 84, 89] and for social influence [29, 50, 52]. While both model families address interactions between individuals, the nature of these interactions varies significantly. Disease transmission is inherently nonreciprocal, in that an infected person can transmit a pathogen to a susceptible contact, but the susceptible person cannot "heal" the infected one. In contrast, social influence is often reciprocal, in that individuals can mutually exert influence on each other. Therefore, the emergent dynamics and state space generated by these two model families can differ fundamentally.

We focus here on so-called agent-based or micro-level models, a category of models which coexist alongside macro-level and meso-level models. Macro-level, compartmental models represent only collective outcomes like the *fraction* of individuals in available disease states, such as the proportion of susceptible S, infected I, and recovered R agents in the classical SIR model [84]. In this model, infection dynamics is determined by the transition rates between the compartments: the contact or transmission rate β to get infected, the rate γ to recover, and the rate ν of waning immunity. To incorporate behavior, these models can in principle include compartments and bridge functions that capture how strongly individuals engage in protective behavior on average, or introduce compartments for people who protect with different mitigation strengths [18, 34, 37, 44, 60, 140]. In contrast, micro-level models formally represent every individual member of the population and its individual health state, opinion, and protection preference, making it possible to investigate the impact of network structures on dynamics. Meso-level models study subpopulations such as local clusters in the network. Finally, there are hybrid models that combine both macro and micro-level dynamics, with behavior typically being modeled at the micro-level and disease spread at the macro-level. [149].

Overall, the distinction between compartmental (macro-) and agent-based (micro-) models is not clear-cut. They rather present a continuous spectrum: One can increase the number of compartments, to, e.g., represent increasingly more combinations of preferences, opinions, disease, or

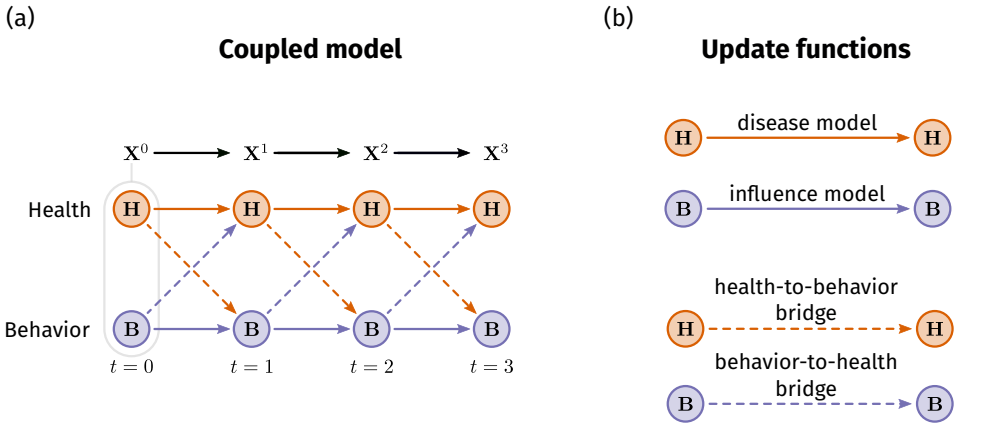


Figure (2) **General framework of coupled infectious disease and behavior dynamics.** (a) **H** and **B** are vectors presenting the agents' health and behavior states, respectively, and $\mathbf{X^t}$ presents the complete state of the system at time step t. Arrows represent the four update functions of coupled models of disease and behavior dynamics, as listed in panel (b).

immunity states. However, this leads to a rapid rise in the number of compartments due to combinatorial explosion, potentially resulting in compartmental models having as many compartments as there would be agents in an equivalent micro-model. The distinctive strength of agent-based models lies in the explicit representation of each agent and its contact structures. Conversely, compartmental models offer an advantage in terms of differentiability: each compartment represents a continuous-valued fraction of the population. This characteristic proves advantageous, for instance, when inferring model parameters from data through Bayes inference [105].

Formally, the dynamics of a population of N individuals i can be described as a N-dimensional random process $\mathbf{X}$. It is convenient to describe the random process as the combination of two interacting contributions (Fig. 2a): the health state $\mathbf{H}$ and the behavioral state $\mathbf{B}$. Here, H_i^t and B_i^t present the health and behavioral status of the agent i at the time-step t. Each of the variables H_i^t and B_i^t can adopt values from the sample spaces Ω_H and Ω_B, respectively, which can be chosen flexibly. For example, in a model in which agents make a binary choice between two behavioral options, as in the decision whether to get vaccinated or not, Ω_B may be equal to $\{0, 1\}$. In contrast, if the behavior is a continuous quantity, for example, the degree to which agents reduce contacts, the sample space presents an interval $\Omega_B = [0, 1]$. In the case of disease models, the sample space of health states is usually discrete, e.g., $\Omega_H = \{\mathrm{S}, \mathrm{I}, \mathrm{R}\}$ for classical SIR dynamics.

We let $\Omega_X = \Omega_H^N \times \Omega_B^N$ denote the whole sample space of the coupled infectious disease-behavior system, with $\mathbf{X}$ denoting a certain state in this space

$$\mathbf{X}^t := \begin{pmatrix} \mathbf{H}^t \\ \mathbf{B}^t \end{pmatrix}. \tag{1}$$

In most agent-based models, the state of every agent at a certain point in time only depends on the states of the agents in the previous time-step [12, 76]. Therefore, we can express the dynamics of the system as a Markov chain:

$$\cdots \to \mathbf{X}^t \to \mathbf{X}^{t+1} \to \ldots \tag{2}$$

The transition probability from $\mathbf{X}^t$ to $\mathbf{X}^{t+1}$ can be broken down into four update functions, as illustrated in Fig. 2b. In the most simple case, the updated health state $\mathbf{H}^{t+1}$ depends on the previous health state $\mathbf{H}^t$; similarly, the updated behavior $\mathbf{B}^{t+1}$ typically depends on $\mathbf{B}^t$. We call the set of modeling assumptions which govern those transitions the *disease model* and the *influence*

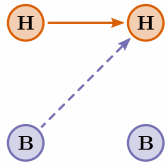
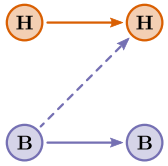
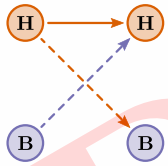
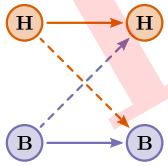
Illustration	Description
	Classical disease models *without protective behavior* [11, 23, 42, 47, 110].
	Disease dynamics with *constant protective behavior* [24, 27, 28, 68, 79, 119].
	Disease dynamics with protective behavior. Agents exert *influence* on each others' behavior [35, 72].
	Disease dynamics with protective behavior but without social influence. Behavior affects disease states, and the state of the pandemic changes behavior [7, 33, 104, 112, 113, 143, 148].
	Fully coupled infectious disease and behavior dynamics [45, 55, 58, 64, 73, 75, 95, 96, 106, 114, 124, 127, 137, 141, 142, 149, 151].

Table (1) **Categories of coupled disease-behavior models.** The categorization depends on which of the four update functions from Fig. 2b are included. We only show those categories that are relevant for the study of disease dynamics.

model, respectively. The disease model determines how a pathogen can move from one agent to another. Central assumptions of this update function, for instance, determine who can infect whom, and how quickly agents recover from an infection. Similarly, the influence model describes how the protective behavior of an agent is influenced by other actors such as other agents, media, or officials.

In a coupled model, both processes may also influence each other. A model includes a *health-to-behavior bridge* if $\mathbf{B}^{t+1}$ also depends on $\mathbf{H}^t$. This update function captures how agents adjust their behavior to the local or global dynamics of the disease, including, for instance, the assumption that agents get vaccinated when the fraction of infected agents in the population exceeds a given threshold. Finally, if $\mathbf{H}^{t+1}$ depends on $\mathbf{B}^t$, the model includes a *behavior-to-health bridge* which specifies how individual behavior affects the processes of disease transmission or recovery from an infection. For instance, a very prominent assumption is that agents who decided to get vaccinated have a decreased probability of catching a disease after a contact with an infected agent.

We can categorize existing models of coupled infectious disease and behavior dynamics according to what update functions are included in a given model. In total, there are $2^4 = 16$ possible tuples of update functions; however, only few of those are relevant for the present review. First, any model of an infectious disease has to include a disease model; otherwise, the agents would not be able to transmit the disease. Second, coupled models without a behavior-to-health bridge are not relevant, since the literature typically seeks to predict the course of an epidemic influenced by protective behavior. In total, this leaves four categories of coupled behavior-disease models, which are shown in Table 1 together with a list of typical implementations from the literature. For example, the second row of the table represents models in which disease dynamics is influenced by protective behavior, but the behavior remains constant since the agents do not influence each other [24, 68, 79, 119]. The model we used to generate Fig. 1 falls into this category.

We acknowledge that modelers may hold differing views on whether specific behaviors should be classified as a health state $\mathbf{H}^t$ or a behavioral variable $\mathbf{B}^t$. Consider the case of vaccination: On the one hand, opting for vaccination clearly constitutes a behavioral choice. Moreover, individuals may engage in discussions regarding vaccination decisions and exert social influence on each other, a process that would be formally incorporated into the influence model (Fig. 2b). On the other hand, within a macro-model with a compartment for vaccinated individuals, vaccination status would be regarded as a health state. Additionally, individuals might base their vaccination decisions on factors such as the risk of infection, which hinges on the vaccination rate within the population. This distinction can be explicitly captured in the health-to-behavior bridge rather than the influence model, without compromising the model predictions. The choice between these interpretations often relies on personal preferences and the disciplinary background of the modelers. However, there are instances in which the distinction becomes crucial. For example, an agent might choose to receive a vaccine but, due to scarcity or governmental restrictions, fail to undergo the vaccination process [115]. Consequently, the social influence exerted by this agent on others may deviate from the agent's impact on the population's health state. In such cases, it becomes imperative to carefully differentiate between the decision to get vaccinated and the health state of being vaccinated or not.

2.2 Updating schedule

The predictions of many complex-system models depend on the exact ordering of updates [32, 59, 74, 122, 129]. Updating schemes can be broadly categorized along two axes. First, one can distinguish between updating a single agent chosen uniformly at random during a time-step, some subset of agents, or all agents. Second, if multiple agents are updated, their states can change either synchronously or asynchronously. In the first case, the agents are updated in parallel and cannot see the updated states of other agents, while in the second, the agents are updated in a sequence and can react to changes in other agents that have already been updated. If only one agent is updated during a time-step, synchronous and asynchronous schemes are identical.

Updating only one agent at a time implies that only a single element of the health-state vector **H** and the behavior vector **B** associated with the chosen agent changes from one time-step to the other. This allows for analytical solutions using Markov chain tools [15] as well as mean-field-like continuous-time formulations in the analysis of the model [62]. When updating all agents synchronously, **H** and **B** may undergo abrupt changes during a single time step, and complex transient patterns such as those known from cellular automata [147] may emerge. In the context of opinion dynamics, it has been shown that parallel update may lead to frustration, making convergence more difficult [13, 128].

Another facet of the updating schedule pertains to the sequence and frequency of health and behavior updates. This decision depends on the specific kind of behavior and disease under consideration, as both processes can unfold over significantly different timescales. An example of such separation of timescales is provided by models in which agents are faced with the decision whether to vaccinate against influenza [18, 91, 136]. This decision is made every year before the flu season starts and is based on the agent's experiences during the previous influenza wave. In those models, behavior dynamics plays out on a characteristic timescale of years, while the course of an infection takes only a few days or weeks. On the other hand, the decision whether to wear a mask or not can be reevaluated multiple times every day, resulting in similar time scales as in disease models. Finally, in scenarios like HIV modeling, disease progression occurs over a slow timescale while behavior operates on a faster one.

Empirical information can help determine exact time scales of the disease and behavior dynamics. For many diseases and populations, for instance, empirical research provides estimates of the basic reproduction number R_0, the expected number of cases directly generated by one case. Knowing R_0 of the context under investigation, the average number of network contacts an agent can infect in a model time step, and the length of the time span an individual is infectious, one can determine the physical meaning of a time-step.

2.3 Representing social networks

The framework does not pose restrictions on the assumed structure of the social network underlying the two processes. In many contributions, a single layer network is employed, assuming that disease transmission and social influence are acting on the same network [35, 95, 104, 106, 114, 153]. Other modeling work assumes that social influence and infection dynamics play out on a multiplex network: a network consisting of two layers that share the set of nodes but may have different sets of edges [3, 58, 151]. This implements, for instance, that individuals exert influence on each others' behavior in a setting such as online communication where the transmission of infection is not possible. The structure of transmission related networks and social networks may be very different. However, networks of disease transmission and networks of behavior influence are also not unrelated. Aksoy, for instance, documented significant social influence on social-distancing behavior within families during the COVID-pandemic in the UK [2]. Families, in other words, are not only a stetting where diseases spread, they are also a source of influence on protective behavior. What is more, Fukuda and colleagues demonstrated that model predictions about disease dynamics can depend on the degree to which the social-influence network and the disease-transmission network overlap, providing a strong argument to carefully consider how networks are modeled [54].

For any single layer in the network, different ways of generating the graph are possible. Popular choices are fully connected networks [114], lattices [153], scale-free networks [54], small-world networks [27, 119], real networks [4, 104, 124, 150], and pseudo-empirical networks [35, 88]. Recently, empirical data on multiplex networks has become available [120]. Within networks, edges between nodes may carry weights to represent closeness of contacts in the disease networks or the strength of influence in the behavior network [150, 151].

The structure of the social-influence network also depends on the behavior under study. Some

forms of behavior can be observed by others during face-to-face interaction. For instance, it is easy to see whether an interaction partner is wearing a mask. However, one cannot directly observe whether the person has been vaccinated or has recently conducted a self-test, which implies that social influence by one's local social network is more restricted in this case. Strikingly, the opposite can be true for collective behavioral patterns. In many countries, officials [45] and the media [63] regularly disclose the population's global vaccination rate and the number of conducted tests per day. On the other hand, it is much less common to assess and disclose the rate of mask wearing in a population. Whether agents are influenced by local or global information, in turn, has important consequences [4], since local information supports the formation of internally homogeneous but mutually distinct clusters in a network [16]. When all agents are affected by the same global information, in contrast, local clustering is less likely [99, 124]. Providing individuals with global information has also been used as an intervention strategy to increase compliance with social norms [99, 116, 123]. While many existing models abstract away from global influence [3, 106, 114, 149, 150, 153], some models explicitly include it [19, 53, 106, 117].

2.4 Existing disease models

Each of the four update functions identified in Fig. 2b has been implemented in various ways. The number of different disease models, however, is relatively low. There are disease models that describe agents by only two possible health states, assuming that agents are either susceptible or infected [104, 110, 148, 151]. Other models add more possible states, including that agents can also be described by a recovered or exposed state [3, 68, 149, 153], or represent vaccinated agents by a dedicated compartment [114, 142]. The choice of the infection model appears to be mainly dependent on the specific disease studied by modelers.

2.5 Existing behavior-to-health bridges

Contributions to the literature make diverse assumptions about the behavior-to-health bridge, the update function determining how agents' behavior affects its health status. This diversity mainly results from the different forms of protective behavior represented in the models. A first category of behavior-to-health bridges implements that an agent who has decided to protect will experience a reduced susceptibility for the remainder of the modeled time frame [3]. This assumption is particularly applicable to diseases where vaccines confer long-term immunity. Alternatively, various models include the assumption that agent's behavior has a non-permanent effect on disease dynamics [106, 114, 153]. This can have alternative interpretations. First, some authors treat one season as a single time-step in which individuals make the decision whether to protect or not, e.g., by getting vaccinated [35, 53, 54]. Second, individuals may change their decision to wear a mask multiple times every day. In this case, however, the interpretation of the duration of a modeled time-step changes compared to the first interpretation, in that a time-step now corresponds to a single encounter between an agent and its social contacts.

Another category of behavior-to-health bridges are models assuming adaptive networks. An adaptive network is a network whose topology changes in time dynamically. The majority of these models assume that susceptible individuals will try to avoid contact with individuals they deem infected, either by rewiring their network connection from an infected individual to a healthy one [67], a random one [118, 152], or by temporarily interrupting contacts with an infected connected individual [135]. In [108], a more psychologically realistic wiring function is considered, which weighs the risk of maintaining a connection to an infected neighbor against the social benefit of this connection to the well-being of the agent. Another class of models that relies on adaptive networks are social distancing models, where a fraction of all the connections between agents is removed as a response to the disease prevalence. The reduction in contacts can be interpreted either as social distancing or as movement restrictions [102, 138].

A third category of behavior-to-health bridges assumes that agents' behavior affects infectiousness instead of, or in addition to, their susceptibility [56]. This might represent that agents get tested and self-isolate if the result is positive [65, 87, 107]. Likewise, taking antiviral drugs, or wearing a mask reduces the probability of an infected agent to infect others [96].

An alternative behavior-to-health bridge that has received limited attention includes that protective behavior can speed up recovery after infection [56, 125]. Agents may be recovering at a higher rate because they better recognize symptoms and seek medical treatment earlier.

While the vast majority of existing models included only a single kind of behavior, there are also models that assume that agents choose between different types of behavior and that each behavior implies a different disease bridge. This is not only a realistic assumption in many contexts, it can also generate interesting disease dynamics. For example, Zhang and colleagues [153] show that increasing the effectiveness of protective behavior can backfire. On the one hand, increased effectiveness of mask-wearing can convince individuals to invest into this behavior rather than staying unprotected. On the other hand, it might also demotivate investment into even more protection (e.g. vaccination), which comes at higher costs. This can lead to higher incidence of infection.

2.6 Existing influence models

Existing influence models can be grouped into two main categories. First, there are what we denote "*behavior models*" [53, 54, 106, 150, 153]. These models describe agents only in terms of their behavior, capturing, for instance, whether a person is wearing a mask, stays at home, or gets vaccinated. Next, modelers assume that agents observe this behavior in others, and are socially influenced by these observations. Second, there are "*internal-state models*" that explicitly represent the cognitive determinants of behavior, such as opinions, beliefs, or feelings of fear [3, 114]. These internal states, rather than the actual behavior are subject to social influence. Thus, agent behavior is changing, because the cognitive determinants are adjusted.

An often critical difference between internal states and behavior is that internal states are often more nuanced than actual behavior. That is, while individuals' opinions towards protective behavior can vary on a scale from negative to positive evaluations [3, 114], behavior is often discrete in that individuals either engage in a given behavior or not. A person, for instance, may wear a mask while it is actually not totally convinced of the protective effect. What others observe and potentially condition their own behavior on, however, is another person wearing a mask. This effect has been shown to foster processes of collective opinion extremization and opinion polarization with strong clustering in the network [16, 50, 99].

There are various reasons for why modelers may decide to not represent internal states and model only behavior. First, the decision to not incorporate internal states may result from the desire to develop simple representations of human behavior, which is understandable given that models combining influence and disease dynamics are highly complicated. Second, there are situations where behavior is socially influenced without changes in underlying internal states. If a person with a negative opinion towards mask wearing is working together with others who always wear a mask, the person will likely conform and also start wearing a mask in this context even when its opinions have remained unchanged [5]. Social influence, in this case, does not mean that individuals exert influence on each other's motives to behave in a certain way but that individuals influence the social context in which others chose a behavioral response to the behavior of others. This can have repercussions on disease dynamics, because this person may not wear a mask in context where others are not wearing a mask.

2.6.1 Behavior models

Existing behavior models can be further categorized into two subcategories: models of *rational decision-making* and models assuming *success-driven imitation*. Being inspired by game-theory,

models based on *rational decision-making* typically rely on very strong assumptions about rationality, implementing that agents' behavior is perfectly consistent with their preferences and perceived restrictions [19, 30, 35, 73, 75, 104, 117, 124]. Bauch and Earn [19], like many other authors, studied a model where agents compare the expected costs of getting vaccinated with the expected costs of becoming infected and chose to vaccinate or not depending on which of the two expectations is higher [104, 150]. Importantly, it is added that the infection probability agents perceive depends on the share of vaccinated agents in the neighborhood, which implements social influence, in that agents respond to the behavior in their social environment. When many contacts of an agent are vaccinated, the added benefit of also getting vaccinated are low, which implies that the agent will not get vaccinated and free-ride on the herd immunity created by others. Likewise, when very few contacts of an agent are vaccinated, chances of infection are high and agents will choose to get the vaccine. There are also game-theoretical models that deviate from the assumptions that perceived infection risks are a function of others' decisions to protect. Huang et al., for instance, included that perceived infection risks depend on disease prevalence [73].

The second subcategory of behavior models assumes the same determinants (preferences and beliefs) of behavior but a very different decision-making process [18, 37, 53–55, 75, 91, 104, 143, 153]. Models of *success-driven imitation* are also inspired by the game-theoretical literature, but drop the assumption of perfect rationality and rather assume a myopic decision-making process [69]. In a nutshell, models of success-driven imitation presume that agents observe the behavior and the past payoffs of their network contacts. Next, it is assumed that agents tend to adopt the behavior of neighbors who experienced higher payoffs in the past, imitating those who have displayed successful behavior in the past. Since others' past success depends in these models on whether or not they experienced costs of protection or costs of infection, agents decide to vaccinate when many network contacts who did not protect got infected and when contacts who did protect did not experience the relatively high costs of infection.

A slight variation of success-driven imitation has been proposed by [104]. In this model, agents imitate the behavior of a network contact, when they expect that the neighbor's past behavior will be more rewarding than their own past behavior. That is, agents do not compare past payoffs but expected future payoffs when they decide whether to imitate or not.

2.6.2 Internal-state models

Models belonging to the second main category of influence models explicitly represent the *internal states* underlying agent behavior. Unlike in the behavior models, agents change their behavior, because these internal-states have changed. We have identified four subcategories in the literature: awareness models, assimilation models, reinforcement models, and repulsion models.

In representatives of the first subcategory, *awareness models*, social influence is implemented like a spreading process [46, 58, 63, 64, 95, 127, 137, 142, 149, 151] similar to the spreading process of the disease. That is, every agent is described by a variable quantifying the agent's awareness of the disease and communicates it to other agents. The concept of awareness has also been described as a feeling of fear of infection. In models such as [63], the awareness is a binary state, where an unaware agent can become aware depending on the fraction of aware and infected neighbors. In other models, awareness is measured on a scale representing the shortest distance to an index case [58] or the number of times an agent got notice of a severe case [149]. Often in these models, the quality of awareness decreases upon spread to another agent. Unlike in all other influence models we summarize here, this form of social influence is nonreciprocal in that an Agent A can make an Agent B aware of an infection case. B, however, cannot make A unaware of this event. Furthermore, awareness is also often assumed to fade over time. Finally, models assume awareness translates into protective behavior.

Awareness models have been extended to also capture awareness of a protective behavior's side-effects such as vaccine side-effects [33, 149], which adds another spreading process. The more aware

agents are of side-effects, the less likely they will protect, according to these models.

The second subcategory are *assimilation models* [4, 27, 28, 45, 95, 96, 104, 106, 143, 150], a model family that shares key assumptions with the most classical models of opinion dynamics [1, 50–52]. In these models, agents are described by a continuous opinion value describing their attitude towards protection. Agents' protective behavior is a function of this opinion. Next, it is assumed that agents' opinions are open to social influence typically implemented as averaging. That is, agents tend to adjust their opinions in a way that they grow more similar to their network contacts. A seminal implication of this assimilative form of influence is that populations characterized by a connected network will eventually reach a state of opinion consensus [1, 40, 50].

Reinforcement models are very similar to assimilation models but they add a critical aspect concerning the opinion update [3, 114, 141]. In contrast to the assumption of averaging, opinion reinforcement implies that agents adopt more extreme views when a contact agrees with them. This can result from the communication of persuasive arguments [25, 103]. When two interaction partners hold positive opinions towards mask wearing but base this position on different arguments (e.g. "masks protect myself" and "masks protect my contacts"), than communication of these arguments will provide both persons with new reasons to be positive towards mask wearing. Reinforcement may also result from social approval [71], for instance, when two individuals mutually support their decision to wear a mask. When coupled with homophily, the tendency to interact with agents holding similar views [100], opinion reinforcement can generate opinion polarization [16, 98]. Homophily is a very strong and well-documented force in human behavior and is reinforced by personalization algorithms installed on online social networks [9, 77, 81, 109].

The fourth subcategory of internal-state models are called *repulsion models* [125] and also include an assumption that has received much attention in the literature on social-influence models [49, 94]. Repulsion is the counterpart of assimilation and includes that individuals may dislike sources of influence who hold different views or who they distrust and may, as a consequence, adopt opinions and behavior that increases distances to these sources. While empirical research testing this assumption is not conclusive, recent findings from research on online social networks documented repulsion in online communication settings [8, 83, 90, 132]. Moreover, motivated cognition and confirmation biases [93, 131], where individuals downgrade new information that challenges their current opinion, may also result in repulsive interaction [17, 39]. Repulsion received much attention, because it is also able to generate opinion polarization. However, unlike opinion reinforcement, it fosters opinion polarization when homophily is weak and individuals frequently interact with others who disagree [81, 98]. As a consequence, interaction activating repulsion may be actually rare in personalized communication settings like online social networks.

2.6.3 The context of social influence

Independent of what model of influence is implemented, there are contextual factors that can have impact on the dynamics of behavior. For instance, the choice of the protective behavior represented in the formal model has important repercussions for the influence model, since humans can observe others' behavior or its determinants only in specific social contexts. Mask wearing, for instance, can be observed in one's *local* network but there is limited information about the *global* distribution of the behavior on a population level. Behavior like vaccination and self-testing cannot be observed locally but in many contexts governments assess and disclose global information. During the COVID-19 pandemic, for instance, many governmental agencies regularly published the population's global vaccination rate and the daily number of conducted tests. It is much less common to assess and disclose the rate of mask wearing in a population. Modeling work and empirical research shows that this can have decisive impact on behavior and disease dynamics, since local social influence promotes the clustering of behavior in networks [14, 99, 116, 123, 124].

Likewise, the predictions of influence models can strongly depend on whether agents communicate in a one-to-one, a one-to-many, or a many-to-one regime. Many influence models assume one-to-

one influence, where in an influence event one agent causes an opinion change in a single other agent [6]. It turns out, that opinion polarization is more likely when models assume one-to-many communication, a regime that is typical for many online social networks, where one user emits a messages to all of its followers at the same time [82]. Influences dynamics are also altered when models include many-to-one communication where agents updating their opinion do not consider the input of a single interaction partner, but their whole ego network [48].

2.7 Existing health-to-behavior bridges

The health-to-behavior bridge specifies how the vector of health states in the population affects the behavior update of the agents. In particular, this implements that agents adjust their protective behavior to disease prevalence. However, as sketched above, it can also be useful to treat for instance the share of vaccinated individuals as a health state, which means that also behavioral reactions to the overall protection state of a population, such as the vaccination rate, can be treated as a health-to-behavior bridge. To avoid repetition, we refer to Section 2.6.1 for a summary of assumptions about how agents respond to others' protective behavior.

In existing models, there are three main categories of health-to-behavior bridges. First, is a considerable number of models that do not include any health-to-behavior bridge [27, 28, 35, 72]. In these models agents do not react to the disease dynamic. Second, there are various models assuming that agents tend to protect more when disease prevalence high, which is sometimes referred to as "prevalence-elastic behavior" [4, 45, 53–55, 58, 63, 64, 73, 75, 91, 95, 96, 114, 124, 127, 141, 143, 149, 151, 153].

This health-to-behavior bridge is often not included separately, but integrated in the agent's decision making process, which we reviewed in the Section 2.6. In some game-theoretical models, for example, an agent's perceived risk of infection is a function of the share of infected agents in the population [73, 112]. Similarly, in awareness models [58, 63, 149] a higher incidence also increases the number of agents that are aware of the disease. Awareness can subsequently spread in the population, leading to more people engaging in self-protecting behavior. In success-driven imitation models [53–55, 75, 91, 143, 153], agents imitate the behavior of successful others. When prevalence is high, agents who protected themselves will have relatively high payoffs since they managed to avoid infection, unlike unprotected agents. When prevalence is low, in contrast, these agents will have a relatively low payoff, since they payed the costs of protection and experienced the same health as unprotected agents. As a consequence, prevalence affects whether protected or unprotected agents are being imitated.

A third category of health-to-behavior bridges implements that high prevalence can sometimes lead to less protective behavior [3, 33, 149]. In Alvarez et al. [3], for instance, the authors assume that vaccines reduce the personal susceptibility only to a limited degree. An agent who decides to vaccinate, but nevertheless becomes infected turns frustrated and develops an extreme negative opinion towards vaccination. Subsequently, frustrated agents will share their negative opinions with others. Since infection despite vaccination is more likely when prevalence is high, a high prevalence can lead to less protective behavior.

A second dimension where health-to-behavior bridges differ is the range of the influence. In several models agents respond to the local health-state of their network neighborhood [53, 54, 73, 106, 151, 153], in [63, 104], for example the probability of vaccinating depends on the fraction of vaccinated neighbors. In other models, agents are influenced by the global fraction of infected agents in the whole population [106, 114]. Some models also consider an interesting intermediate step, where each severe case influences the whole network, but the strength of this influence decrease exponentially with the distance from the original severe case [58, 149].

Local and global effects of disease prevalence on behavior can have different consequences [4, 148]. Since agent-based models of disease dynamics generate localized outbreaks of the disease,

local health-to-behavior bridges imply that agents located in parts of the network distant to an outbreak will not adjust their behavior and may be unprepared when the disease reaches them. At the same time, agents located inside an infected segment of the network will protect when reacting to local information, which can have substantial impact on disease dynamics. In contrast, when agents react to global dynamics, local reactions tend to be weaker, which hampers the effect of protection.

3 *Comparing* models using influence-response functions

In the preceding section, we cataloged an extensive array of competing model assumptions within the existing literature. Modelers have experimented with various alternative approaches to implementing each of the four update functions. The behavioral part, in particular, is characterized by a large and growing number of conflicting assumptions. This mirrors the ongoing discourse about human behavior and decision-making in the social sciences, a domain often criticized for its inability to converge on a shared set of behavioral assumptions [10, 38, 61, 101, 134, 146].

To be sure, we do not criticize any individual contribution to the literature, and we applaud all efforts to explore additional models or variations of existing ones. However, the collective body of literature provides limited guidance for modelers who aim to develop a valid model of coupled disease and behavior dynamics due to its extensive diversity. Furthermore, many of the behavior theories differ significantly in their fundamental assumptions. For example, game-theoretical models assume perfect rationality, while success-driven imitation assumes a myopic decision maker. Consequently, selecting appropriate model assumptions for a given application is exceedingly challenging. The existing literature falls short in delineating where behavior theories diverge and whether and when these differences result in disparate model predictions about disease dynamics.

Here, we elaborate an approach to comparing behavior models that was proposed by López-Pintado and Watts [92]: influence-response functions. In their original contribution, an influence-response function (IRF) returns an agent's probability of adopting a binary behavior given a weighted average of behaviors of other agents. That is, rather than representing the actual individual decision-making process, the psychological complexity is boiled down into a function capturing the cause of complexity, the interaction between agents.

When an IRFs accurately reflects the assumed decision process, model predictions are obviously unaffected. The central advantage of IRFs, however, is that they are a tool to compare competing behavior theories in terms of their implications for the interaction between the microscopic entities of the complex system. As López-Pintado and Watts illustrated, sometimes even seemingly different behavior theories actually imply the same IRFs and, as a consequence, generate the exact same macroscopic dynamics. Likewise, seemingly similar behavior theories can translate into very different IRFs. These differences have the potential to generate different macroscopic dynamics and, thus, help understand why different micro-theories generate different macro-predictions.

To render the concept of IRFs applicable to models of coupled disease-behavior dynamics, we expand upon the original definition presented in [92]. Subsequently, we offer various examples illustrating how existing models can be expressed using IRFs, encompassing both purely behavioral models and those entwining behavior and disease.

3.1 Definition of influence-response function

In the original work by López-Pintado and Watts [92], agents can choose between two behavior options $\Omega_B = \{0, 1\}$. The authors formally defined an influence-response function as the probability that an agent i chooses behavior 1 at time-step $t + 1$, given that the agent receives a certain social signal s_i^t from other agents

$$\mathrm{IRF}_{\mathrm{LP}}(s_i^t) \equiv P\left(B_i^{t+1} = 1 \mid s_i^t\right). \tag{3}$$

The social signal is defined as a weighted sum $f_i : \Omega_B^{N-1} \to [0,1]$ of behaviors of other agents

$$s_i = f_i\left(\mathbf{b}_{-i}^t\right) := \sum_{j \neq i} w_{ij} b_j^t, \tag{4}$$

where $\mathbf{b}_{-i}^t$ is the vector of all behaviors except that of the agent i.

In order to extend IRFs to models of coupled behavior-disease dynamics, we relax some of the assumptions made in the original work. First, we generalize the framework to accommodate non-binary behavior, since the behavior in some of the influence models we described in Sec. 2.6 depends on an internal state that is often modeled by a continuous variable. Second, we add the health-to-behavior bridge by including the health status of the population into the right hand-side of the IRF. In many such coupled models, behavior often depends on multiple observables of the system; for example, the global number of infected individuals and the local fraction of neighbors adhering to the same protective behavior as the agent. Therefore, we also allow any observable of the system to function as the social signal.

To begin, recall that the dynamics of the system, can be described as a Markov chain $\{\mathbf{X}^t\}$. Therefore, the evolution of the system can be fully desctribed using the transition probability $P\left(\mathbf{X}^{t+1} = \mathbf{x} \mid \mathbf{X}^t = \mathbf{x}^t\right)$. Depending on the updating schedule, a single transition can affect only behavior or only health, only the state of a single agent or the states of all agents at once, or any combination of the above. All these cases can be captured by the transition probability. However, this flexibility comes at a cost of an exploding number of dimensions, as the transition probability has to be specified for all possible transitions between all possible states of the system. For example, even for a simple SIS disease dynamic coupled with binary behavior, there are 2^{4N} different possible transitions. To reduce the dimensionality of the problem, we will make a series of approximations and projections which will map the transition probability into a lower-dimensional space, while still retaining as much of its flexibility as possible. For the sake of simple notation, we will assume that the sample spaces Ω_H and Ω_B are countable; if behavior is a continuous variable, the same arguments apply if the corresponding sums are replaced by integrals.

First, we will assume that the agents are updated independently of each other, i.e., that the transition probability can be re-written as

$$P(\mathbf{H}^{t+1} = \mathbf{h}, \mathbf{B}^{t+1} = \mathbf{b} \mid \mathbf{x}^t) = \prod_{i=1}^{N} P(H_i^{t+1} = h_i, B_i^{t+1} = b_i \mid \mathbf{x}^t). \tag{5}$$

This assumption is trivially fulfilled in models which only update a single agent during a time-step, and it is typically assumed in models with synchronous updating. If multiple agents are updated asynchronously, each time-step can be formally split into several substeps, such that just one agent is updated at a time.

In the next simplification step, we disentangle the update of health and behavior. We can define the marginal probability that an agent i will adhere to behavior b_i regardless of its health state:

$$P(B_i^{t+1} = b_i \mid \mathbf{x}^t) = \sum_{h \in \Omega_H} P\left(H_i^{t+1} = h, B_i^{t+1} = b_i \mid \mathbf{x}^t\right), \tag{6}$$

and, similarly, for the marginal probability of the health state

$$P(H_{t+1,i} = h_i \mid \mathbf{x}^t) = \sum_{b \in \Omega_B} P\left(H_i^{t+1} = h_i, B_i^{t+1} = b \mid \mathbf{x}^t\right). \tag{7}$$

As our primary interest lies in deriving the IRF for behavior, we will focus on Eq. 6 in the following. However, note that Eq. 6 and 7 are not statistically independent in some models. As an example, consider the model proposed by [63] which examines how awareness of a pandemic spreads within a society. In this model, agents can be in one of two behavior states: "aware" and "unaware".

Agents become aware of the pandemic either through social connections or through becoming infected, meaning that an agent cannot be infected while remaining unaware. Consequently, the new behavior and health states are correlated, meaning that we lose some of the information about the full model by considering Eq. 6 separately from Eq. 7. However, this separation allows us to arrive at a simple method for comparing and visualizing various models.

With those two steps, we have already greatly reduced the dimensionality of the problem: Eq. 6 only requires specifying the transition probability for $|\Omega_H|^N|\Omega_B|^N \times |\Omega_B|$ cases. However, even this is prohibitively challenging to visualize. To reduce the dimensionality even further, we will use the fact that the update rule rarely depends on the state of the whole system. Instead, in most models, agents make decisions based on two pieces of information: their own state and some observable of the system such as the number of infected agents or the fraction of neighbors who express the same behavior. Let $x_i^t = (h_i^t, b_i^t)$ denote the state of the agent at time-step t and let $f_i(\mathbf{x})$ be an observable of the system for the agent i. Then, we can perform a transformation of variables and express the transition probability in terms of those two pieces of information

$$P(B_i^{t+1} = b_i \mid s_i, x_i^t) = \frac{\sum_{\mathbf{x}\in\tilde{\Omega}_X} P(B_i^{t+1} = b_i \mid \mathbf{X}^t = \mathbf{x}) P(\mathbf{X}^t = \mathbf{x})}{\sum_{\mathbf{x}\in\tilde{\Omega}_X} P(\mathbf{X}^t = \mathbf{x})}, \tag{8}$$

where $\tilde{\Omega}_X$ is the subset of Ω_X where $f_i(\mathbf{x}) = s_i$, and $x_i = x_i^t$. The probability of every state $P(\mathbf{x})$ depends on how the model samples the probability space.

Finally, we define the IRF with respect to the observable f_i as the left hand-side of Eq. 8

$$\mathrm{IRF}(b_i, s_i, x_i^t) \coloneqq P(B_i^{t+1} = b_i \mid s_i, x_i^t). \tag{9}$$

As in the previous simplification step, the reduction of the update rules to a single signal may come at a loss of information. Nevertheless, the IRF as defined in Eq. 9 is always well-defined.

The IRF can be approached numerically with a Monte Carlo algorithm. The goal is to sample the probability distribution of an agent's behavior for the state of interest x_i^t and a set of signals $\{s_i\}$. The main difficulty of approaching Eq. 8 numerically is to approximate the distribution $P(\mathbf{x})$; this is particularly challenging because the section of the phase space explored by models is time-dependent. Since the goal of our approach is not to propose an alternative algorithm to simulate the evolution of the system, but to build a tool to compare how different models react to social stimuli, one can fix the probability distribution of the other agents, $P(\mathbf{x}_{j\neq i})$, and use that to simulate the behavior of agent i.

This approach was used to produce all IRFs in the subsequent subsections (Fig. 3, 4, and 5). A more detailed step-by-step explanation of this approach is described in the SI, sec. 7.2.1. Additionally, the SI contains information on the signal s_i and the distribution of states of the other agents $P(\mathbf{x}_{j\neq i})$ for all IRFs shown.

3.2 Influence-response functions for influence models

Before addressing models that assume that protective behavior is conditioned on the populations' health state, we show in this section how *behavior* models from the literature can be expressed using IRFs. Our framework also allows one to use the same formalism for *internal-state* models.

Fig. 3 illustrates the IRFs of three behavior models. Panels a and b show two IRFs studied by López-Pintado and Watts [92] to demonstrate that these models generate very different macroscopic dynamics. The IRF in panel a is obtained from an upward sloping threshold model, where a given behavior becomes more likely as the number of network neighbors displaying the behavior rises. This IRF can be derived, among others, from game-theoretical models of coordination in networks. For instance, communication technology such as online social networks are more useful when many contacts are using the same platform. Panel b shows the opposite case in which a behavior is less

likely to be adopted when many contacts have already adopted it. This IRF is typical for public-good games, a setting in which an agent can free-ride off other individuals who adopt a costly behavior.

Panel c shows an IRF of a coupled disease-behavior model where each agent randomly chooses between rational decision-making and imitation, similar to [150]. In essence, this model combines the upward and downward threshold models shown in Fig. 3a and b. Specifically, an agent's probability of engaging in self-protecting behavior, in this case, receiving a vaccination, depends on the average behavior of their local neighbors in the previous time step. Each time step can be interpreted as a season of a flu epidemic. The agents can follow one of two strategies: act rationally, by taking a vaccine if the fraction of vaccinated neighbors in the previous season was below a certain threshold (dark green), or imitate the dominant behavior of others (light green). In the second case, the probability of taking the vaccine increases with the fraction of vaccinated neighbors. In the intermediate regime where the choice between rational and imitative behavior is random, the IRF exhibits a discontinuity.

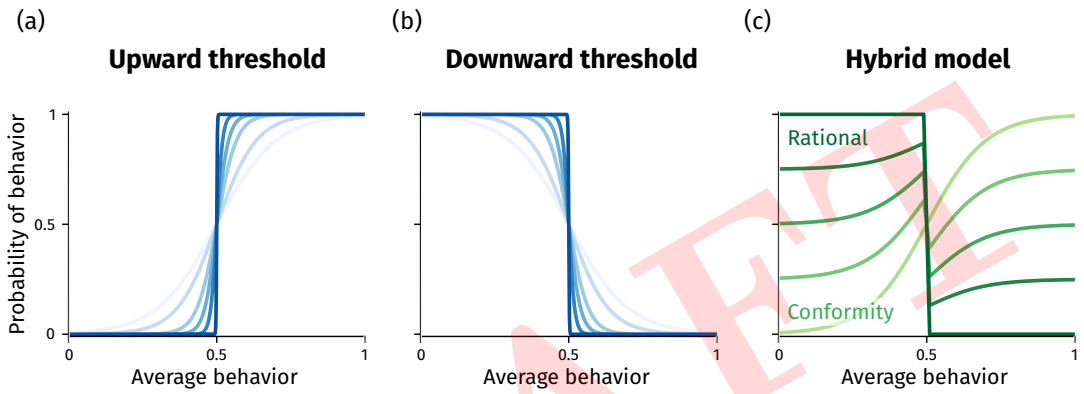


Figure (3) **Three examples of behavioral influence-response functions (IRFs).** The probability of expressing a binary behavior is a function of the average behavior in the relevant population. Upward (a) and downward (b) threshold models implemented by logistic functions. (c) hybrid model, similar to the model by [150], where agents randomly chose between two decision-making options: rationality (shown dark green) and imitation (shown in light green); details on the derivation of these IRFs are in the SI section 2.2.

In Fig. 4, we show the IRFs for the three competing internal-state models reviewed above: assimilation, reinforcement, and repulsion models. Panels a-c of Fig. 4 show how an agent i shifts its opinion from its initial position $b_i^t = 0.2$ after interaction with different neighbors j. According to the assimilation model, i always moves closer to j. Reinforcement, in contrast, implies that i adopts a more extreme opinion when j holds a similar view. Repulsion can also generate shifts towards more extreme opinions; however, unlike in the reinforcement model, these shifts occur when agents with very different views interact. These differences are also visible in the IRFs of the three models, as panels d, e, and f show. Again, we focus on the scenario where agent i holds an opinion of $b_i^t = 0.2$ (see dotted lines). The solid lines show the expected updated opinions after interaction. We use the average opinion of the neighbors of agent i as an observable of the system

$$s_i := f_i(\mathbf{b}) := \frac{1}{k_i} \sum_{j=1}^{N} A_{ij} b_j, \tag{10}$$

where A_{ij} is the adjacency matrix of the underlying network.

The IRF of assimilation models (panel d) is sloping linearly upward since assimilation is implemented as simple averaging. Panel e shows the IRF of a reinforcement model, which is also

upward-sloping. The main difference compared to assimilation, however, is that agent i's opinion is reinforced by s_i when the initial opinion distance is low. In fact, agent i's opinion may even overshoot, adopting an opinion value closer to zero than s_i. Repulsion models have a very different IRF (see panel f). Repulsion is often implemented as a weighted average with positive weights when agents agree (assimilation) and negative weights when agents disagree too much (repulsion). Accordingly, the IRF is upward-sloping when i agrees with its contacts and downward-sloping when they are too dissimilar. For a more comprehensive comparison of the three models, we present vector fields in panels g-i of Fig. 4. In these panels, arrows depict shifts in opinion for agent i depending on its position before the opinion update and the local average opinion s_i.

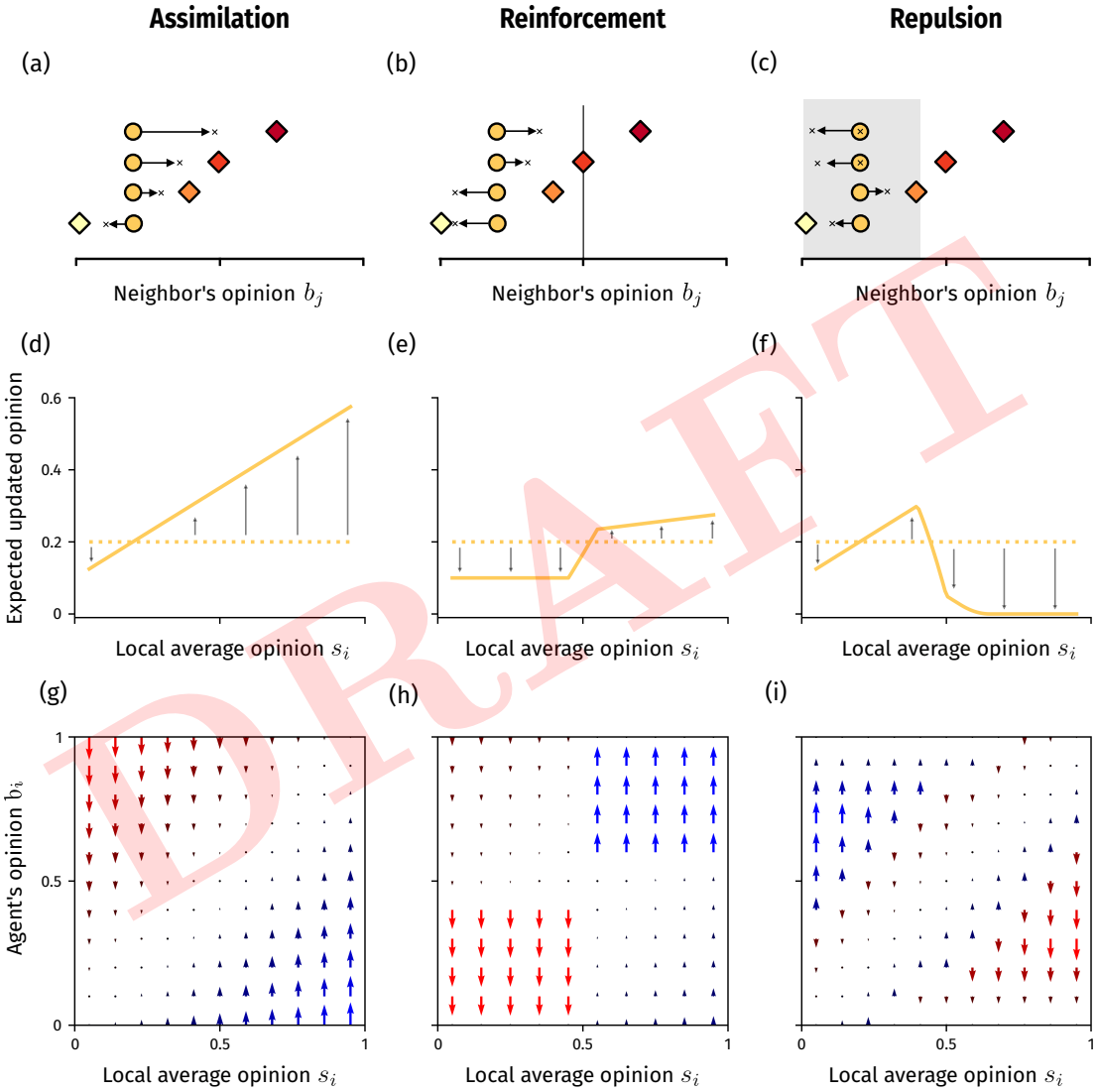


Figure (4) **Influence-response functions of internal-state models** assuming assimilation (a, d, g), reinforcement (b, e, h), and repulsion (c, f, i). First row: agent i (circle) with an initial opinion of $b_i^t = 0.2$ is influenced by agent j (diamond). The arrows illustrate how the opinion of agent i updates after interaction, with opinion shown on x-axis. Second row: expected updated opinion of an agent i with an initial opinion $b_i^t = 0.2$ (dotted line) depending on the local average opinion s_i in i's neighborhood. Third row: vector fields for the update of i's opinion in the phase space spanned by local average opinion s_i and i's previous opinion b_i^t. For details see SI Section 2.3.

3.3 Influence-response functions including effects of health states

As shown in Sec. 2, existing models of coupled disease and behavior dynamics differ not only in terms of the influence model but also in their assumptions about agents' responses to the heath state of the population (heath-to-behavior bridge). The models presented in [18] and [3], for instance, focus on a scenario where agents decide between getting vaccinated or not. Vaccination probability of an agent depends on the fractions of infected (I) and vaccinated (V) neighbors. The framework we propose allows the comparison of these assumptions. To this end, we define two signal functions $f_i^{(\mathrm{I})}(\mathbf{x})$ and $f_i^{(\mathrm{V})}(\mathbf{x})$ as

$$\begin{aligned} s_i^{(\mathrm{I})} &:= f_i^{(\mathrm{I})}(\mathbf{x}) = \frac{1}{k_i}\sum_j \mathbf{A}_{ij}\delta_{x_j,\mathrm{I}} \\ s_i^{(\mathrm{V})} &:= f_i^{(\mathrm{V})}(\mathbf{x}) = \frac{1}{k_i}\sum_j \mathbf{A}_{ij}\delta_{x_j,\mathrm{V}}, \end{aligned} \tag{11}$$

that return the fraction of neighbors in the state I or V, respectively.

First, we will examine how the probability of behavior depends on the number of infected agents, either in the neighborhood of an agent i or in the global population. Fig. 5, panels a, b, and c, show the IRFs of three existing models: a success-driven imitation model [18] in Fig. 5a, an awareness-spreading model [64] in Fig. 5b, and a model which includes negative feedback from vaccinated who nevertheless got infected [3] in Fig. 5c. All plots shown in Fig. 5 are obtained by simplifying the original models, as detailed in the SI 7.3.

Although the success-driven imitation model (panel a) and the awareness model (panel b) are based on very different behavior theories of decision-making, their IRFs are markedly similar. In both models, the probability of adopting protective behavior increases with the number of infected neighbors. In both models, it is also possible that agents do not respond to the fraction of infected agents in the population. In the success-driven imitation model, this happens when the agent i is susceptible. Such an agent has obtained maximal payoff in the previous round (zero costs of infected and protection); therefore, it does not reconsider the decision. Likewise, in the awareness model by [64], aware agents protect unconditionally.

Unlike the success-driven imitation model and the awareness model, the model by Alvarez-Zuzek [3] shows a downward-sloping IRF: under certain conditions the average local opinion about vaccines decreases with the number of vaccinated infected neighbors. This is possible because the more infected neighbors there are, the higher the chances that a vaccinated agent becomes infected despite self-protection. This decreases the local average opinion about vaccines. Similarly, as the fraction of vaccinated neighbors increases from $s_i^{(\mathrm{V})} = 0.25$ to $s_i^{(\mathrm{V})} = 0.99$, the likelihood of one of them getting infected also increases, leading to a faster decline in the local average opinion about vaccines.

Following the same approach, one can also detail how the probability of behavior depends on the fraction of vaccinated neighbors $s_i^{(\mathrm{V})}$. The resulting IRFs for three models are shown in Fig. 5d, e, and f. The specific models we consider are a success-driven imitation model [18] in Fig. 5d, a game-theory model [104] in Fig. 5e, and a negative-feedback model [3] in Fig. 5f.

The dependency of behavior upon the fraction of vaccinated neighbors is very diverse. In success-driven imitation models, Fig 5d, the probability of protective behavior increases with the fraction of vaccinated individuals at a fixed incidence because it is more likely to randomly select a vaccinated agent to copy if more agent are vaccinated. In game-theoretical models where agents try to maximize their payoff by evaluating the probability of becoming infected, the probability of vaccinating drops suddenly when enough neighbors have been vaccinated, as shown in Fig. 5e. Finally, in negative feedback models, Fig. 5f, the IRF depends on the area of the parameter space one is exploring. If disease prevalence is low, increasing the fraction of vaccinated neighbors increases the fraction of healthy vaccinated neighbors. Conversely, for high disease prevalence, increasing the fraction

of vaccinated neighbors mostly increases the fraction of vaccinated neighbors who got infected, decreasing the overall opinion about vaccinations.

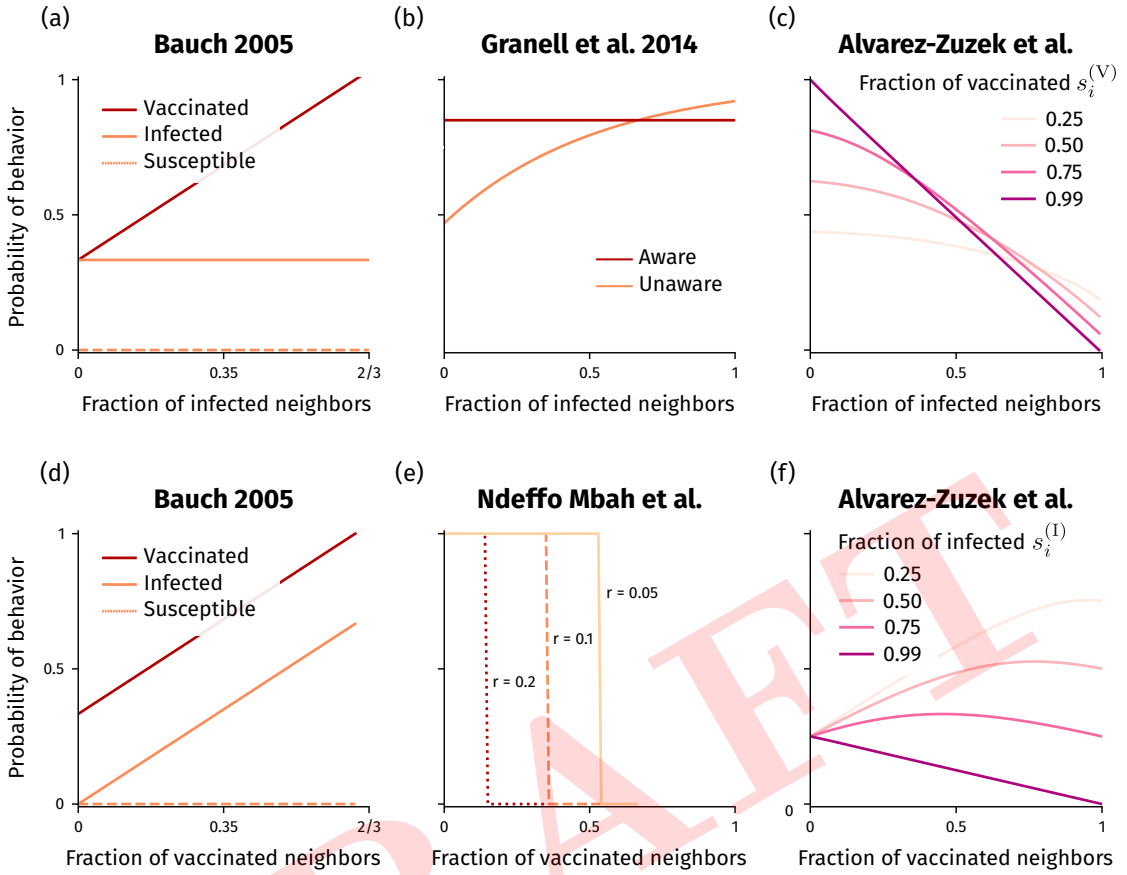


Figure (5) **Influence-response functions (IRFs) from existing models of coupled disease-behavior dynamics.** First row (a-c): probability of protective behavior as a function of the fraction of infected neighbors in the population; second row (d-f): protection probability as a function of the fraction of vaccinated neighbors; a and d: examples of the success-imitation game described in [18] where a third of the agents are vaccinated (in a) and infected (in d), and the fraction of infected (a) and vaccinated (d) is varied from 0 and 2/3; b: the awareness model described in [63]; c and f: model described in [3], where vaccines are not perfect, but reduce susceptibility by 90%, the probability of vaccination is shown as a function of the fraction of infected neighbors in c and as a function of the vaccinated neighbors in f, for different fractions of vaccinated and infected respectively; e: IRF of a simplification of the model described in [104], where agents vaccinate when the risk of becoming infected is too high, and the risk decreases as a function of the vaccinated neighbors, the same curve is shown for three ratios between the cost of vaccination and the cost of the infection. Details on the derivation of these IRFs are in the SI Section 7.3.

In summary, the influence-response functions shown in Figures 4 and 5 illustrate our main argument in favor of using these functions: they allow to compare influence models and health-to-behavior bridges, even when the underlying behavior theories are challenging to compare. Sometimes even theories rooted in vastly different assumptions about decision-making actually imply remarkably similar IRFs. This suggests that their macroscopic implications about heath dynamics may also be comparable. However, sometimes IRFs also turn out to be fundamentally different, as in Fig. 5d and e. These differences have the potential to generate fundamentally different behavior and health dynamics since they originate from the fundamental source of complexity: the interaction between microscopic entities.

3.4 Testing influence-response functions

Influence-response functions are not only a powerful tool to compare theories of behavior; they can also be empirically tested in a more straightforward way than the theories they have been derived from. Existing behavior theories depend on latent psychological determinants such as opinions, perceived risks, and preferences, which require complex measurement tools. In contrast, the input variables to influence-response functions, such as neighbors' behavior and disease prevalence, are often directly observable and easier to quantify. In fact, social sciences provide a powerful set of approaches of quantitative empirical research to test IRFs. There are at least four well elaborated approaches with complementary strengths and weaknesses.

First, there are various sources of observational data about protective behavior. During the COVID-19 pandemic, for instance, various regional, national, and international dashboards provided large-scale administrative data about vaccine uptake [43]. Likewise, researchers gathered information from mobile-phone data on reductions in mobility in contact frequency [36, 102, 121]. This information can be pooled with information about local disease prevalence to test health-behavior-bridges. Unfortunately, many forms of protective behavior are hard to record or come with privacy restrictions that make data totally unavailable or limit access to high aggregation levels.

Second, surveys are a frequently applied method. During the COVID-19 pandemic, for instance, the "COVID-19 behavior tracker" conducted repeated and representative online surveys measuring many forms of protective behavior in various countries [78]. Together with information about disease prevalence and governmental information, this data can also be used to test health-behavior-bridges. In addition, there are various survey methods allowing to assess participants' risk perceptions, opinions about protection behavior, behavior norms, and perceptions about others' behavior [2]. Furthermore, there are sophisticated statistical methods to model the relationships between behavior and its determinants [85]. Obviously, surveys rely on the often problematic assumption that participants are able and motivated to accurately reflect about their opinions and behavior. What is more, surveys also provide observational data, which limits the testing of causal statements.

Third, there are vignette studies, an approach that introduces experimentation into surveys [133]. Vignette designs do not measure actual behavior but self-reported intentions in a hypothetical situation. Bicchieri et al., for instance, confronted survey respondents with multiple descriptions of different social settings and asked what behavior was considered most appropriate for each of them [21]. This approach allowed them to test which aspects of a given setting increased the motivation to engage in protective behavior.

A fourth approach to test IRFs are incentivized experiments along the lines of behavioral game-theory [26]. In this empirical paradigm, participants are provided with monetary incentives that simulate the decision problem faced by individuals deciding whether to protect or not. In fact, this paradigm has been used to implement the same payoff structure as assumed in the game-theoretical models described above [22]. The advantage of this approach over surveys is that participants' responses carry real-world significance as their decisions directly translate into monetary consequences. Combined with randomized control trials manipulating the behavior of others [116] and information about local and global prevalence, this approach provides a means to directly test IRFs while reducing spurious causality. However, behavioral experiments are often criticized for studying behavior in highly stylized, artificial settings, which raises concerns about the external validity of this approach.

Given that each empirical method has its own set of strengths and weaknesses, it is crucial to employ them in tandem. Similarly, it is essential to test IRFs across various countries, cultural contexts, pandemic phases, and diverse communities within a population, since factors such as age, education, socioeconomic status, and religion can significantly influence individuals' decisions regarding self-protection [126].

4 *Selecting* coupled models

Above, we elaborated a framework to *describe* models of coupled infectious disease and behavior dynamics. This involved identifying four update functions and scrutinizing the implementation of each in existing models. Despite two decades of research yielding a diverse array of compelling individual models, the substantial quantity and diversity of alternative model assumptions have resulted in an inconsistent literature. In particular, we have documented a vast array of approaches to modeling individuals' decision-making regarding protective behavior, and how this decision is influenced by others' choices and the prevalence of the disease.

Next, we advocated an approach to *compare* competing models of behavior put forward in the social sciences [92, 97]. This approach suggests that modelers investigating the macroscopic outcomes of interdependent decision-making on the micro-level should abstain from formally representing the psychological processes driving micro-behavior without demonstrating the necessity of these micro-level assumptions for generating distinct macro-dynamics. Instead, modelers should encapsulate the formal relationship between an agent's behavior and the behavior of others in what is referred to as an influence-response function (IRF).

However, to make IRFs applicable to the literature on coupled models of disease and behavior dynamics, the concept needed to be extended in two ways. First, we included that an agent's decision might also depend on the state of the disease dynamics, incorporating the health-to-behavior-bridge in the IRF. Second, unlike Lopez-Pintado and Watts who studied a decision problem where agent's behavior was a function of the average behavior in the population, we extended the concept to cover any combination of the local average behavior, Fig. 3 and 4, fraction of infected neighbors, Fig. 5a, b, c, and fraction of vaccinated neighbors, Fig. 5d, e, f, and of other scalar observables.

While we do not dispute psychological realism in general, abstract IRFs come with three main advantages. First, simple descriptions of the complexity within individuals make it easier to understand the complexity arising from the interaction between individuals. Models of coupled infectious disease and behavior dynamics combine two model families that generate complex dynamics already when studied in isolation. Keeping micro-models simple when coupling helps to understand the results of the coupled system.

Second, IRFs derived from alternative coupled models can be compared in a straight-forward way. In contrast, the decision-making models underlying these functions often differ on various fundamental assumptions about how individuals make decisions, which makes it hard to ascertain how these differences translate into different protective behavior. With examples, we illustrated that sometimes seemingly different behavior theories actually imply similar IRFs and vice versa [92, 97].

Third, competing IRFs can be put to the test with existing empirical approaches, as we discussed in Sec. 3. Therefore, IRFs guide empirical research selecting empirically accurate model assumptions, despite being abstract descriptions of a complex decision processes.

We argue that IRFs can help move the literature on models coupling disease and behavior dynamics into a process of systematic empirical validation, guiding empirical research testing alternative influence models and health-to-behavior bridges. To this end, we propose three recommendations guiding modelers introducing new models.

Recommendation 1 *Build, as much as possible, on existing models.*

While we do not dispute the epistemological value of novel models and recombination of existing model assumptions, we criticize that many contributions to the literature introduce new models without demonstrating whether existing models would have done the same job or whether the presented findings follow only from the newly introduced model. As a consequence, we recommend future modelers to depart from one of the many existing models when they seek to demonstrate a given effect, unless this is not possible in that the new effect cannot be generated by an existing model. Ideally, this will help the community to identify a set of standard models that are well understood

and serve as a reference point to demonstrate new findings. Likewise, when an innovative model assumption is introduced, the remaining update functions should remain unchanged compared to an earlier publication, in order to demonstrate that the novel assumption is responsible for the new findings and not a change in some other model aspect. For instance, existing contributions do not only apply different models of behavior, they often also study dynamics in different network topologies. While it is extremely valuable to study effects of network topologies, it is also important to make one's work comparable to existing work.

Recommendation 2 *When you introduce a new influence model or health-to-behavior bridge, identify the influence-response functions and show where they differ from existing models.*

In Section 3, we provided examples of how fundamentally different behavior theories can be compared in terms of what matters for coupled behavior-disease dynamics. Deriving IRFs from a complicated model of individual decision-making can be challenging. While analytical solutions are very desirable, they sometimes force one to make very strong simplifying assumptions. In this case, it can be insightful to simulate the model and statistically infer IRFs from the observed agent behavior.

Recommendation 3 *If your behavioral model implies an influence-response function that deviates from other models and if this difference alters disease dynamics, test these functions empirically against each other.*

Ultimately, the process of selecting model assumptions is an empirical exercise. However, empirical research must be guided by theoretical foundations. Influence-response functions serve to highlight differences between micro assumptions. If subsequent theoretical exploration demonstrates that these distinctions also manifest in varying macro-predictions concerning disease dynamics, empirical research becomes imperative to select models. As detailed in Section 3, we have outlined alternative social-scientific approaches for directly testing IRFs. The application of these approaches across diverse social and cultural contexts, coupled with concurrent formal modeling, is an intricate and interdisciplinary endeavor. Nevertheless, it holds the potential to elevate the literature on models of coupled infectious disease and behavior dynamics from a collection of independent models to a disciplinary canon capable of reliably informing political decision-making during a pandemic.

5 Acknowledgements

We thank Kai Nagel, Andreas Flache, and Hendrik Nunner for helpful feedback.

6 Funding

All authors acknowlegde funding from the German Federal Ministry for Education and Research for the infoXpand project (031L0300A,031L0300C), within the MONID consortium.

VP is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC 2067/1- 390729940, and by the CRC1528 - "Cognition of Interaction".

References

[1] Robert P Abelson. "Mathematical models of the distribution of attitudes under controversy". In: *Contributions to mathematical psychology* (1964).

[2] Ozan Aksoy. "Within-family influences on compliance with social-distancing measures during COVID-19 lockdowns in the United Kingdom". In: *Nature Human Behaviour* (2022), pp. 1–9.

[3] Lucila G Alvarez-Zuzek et al. “Epidemic spreading in multiplex networks influenced by opinion exchanges on vaccination”. In: *PloS one* 12.11 (2017), e0186492.

[4] Michael A Andrews and Chris T Bauch. “The impacts of simultaneous disease intervention decisions on epidemic outcomes”. In: *Journal of theoretical biology* 395 (2016), pp. 1–10.

[5] Solomon E Asch. “Opinions and social pressure”. In: *Scientific American* 193.5 (1955), pp. 31–35.

[6] Robert Axelrod. “The dissemination of culture: A model with local convergence and global polarization”. In: *Journal of conflict resolution* 41.2 (1997), pp. 203–226.

[7] Franco Bagnoli, Pietro Lio, and Luca Sguanci. “Risk perception in epidemic modeling”. In: *Physical Review E* 76.6 (2007), p. 061904.

[8] Christopher A Bail et al. “Exposure to opposing views on social media can increase political polarization”. In: *Proceedings of the National Academy of Sciences* 115.37 (2018), pp. 9216–9221.

[9] Eytan Bakshy, Solomon Messing, and Lada A Adamic. “Exposure to ideologically diverse news and opinion on Facebook”. In: *Science* 348.6239 (2015), pp. 1130–1132.

[10] Stefano Balietti, Michael Mäs, and Dirk Helbing. “On disciplinary fragmentation and scientific progress”. In: *PloS one* 10.3 (2015), e0118747.

[11] W Marijn van Ballegooijen and Maarten C Boerlijst. “Emergent trade-offs and selection for outbreak frequency in spatial epidemics”. In: *Proceedings of the National Academy of Sciences* 101.52 (2004), pp. 18246–18250.

[12] Sven Banisch. *Markov chain aggregation for agent-based models*. Understanding Complex Systems. Springer, 2016.

[13] Sven Banisch. “Unfreezing social dynamics: Synchronous update and dissimilation”. In: *Proceedings of the 3rd World Congress on Social Simulation*. 2010.

[14] Sven Banisch and Tanya Araújo. “Who replaces whom? Local versus non-local replacement in social and evolutionary dynamics”. In: *Discontinuity, Nonlinearity, and Complexity* 2.1 (2012), pp. 57–73.

[15] Sven Banisch, Ricardo Lima, and Tanya Araújo. “Agent based models and opinion dynamics as Markov chains”. In: *Social Networks* 34.4 (2012), pp. 549–561.

[16] Sven Banisch and Eckehard Olbrich. “Opinion polarization by learning from social feedback”. In: *The Journal of Mathematical Sociology* 43.2 (2019), pp. 76–103.

[17] Sven Banisch and Hawal Shamon. “Biased processing and opinion polarisation: experimental refinement of argument communication theory in the context of the energy debate”. In: *Sociological Methods and Research* (2023).

[18] Chris T Bauch. “Imitation dynamics predict vaccinating behaviour”. In: *Proceedings of the Royal Society B: Biological Sciences* 272.1573 (2005), pp. 1669–1675.

[19] Chris T Bauch and David JD Earn. “Vaccination and the theory of games”. In: *Proceedings of the National Academy of Sciences* 101.36 (2004), pp. 13391–13394.

[20] Jamie Bedson et al. “A review and agenda for integrated disease models including social and behavioural factors”. In: *Nature human behaviour* 5.7 (2021), pp. 834–846.

[21] Cristina Bicchieri et al. “In science we (should) trust: Expectations and compliance across nine countries during the COVID-19 pandemic”. In: *PloS one* 16.6 (2021), e0252892.

[22] Robert Böhm, Cornelia Betsch, and Lars Korn. “Selfish-rational non-vaccination: Experimental evidence from an interactive vaccination game”. In: *Journal of Economic Behavior & Organization* 131 (2016), pp. 183–195.

[23] Camila Buono et al. “Epidemics in partially overlapped multiplex networks”. In: *PloS one* 9.3 (2014), e92200.

[24] Donald S Burke et al. “Individual-based computational modeling of smallpox epidemic control strategies”. In: *Academic Emergency Medicine* 13.11 (2006), pp. 1142–1149.

[25] Eugene Burnstein and Amiram Vinokur. “Testing two classes of theories about group induced shifts in individual choice”. In: *Journal of experimental social psychology* 9.2 (1973), pp. 123–137.

[26] Colin F Camerer. *Behavioral game theory: Experiments in strategic interaction*. Princeton university press, 2011.

[27] Ellsworth Campbell and Marcel Salathé. “Complex social contagion makes networks more vulnerable to disease outbreaks”. In: *Scientific reports* 3.1 (2013), pp. 1–6.

[28] Alejandro Carballosa, Mariamo Mussa-Juane, and Alberto P Muñuzuri. “Incorporating social opinion in the evolution of an epidemic spread”. In: *Scientific Reports* 11.1 (2021), p. 1772.

[29] Claudio Castellano, Santo Fortunato, and Vittorio Loreto. “Statistical physics of social dynamics”. In: *Reviews of modern physics* 81.2 (2009), p. 591.

[30] Frederick H Chen. “A susceptible-infected epidemic model with voluntary vaccinations”. In: *Journal of mathematical biology* 53 (2006), pp. 253–272.

[31] Jiangzhuo Chen, Achla Marathe, and Madhav Marathe. “Feedback between behavioral adaptations and disease dynamics”. In: *Scientific reports* 8.1 (2018), p. 12452.

[32] MY Choi and BA Huberman. “Digital dynamics and the simulation of magnetic systems”. In: *Physical Review B* 28.5 (1983), p. 2547.

[33] Flávio Codeço Coelho and Claudia T Codeço. “Dynamic modeling of vaccinating behavior as a function of individual beliefs”. In: *PLoS computational biology* 5.7 (2009), e1000425.

[34] Sebastian Contreras et al. “Low case numbers enable long-term stable pandemic control without lockdowns”. In: *Science advances* 7.41 (2021), eabg2243.

[35] Daniel M Cornforth et al. “Erratic flu vaccination emerges from short-sighted behavior in contact networks”. In: *PLoS computational biology* 7.1 (2011), e1001062.

[36] “COVID-19 Mobility Project”. In: (). `https://www.covid-19-mobility.org/mobility-monitor/`, Accessed: 2023-11-01.

[37] Alberto d’Onofrio, Piero Manfredi, and Piero Poletti. “The impact of vaccine side effects on the natural history of immunization programmes: an imitation-game approach”. In: *Journal of theoretical biology* 273.1 (2011), pp. 63–71.

[38] James A Davis. “What’s wrong with sociology?” In: *Sociological Forum*. JSTOR. 1994, pp. 179–197.

[39] Guillaume Deffuant, Marijn A Keijzer, and Sven Banisch. “Regular access to constantly renewed online content favors radicalization of opinions”. In: *arXiv preprint arXiv:2305.16855* (2023).

[40] Morris H DeGroot. “Reaching a consensus”. In: *Journal of the American Statistical association* 69.345 (1974), pp. 118–121.

[41] Jonas Dehning et al. “Inferring change points in the spread of COVID-19 reveals the effectiveness of interventions”. In: *Science* 369.6500 (2020), eabb9789.

[42] Mark Dickison, Shlomo Havlin, and H Eugene Stanley. “Epidemics on interconnected networks”. In: *Physical Review E* 85.6 (2012), p. 066109.

[43] Ensheng Dong, Hongru Du, and Lauren Gardner. “An interactive web-based dashboard to track COVID-19 in real time”. In: *The Lancet infectious diseases* 20.5 (2020), pp. 533–534.

[44] Philipp Dönges et al. “Interplay between risk perception, behaviour, and COVID-19 spread”. In: *Frontiers in Physics* (2022), p. 68.

[45] Erhu Du et al. “How do social media and individual behaviors affect epidemic transmission and control?” In: *Science of the Total Environment* 761 (2021), p. 144114.

[46] Joshua M Epstein et al. “Coupled contagion dynamics of fear and disease: mathematical and computational explorations”. In: *PloS one* 3.12 (2008), e3955.

[47] Stephen Eubank et al. “Modelling disease outbreaks in realistic urban social networks”. In: *Nature* 429.6988 (2004), pp. 180–184.

[48] Andreas Flache and Michael W Macy. “Local convergence and global diversity: From interpersonal to social influence”. In: *Journal of Conflict Resolution* 55.6 (2011), pp. 970–995.

[49] Andreas Flache and Michael Mäs. “How to get the timing right. A computational model of the effects of the timing of contacts on team cohesion in demographically diverse teams”. In: *Computational and Mathematical Organization Theory* 14.1 (2008), pp. 23–51.

[50] Andreas Flache et al. “Models of social influence: Towards the next frontiers”. In: *Journal of Artificial Societies and Social Simulation* 20.4 (2017).

[51] John RP French Jr. “A formal theory of social power.” In: *Psychological review* 63.3 (1956), p. 181.

[52] Noah E Friedkin and Eugene C Johnsen. *Social influence network theory: A sociological examination of small group dynamics*. Vol. 33. Cambridge University Press, 2011.

[53] Feng Fu et al. “Imitation dynamics of vaccination behaviour on social networks”. In: *Proceedings of the Royal Society B: Biological Sciences* 278.1702 (2011), pp. 42–49.

[54] Eriko Fukuda, Jun Tanimoto, and Mitsuhiro Akimoto. “Influence of breaking the symmetry between disease transmission and information propagation networks on stepwise decisions concerning vaccination”. In: *Chaos, Solitons & Fractals* 80 (2015), pp. 47–55.

[55] Eriko Fukuda et al. “Risk assessment for infectious disease and its impact on voluntary vaccination behavior in social networks”. In: *Chaos, Solitons & Fractals* 68 (2014), pp. 1–9.

[56] Sebastian Funk, E Gilad, and Vincent AA Jansen. “Endemic disease, awareness, and local behavioural response”. In: *Journal of theoretical biology* 264.2 (2010), pp. 501–509.

[57] Sebastian Funk, Marcel Salathé, and Vincent AA Jansen. “Modelling the influence of human behaviour on the spread of infectious diseases: a review”. In: *Journal of the Royal Society Interface* 7.50 (2010), pp. 1247–1256.

[58] Sebastian Funk et al. “The spread of awareness and its impact on epidemic outbreaks”. In: *Proceedings of the National Academy of Sciences* 106.16 (2009), pp. 6872–6877.

[59] ET Gawlinski et al. “Growth of unstable domains in the two-dimensional Ising model”. In: *Physical Review B* 31.1 (1985), p. 281.

[60] Pierre-Yves Geoffard and Tomas Philipson. “Rational epidemics and their public control”. In: *International economic review* (1996), pp. 603–624.

[61] Herbert Gintis. *The Bounds of Reason: Game Theory and the Unification of the Behavioral Sciences-Revised Edition*. Princeton University Press, 2014.

[62] James P Gleeson. “Binary-state dynamics on complex networks: Pair approximation and beyond”. In: *Physical Review X* 3.2 (2013), p. 021004.

[63] Clara Granell, Sergio Gómez, and Alex Arenas. “Competing spreading processes on multiplex networks: awareness and epidemics”. In: *Physical review E* 90.1 (2014), p. 012808.

[64] Clara Granell, Sergio Gómez, and Alex Arenas. "Dynamical interplay between awareness and epidemic spreading in multiplex networks". In: *Physical review letters* 111.12 (2013), p. 128701.

[65] Reuben M Granich et al. "Universal voluntary HIV testing with immediate antiretroviral therapy as a strategy for elimination of HIV transmission: a mathematical model". In: *The Lancet* 373.9657 (2009), pp. 48–57.

[66] Mark Granovetter. "Threshold models of collective behavior". In: *American journal of sociology* 83.6 (1978), pp. 1420–1443.

[67] Thilo Gross, Carlos J Dommar D'Lima, and Bernd Blasius. "Epidemic dynamics on an adaptive network". In: *Physical review letters* 96.20 (2006), p. 208701.

[68] Gerrit Großmann, Michael Backenköhler, and Verena Wolf. "Heterogeneity matters: Contact structure and individual variation shape epidemic dynamics". In: *Plos one* 16.7 (2021), e0250050.

[69] Douglas D Heckathorn. "The dynamics and dilemmas of collective action". In: *American sociological review* (1996), pp. 250–277.

[70] Herbert W Hethcote. "The mathematics of infectious diseases". In: *SIAM review* 42.4 (2000), pp. 599–653.

[71] George C Homans. *Social behavior: Its elementary forms*. Harcourt Brace Jovanovich, 1974.

[72] He Huang, Yahong Chen, and Yefeng Ma. "Modeling the competitive diffusions of rumor and knowledge and the impacts on epidemic spreading". In: *Applied mathematics and computation* 388 (2021), p. 125536.

[73] Jiechen Huang, Juan Wang, and Chengyi Xia. "Role of vaccine efficacy in the vaccination behavior under myopic update rule on complex networks". In: *Chaos, Solitons & Fractals* 130 (2020), p. 109425.

[74] Bernardo A Huberman and Natalie S Glance. "Evolutionary games and computer simulations." In: *Proceedings of the National Academy of Sciences* 90.16 (1993), pp. 7716–7718.

[75] "Investigation of vaccination game approach in spreading covid-19 epidemic model with considering the birth and death rates". In: *Chaos, Solitons & Fractals* 163 (2022), p. 112565.

[76] Luis R Izquierdo et al. "Techniques to understand computer simulations: Markov chain analysis". In: *Journal of Artificial Societies and Social Simulation* 12.1 (2009), p. 6.

[77] Dennis Jacob and Sven Banisch. "Polarization in Social Media: A Virtual Worlds-Based Approach". In: *Journal of Artificial Societies and Social Simulation* 26.3 (2023).

[78] SP Jones. "Imperial College London Big Data Analytical Unit, & YouGov Plc (2020) Imperial College London YouGov Covid Data Hub, v1. 0". In: *YouGov Plc* (2020). https://github.com/YouGov-Data/covid-19-tracker.

[79] Claus Kadelka and Audrey McCombs. "Effect of homophily and correlation of beliefs on COVID-19 and general infectious disease outbreaks". In: *PloS one* 16.12 (2021), e0260973.

[80] Matt J Keeling and Ken TD Eames. "Networks and epidemic models". In: *Journal of the royal society interface* 2.4 (2005), pp. 295–307.

[81] Marijn A Keijzer and Michael Mäs. "The complex link between filter bubbles and opinion polarization". In: *Data Science* Preprint (2022), pp. 1–28.

[82] Marijn A Keijzer, Michael Mäs, and Andreas Flache. "Communication in online social networks fosters cultural isolation". In: *Complexity* 2018 (2018).

[83] Marijn A Keijzer, Michael Mäs, and Andreas Flache. "Polarization on social media: Micro-level evidence and macro-level implications". In: *Chapter 2 in Keijzer M. Opinion Dynamics in Online Social Media. Dissertation University of Groningen* (2022).

[84] William Ogilvy Kermack and Anderson G McKendrick. "A contribution to the mathematical theory of epidemics". In: *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character* 115.772 (1927), pp. 700–721.

[85] Lilian Kojan et al. "Perceptions of behaviour efficacy, not perceptions of threat, are drivers of COVID-19 protective behaviour in Germany". In: *Humanities and Social Sciences Communications* 9.1 (2022), pp. 1–15.

[86] Cristian E La Rocca, Lidia A Braunstein, and Federico Vazquez. "The influence of persuasion in opinion formation and polarization". In: *Europhysics Letters* 106.4 (2014), p. 40004.

[87] Jana Lasser et al. "Agent-based simulations for protecting nursing homes with prevention and vaccination strategies". In: *Journal of the Royal Society Interface* 18.185 (2021), p. 20210608.

[88] Jana Lasser et al. "Assessing the impact of SARS-CoV-2 prevention measures in Austrian schools using agent-based simulations and cluster tracing data". In: *Nature Communications* 13.1 (2022), p. 554.

[89] Teddy Lazebnik. "Computational applications of extended SIR models: A review focused on airborne pandemics". In: *Ecological Modelling* 483 (2023), p. 110422.

[90] Tony Zhiyang Lin and Xiaoli Tian. "Audience design and context discrepancy: How online debates lead to opinion polarization". In: *Symbolic interaction* 42.1 (2019), pp. 70–97.

[91] Xiao-Tao Liu, Zhi-Xi Wu, and Lianzhong Zhang. "Impact of committed individuals on vaccination behavior". In: *Physical Review E* 86.5 (2012), p. 051132.

[92] Dunia Lopez-Pintado and Duncan J Watts. "Social influence, binary decisions and collective dynamics". In: *Rationality and Society* 20.4 (2008), pp. 399–443.

[93] Charles G Lord, Lee Ross, and Mark R Lepper. "Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence." In: *Journal of personality and social psychology* 37.11 (1979), p. 2098.

[94] Michael W Macy et al. "Polarization in dynamic networks: A Hopfield model of emergent structure". In: *Research Handbook on Analytical Sociology*. Ed. by K. Carley Breiger and P. Pattison. Dynamic Social Network Modeling, Analysis: Workshop Summary, and Papers, 2003, pp. 162–173.

[95] Liang Mao. "Modeling triple-diffusions of infectious diseases, information, and preventive behaviors through a metropolitan social network—an agent-based simulation". In: *Applied Geography* 50 (2014), pp. 31–39.

[96] Liang Mao and Yan Yang. "Coupling infectious diseases, human preventive behavior, and networks–a conceptual framework for epidemic modeling". In: *Social science & medicine* 74.2 (2012), pp. 167–175.

[97] Michael Mäs. "Interactions". In: *Research Handbook on Analytical Sociology*. Edward Elgar Publishing, 2021.

[98] Michael Mäs and Andreas Flache. "Differentiation without distancing. Explaining bi-polarization of opinions without negative influence". In: *PloS one* 8.11 (2013), e74516.

[99] Michael Mäs and Karl-Dieter Opp. "When is ignorance bliss? Disclosing true information and cascades of norm violation in networks". In: *Social Networks* 47 (2016), pp. 116–129.

[100] Miller McPherson, Lynn Smith-Lovin, and James M Cook. "Birds of a feather: Homophily in social networks". In: *Annual review of sociology* (2001), pp. 415–444.

[101] Nicos Mouzelis. *Sociological theory: what went wrong?: diagnosis and remedies*. Routledge, 2003.

[102] Sebastian A Müller et al. "Predicting the effects of COVID-19 related interventions in urban settings by combining activity-based modelling, agent-based simulation, and mobile phone data". In: *PloS one* 16.10 (2021), e0259037.

[103] David G Myers. "Polarizing effects of social interaction". In: *Group decision making* 125 (1982), pp. 137–138.

[104] Martial L Ndeffo Mbah et al. "The impact of imitation on vaccination behavior in social contact networks". In: *PLoS computational biology* 8.4 (2012), e1002469.

[105] Radford M Neal et al. "MCMC using Hamiltonian dynamics". In: *Handbook of markov chain monte carlo* 2.11 (2011), p. 2.

[106] Shunjiang Ni, Wenguo Weng, and Hui Zhang. "Modeling the effects of social impact on epidemic spreading in complex networks". In: *Physica A: Statistical Mechanics and its Applications* 390.23-24 (2011), pp. 4528–4534.

[107] Pierre Nouvellet et al. "The role of rapid diagnostics in managing Ebola epidemics". In: *Nature* 528.7580 (2015), S109–S116.

[108] Hendrik Nunner, Vincent Buskens, and Mirjam Kretzschmar. "A model for the co-evolution of dynamic social networks and infectious disease dynamics". In: *Computational Social Networks* 8.1 (2021), pp. 1–33.

[109] Eli Pariser. *The filter bubble: What the Internet is hiding from you*. penguin UK, 2011.

[110] Romualdo Pastor-Satorras and Alessandro Vespignani. "Epidemic spreading in scale-free networks". In: *Physical review letters* 86.14 (2001), p. 3200.

[111] Romualdo Pastor-Satorras et al. "Epidemic processes in complex networks". In: *Reviews of modern physics* 87.3 (2015), p. 925.

[112] Ana Perisic and Chris T Bauch. "A simulation analysis to characterize the dynamics of vaccinating behaviour on contact networks". In: *BMC Infectious Diseases* 9 (2009), pp. 1–15.

[113] Ana Perisic and Chris T Bauch. "Social contact networks and disease eradicability under voluntary vaccination". In: *PLoS computational biology* 5.2 (2009), e1000280.

[114] Marcelo A Pires, André L Oestereich, and Nuno Crokidakis. "Sudden transitions in coupled opinion and epidemic dynamics with vaccination". In: *Journal of Statistical Mechanics: Theory and Experiment* 2018.5 (2018), p. 053407.

[115] Rafael Prieto Curiel and Humberto González Ramírez. "Vaccination strategies against COVID-19 and the diffusion of anti-vaccination views". In: *Scientific Reports* 11.1 (2021), p. 6626.

[116] Heiko Rauhut. "Beliefs about lying and spreading of dishonesty: Undetected lies and their constructive and destructive social dynamics in dice experiments". In: *PloS one* 8.11 (2013), e77878.

[117] Timothy C Reluga, Chris T Bauch, and Alison P Galvani. "Evolving public perceptions and stability in vaccine uptake". In: *Mathematical biosciences* 204.2 (2006), pp. 185–198.

[118] Sebastián Risau-Gusmán and Damián H Zanette. "Contact switching as a control strategy for epidemic outbreaks". In: *Journal of theoretical biology* 257.1 (2009), pp. 52–60.

[119] Marcel Salathé and Sebastian Bonhoeffer. "The effect of opinion clustering on disease outbreaks". In: *Journal of The Royal Society Interface* 5.29 (2008), pp. 1505–1508.

[120] Piotr Sapiezynski et al. "Interaction data from the copenhagen networks study". In: *Scientific Data* 6.1 (2019), p. 315.

[121] Frank Schlosser et al. “COVID-19 lockdown induces disease-mitigating structural changes in mobility networks”. In: *Proceedings of the National Academy of Sciences* 117.52 (2020), pp. 32883–32890.

[122] Birgitt Schönfisch and André de Roos. “Synchronous and asynchronous updating in cellular automata”. In: *BioSystems* 51.3 (1999), pp. 123–143.

[123] P Wesley Schultz et al. “The constructive, destructive, and reconstructive power of social norms”. In: *Psychological science* 18.5 (2007), pp. 429–434.

[124] Anupama Sharma et al. “Epidemic prevalence information on social networks can mediate emergent collective outcomes in voluntary vaccine schemes”. In: *PLoS computational biology* 15.5 (2019), e1006977.

[125] Baike She et al. “On a networked SIS epidemic model with cooperative and antagonistic opinion dynamics”. In: *IEEE Transactions on Control of Network Systems* (2022).

[126] Shazia Sheikh et al. “A report on the status of vaccination in Europe”. In: *Vaccine* 36.33 (2018), pp. 4979–4992.

[127] Tianyu Shi et al. “Effects of asymptomatic infection on the dynamical interplay between behavior and disease transmission in multiplex networks”. In: *Physica A: Statistical Mechanics and its Applications* 536 (2019), p. 121030.

[128] Dietrich Stauffer. “Difficulty for consensus in simultaneous opinion formation of Sznajd model”. In: *Mathematical Sociology* 28.1 (2004), pp. 25–33.

[129] Daniel Strömbom et al. “Asynchrony induces polarization in attraction-based models of collective motion”. In: *Royal Society open science* 6.4 (2019), p. 190381.

[130] Gui-Quan Sun et al. “Pattern transitions in spatial epidemics: Mechanisms and emergent properties”. In: *Physics of life reviews* 19 (2016), pp. 43–73.

[131] Charles S Taber, Damon Cann, and Simona Kucsova. “The motivated processing of political arguments”. In: *Political Behavior* 31.2 (2009), pp. 137–155.

[132] Károly Takács, Andreas Flache, and Michael Mäs. “Discrepancy and disliking do not induce negative opinion shifts”. In: *PloS one* 11.6 (2016), e0157948.

[133] Edgar Treischl and Tobias Wolbring. “The past, present and future of factorial survey experiments: A review for the social sciences”. In: *Methods, data, analyses* 16.2 (2022), p. 30.

[134] Stephen Park Turner and Jonathan H Turner. *The impossible science: An institutional analysis of American sociology*. SAGE Publications, Inc, 1990.

[135] Lucas D Valdez, Pablo A Macri, and Lidia A Braunstein. “Intermittent social distancing strategy for epidemic control”. In: *Physical Review E* 85.3 (2012), p. 036108.

[136] Raffaele Vardavas, Romulus Breban, and Sally Blower. “Can Influenza Epidemics Be Prevented by Voluntary Vaccination?” In: *PLOS Computational Biology* 3.5 (2007), e85. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.0030085. (Visited on 10/19/2023).

[137] Fátima Velásquez-Rojas et al. “Disease and information spreading at different speeds in multiplex networks”. In: *Physical Review E* 102.2 (2020), p. 022312.

[138] Paulo Cesar Ventura et al. “Modeling the effects of social distancing on the large-scale spreading of diseases”. In: *Epidemics* 38 (2022), p. 100544.

[139] Frederik Verelst, Lander Willem, and Philippe Beutels. “Behavioural change models for infectious disease transmission: a systematic review (2010–2015)”. In: *Journal of The Royal Society Interface* 13.125 (2016), p. 20160820.

[140] Joel Wagner et al. “Societal feedback induces complex and chaotic dynamics in endemic infectious diseases”. In: *medRxiv* (2023), pp. 2023–05.

[141] Jinming Wan et al. “Multilayer networks with higher-order interaction reveal the impact of collective behavior on epidemic dynamics”. In: *Chaos, Solitons & Fractals* 164 (2022), p. 112735.

[142] Wei Wang et al. “Asymmetrically interacting spreading dynamics on complex layered networks”. In: *Scientific reports* 4.1 (2014), p. 5097.

[143] Xinyu Wang et al. “Vaccination behavior by coupling the epidemic spreading with the human decision under the game theory”. In: *Applied Mathematics and Computation* 380 (2020), p. 125232.

[144] Zhen Wang et al. “Coupled disease–behavior dynamics on complex networks: A review”. In: *Physics of life reviews* 15 (2015), pp. 1–29.

[145] Dale Weston, Katharina Hauck, and Richard Amlôt. “Infection prevention behaviour and infectious disease modelling: a review of the literature and recommendations for the future”. In: *BMC public health* 18.1 (2018), pp. 1–16.

[146] Richard Whitley et al. *The intellectual and social organization of the sciences*. Oxford University Press on Demand, 2000.

[147] Stephen Wolfram. “Cellular automata as models of complexity”. In: *Nature* 311.5985 (1984), pp. 419–424.

[148] Qingchu Wu et al. “The impact of awareness on epidemic spreading in networks”. In: *Chaos: an interdisciplinary journal of nonlinear science* 22.1 (2012), p. 013101.

[149] Shang Xia and Jiming Liu. “A belief-based model for characterizing the spread of awareness and its impacts on individuals’ vaccination decisions”. In: *Journal of The Royal Society Interface* 11.94 (2014), p. 20140013.

[150] Shang Xia and Jiming Liu. “A computational approach to characterizing the impact of social influence on individuals’ vaccination decision making”. In: *PloS one* 8.4 (2013), e60373.

[151] Haidong Xu, Ye Zhao, and Dun Han. “The impact of the global and local awareness diffusion on epidemic transmission considering the heterogeneity of individual influences”. In: *Nonlinear Dynamics* 110.1 (2022), pp. 901–914.

[152] Damián H Zanette and Sebastián Risau-Gusmán. “Infection spreading in a population with evolving contacts”. In: *Journal of biological physics* 34 (2008), pp. 135–148.

[153] Hai-Feng Zhang et al. “Braess’s paradox in epidemic game: better condition results in less payoff”. In: *Scientific reports* 3.1 (2013), pp. 1–8.

7 Supplementary Information

7.1 Information to reproduce Figure 1

7.1.1 Heterogeneous susceptibility (panels a to d)

Figure 1a-d illustrate the effect of agent heterogeneity concerning how much agents engage in self-protective behavior. To this end, we generated three agent populations with the same average but different variance of susceptibility β_i. Each population is composed by 31×31 agents assigned to cells in a regular cellular automaton with von Neumann neighborhoods. Each agent had a susceptibility β_i drawn from a symmetrical beta distribution with $a = b = 100$ in panel a to obtain an unimodal distribution of the susceptibility, with $a = b = 1$ in panel b to obtain a uniform distribution of susceptibilities, and $a = b = 0.01$ in panel c to obtain a polarized, bimodal distribution of susceptibilities.After that, the central agent was infected and a classical SIR dynamic with fixed recovery rate $\gamma = 0.1$, as described in [111] was simulated. For panels a, b, and c, we selected runs where the

number of recovered agents in equilibrium equaled the median number of recovered agents in 1,000 independent realizations.

7.1.2 Behavioral dynamics (panel e)

In Fig. 1e, we show different dynamics of protective behavior observed in Denmark and the United Kingdom during the COVID-19 pandemic. Data was collected by the COVID-19 behavior tracker [78], a global survey to monitor protective behavior. The behavior scale was calculated as the average of 29 questionnaire items (*i12_health_1* to *i12_health_29*), where each item assessed whether participants engaged in a protective behavior (e.g., mask-wearing, avoiding contact) on a 5-point scale ranging from "always" (coded 5) to "not at all" (coded 1). In every measurement wave, there were at least 945 participants per country. Results are qualitatively the same when only the 16 items are included that were measured in all measurement waves. Each bar shown in the figure is a box plot, where black bars represent the range between the 5^{th} and the 95^{th} percentile. White bars show the interquartile range. The green line indicates the median.

7.2 Derivation of influence-response functions

In this section, we describe the *singe-step* Monte Carlo, the assumptions we made to simplify the models in Fig. 3, 4, and 5, as well as the assumption we made about the state distribution of the other agents $P\left(X_{i\neq j}\right)$.

7.2.1 Single-step Monte Carlo

In this subsection we will describe the algorithm to approximate the expected value of the IRF calculated for the social stimulus s_i:

$$\mathbb{E}_b\left[\text{IRF}\left(b, s_i, x_i\right)\right]. \tag{12}$$

In Eq. 8, we showed that the IRF depends on the probability distribution of the states of all the other agents; furthermore, this probability distribution evolves in time, so approximating it numerically is beyond the scope of this review. Moreover, that would not help with our objective of easily comparing the IRFs of two different models. For this purpose, we expose agent i to a fixed distribution of states $P\left(\mathbf{x}_{j\neq i}\right)$.

For each realization l of the state of the system $\{\mathbf{x}_l\}$, we calculate the corresponding signal $s_i = f_i\left(\mathbf{x}_l\right)$, and the resulting behavior $b_{i,l}$. The average behavior is defined as the average of all behaviors generated by a signal s_i. If the signal s_i is a continuous variable, one can discretize the s-space into bins of the same width.

Summarizing the steps:

1. Make an assumption about how $P\left(\mathbf{X}\right)$ looks like.
2. Generate L realizations of $\mathbf{X}_l$.
3. For each of these realizations, calculate $s_{i,l}$, and b_l. If the signal is continuous, discretize the signal into bins.
4. Calculate the average behavior for the j-th bin of s as $\frac{1}{|\mathcal{L}_j|}\sum_{l\in\mathcal{L}_j} b_l$, where $\mathcal{L}_j$ is the set of samples where the measured signal falls into the j-th bin.

7.2.2 Threshold models (Figure 3)

In Fig. 3 we considered discrete, binary behavior with $\Omega_B = \{0, 1\}$. Panels a and b show two typical IRFs that are discussed in the original paper of Lopez-Pintado and Watts, [92]. First, in

upward threshold models an agent adopts the behavior if a sufficient fraction of neighbors shows the behavior, like in the threshold models by Granovetter [66]. Second, in downward threshold models the behavior becomes less likely if many others already show it capturing effects like free-riding. Panel c of Fig. 3 shows a combination of the two similar to the model by [150].

The only information agent i needs to decide whether to adhere to behavior $b_i^{t+1} = 1$ is the fraction of neighbors expressing the same behavior. Therefore, the observable f_i is equal to

$$s_i := f_i(\mathbf{b}) = \frac{1}{k_i} \sum_{j=1}^{N} A_{ij} b_j, \tag{13}$$

where k_i is the degree of node i and A_{ij} the adjacency matrix of the social network. The IRF may be implemented as a sharp threshold function or as a soft transition often realized by a logistic function in the literature. In general, such an IRF can then be written as

$$\mathrm{IRF}_{\mathrm{LP}}(b_i^{t+1} = 1, s_{t,i}) \equiv P\left(B_i^{t+1} = 1 \mid s_i^t\right) = \frac{1}{1 + e^{-\lambda(s_i - \tau)}}, \tag{14}$$

where τ is the threshold value which has been set to $\tau = 1/2$ in Fig. 3. The parameter λ determines the slope of the logistic function: a small absolute value of λ corresponds to a gradual transition and a large absolute value to a sharp step. Furthermore, a positive value of λ leads to an upward logistic function, as in Fig. 3a, and a negative value to a downward logistic function, as in Fig. 3b. Specifically, we used $\lambda = \pm 10$ (light blue), ± 15, ± 30, ± 50 and ± 100 (dark blue), with positive values in panel a and negative in panel b.

Fig. 3c is the combination of these two basic models. For instance, in the setting of [150] the choice to get a vaccine depends in two complementary ways on the vaccination decisions of others, because agents decide for or against a vaccine depending on an assessment of personal costs (downward threshold, free riding) and social impact (soft upward function, conformity) through the behavioral choices of others. Such a decision rule can be captured by an upward threshold model with a low absolute value of $\lambda_{\mathrm{up}} > 0$, and a downward one with a high absolute value of $\lambda_{\mathrm{down}} < 0$:

$$\mathrm{IRF}(b_i^{t+1} = 1, s_i) = \frac{p}{1 + e^{-\lambda_{\mathrm{up}}(s_i - \tau)}} + \frac{1 - p}{1 + e^{-\lambda_{\mathrm{down}}(s_i - \tau)}}, \tag{15}$$

where p is a parameter governing the relative importance of the two mechanisms.

7.2.3 Assimilation, reinforcement, and repulsion (Figure 4)

In Fig. 4, we compare assimilation, reinforcement, and repulsion, the three classical models of opinion dynamics [50]. In each time step, the opinion of an agent i is updated after a comparison with a random connected agent j according to

$$b_i^{t+1} = b_i^t + \mu g\left(b_i^t, b_j^t\right), \tag{16}$$

where $g(b_i, b_j)$ is a function which captures how the opinion changes depending on the opinions of both agents, and $\mu \in [0, 1]$ determines the size of the update.

For assimilation (panels a, d and g), it is equal to

$$g\left(b_i, b_j\right) = b_j - b_i. \tag{17}$$

This means that agent i approaches the opinion of j in proportion to the distance between the two. These averaging models are amenable to analytical solutions and their convergence properties are well-understood [40, 51, 52].

In order to compute the expected update for an agent with a given opinion (panel d), we use the

average opinion of the neighbors of agent i as an observable social signal:

$$s_i := f_i(\mathbf{b}) := \frac{1}{k_i} \sum_{j=1}^{N} A_{ij} b_j. \tag{18}$$

For this class of models based on averaging, it is possible to show that the expected updated behavior of i is given by

$$\mathbb{E}[B_{t+1,i} \mid b_i^t, s_i^t] = (1-\mu) b_i^t + \mu s_i^t. \tag{19}$$

That is, agent i will move closer to the average opinion in its neighborhood and μ decides how much weight is given to its current opinion.

For reinforcement (panel b, e and h) and repulsion (panel c, f and i) the opinion update functions g are nonlinear. Reinforcement is realized, for instance, by

$$g(b_i, b_j) = \begin{cases} +1 & \text{if } b_i > \frac{1}{2} \text{ and } b_j > \frac{1}{2} \\ -1 & \text{if } b_i < \frac{1}{2} \text{ and } b_j < \frac{1}{2} \\ \frac{1}{2}(b_j - b_i) & \text{otherwise.} \end{cases} \tag{20}$$

Here, we assume that an opinion value of 1/2 defines a neutral opinion dividing the opinion scale into "in favor" ($b_i > 1/2$) and "opposed" ($b_i < 1/2$), as done in [3, 114, 141]. If the two agents are on the same side of the opinion scale, agents move further towards a more extreme opinion. The magnitude of this effect is equal to μ.

Models of repulsion assume that agents are drawn further apart if their opinions are very different. These models typically contain a threshold ϵ which is compared to the distance $|b_j - b_i|$ between the two agents in opinion space. That is, they also integrate assimilative influence (as specified above) when their opinion distance is smaller than ϵ. Negative influence and repulsion set in when $|b_j - b_i| > \epsilon$ so that g can be written as

$$g\left(b_i^t, b_j\right) = \begin{cases} b_j - b_i & \text{if } |b_j - b_i| \le \epsilon \\ b_i - b_j & \text{if } |b_j - b_i| > \epsilon \\ 0 & \text{otherwise.} \end{cases} \tag{21}$$

Despite (d) being analytically solvable, to ensure that panels d, e and f in Fig. 4 are comparable, we followed the single-step Monte Carlo we described above (SI section 2.1) for all three of them. We binned the signal space in 100 bins from 0.05 to 0.95, and for each of these bin we extracted 10^5 neighbors iid in an interval of radius 0.05 from the center of the bin. For each of these trials, we calculated the updated value of the opinion of agent i, and took its average.

In Fig. 4d, we set $b_i^t = 0.2$, and $\mu = 0.5$. In Fig. 4e, we set $b_i^t = 0.2$, $\mu = 0.1$, and finally in Fig. 4f, we set $b_i^t = 0.2$, $\epsilon = 0.25$, and $\mu = 0.5$.

7.3 Coupled models

In this section, we describe the steps we took to obtain four classical models from the literature in Fig. 5. For each model, we made the assumption that the susceptibility to the disease is $\beta = 0.1$. Additionally, in panels a and b of Fig. 5, we assumed that a third of the individuals are vaccinated. Conversely, in panel c, we varied the fraction of vaccinated from $s^{(\mathrm{V})} = 0.25$ to $s^{(\mathrm{V})} = 0.99$. In panels d and e of Fig. 5, we assumed that a third of the individuals are infected, while in panel f, we varied the fraction of infected from $s^{(\mathrm{I})} = 0.25$ to $s^{(\mathrm{I})} = 0.99$.

The results presented were derived using the single-step Monte Carlo approach outlined earlier. For the simulations, we consider simplified versions of the models in which neighbors were generated deterministically. Consequently, there was no need to average across multiple trials as we generated the neighbors only once. In the following sections, we will explain the assumptions we made to

simplify each model and to generate a reasonable distribution of neighbors.

7.3.1 Bauch 2005

In Bauch 2005 [18], every agent has the freedom to decide whether to vaccinate or not. Opting for vaccination incurs a cost, denoted as C_V. If the agent decides not to vaccinate and avoids the infection, no cost will incur, but if the decision to forego vaccination results in an infection, the agent will pay a higher cost $C_I > C_V$. The model assumes that the vaccination is perfect, meaning that an agent cannot get infected when vaccinated. At the end of an epidemic wave, each agent will have paid one of three costs: 0 if it was neither vaccinated nor infected, C_I if it wasn't vaccinated and got infected, and C_V if it was vaccinated.

The model considers multiple successive waves of the same disease, with agents reassessing their vaccination decisions after each wave. To make a decision, an agent i compares the cost it paid in the previous wave with that of a random neighbor j in the same wave. The probability of adopting the behavior b_j of the neighbor is obtained via a logistic map (often referred to as Fermi function in the literature) of the difference of costs

$$P\left(b_i \to b_j\right) = \frac{1}{1 + e^{-\lambda(b_j - b_i)}}. \tag{22}$$

To simplify the model, we made one main assumption: we set $\lambda = \infty$, so that agent i will always copy the behavior of agent j if that behavior led to a lower cost for agent j.

In Fig. 5a, we fixed the fraction of vaccinated neighbors to $s^{(\mathrm{V})} = 1/3$, and we varied the fraction of infected neighbors from $s^{(\mathrm{I})} = 0$ to $2/3$; these two assumptions fully determine the important aspects of the states of the other agents $\mathbf{X}_{j\neq i}$. In this scenario, the probability that agent i will vaccinate in the next wave depends on the behavior of agent i in the previous wave as well as on the fraction of infected and vaccinated neighbors in the previous wave.

The probability that an agent who was vaccinated in the previous time step will vaccinate again is equal to the probability that such agent will interact with a vaccinated or with a recovered agent. Conversely, if the agent was not vaccinated, it will not vaccinate again if it did not develop an infection; if it did, it will vaccinate with a probability equal to the fraction of vaccinated agents.

In Fig. 5a, we show the IRF of this model with respect to the fraction of infected neighbors $s^{(\mathrm{I})}$ and a fixed fraction of vaccinated agents $s^{(\mathrm{V})} = 1/3$:

$$\mathrm{IRF}\left(V', s^{(\mathrm{I})}, x\right) = \begin{cases} 0 & \text{if } x_i = S \\ s^{(\mathrm{V})} & \text{if } x_i = R \\ s^{(\mathrm{V})} + s^{(\mathrm{I})} & \text{if } x_i = V, \end{cases} \tag{23}$$

where V' represents that agent i will decide to vaccinate, and $x_i \in \{V, S, R\}$ is the state of agent i at the end of the previous wave.

Similarly, in Fig. 5d, we show the IRF of this model with respect to the fraction of vaccinated neighbors and a fixed fraction of infected agents, $s^{(\mathrm{I})} = 1/3$:

$$\mathrm{IRF}\left(V', s^{(\mathrm{V})}, x_i\right) = \begin{cases} 0 & \text{if } x_i = S \\ s^{(\mathrm{V})} & \text{if } x_i = R\,, \\ s^{(\mathrm{V})} + s^{(\mathrm{I})} & \text{if } x_i = V \end{cases} \tag{24}$$

which is formally the same, with the only difference that now $s^{(\mathrm{I})}$ is kept constant, and $s^{(\mathrm{V})}$ varied from 0 to $2/3$.

In Fig. 5a and c, we plot $\mathrm{IRF}\left(V', s^{(\mathrm{I/V})}, \mathrm{S}\right)$ with a dashed orange line, $\mathrm{IRF}\left(V', s^{(\mathrm{I/V})}, \mathrm{R}\right)$ with a solid orange line, and $\mathrm{IRF}\left(V', s^{(\mathrm{I/V})}, V\right)$ with a dark red line.

7.3.2 Granell 2014

In Granell 2014 [63], an awareness of the disease spreads in parallel to the disease itself through the population. The disease model is the classical SIS model, in which infected agents immediately return to the pool of susceptibles upon recovering. The behavior dynamic follows a similar mechanism called UAU (Unaware-Aware-Unaware), where unaware agents can become aware with probability $\zeta = 0.025$ either by contracting the disease or by being convinced by other aware agents, and aware agents return to the unaware status with probability $\theta = 0.15$. Aware agents are assumed to engage in more self protecting behavior, leading to a lower susceptibility β.

The other simplification to the model we made to arrive at Fig. 5b is to fix the fraction of unaware susceptible individuals to 3/4. Here we chose as signal the fraction of infected neighbors $s^{(\mathrm{I})}$. This allows us to write

$$\mathrm{IRF}\left(A, s^{(\mathrm{I})}, x_i\right) = \begin{cases} 1-\theta & \text{if } x_i = \mathrm{A} \\ 1-(1-\zeta)^{k_i\left(1+\frac{3}{4}s^{(\mathrm{I})}\right)} & \text{if } x_i = U, \end{cases} \tag{25}$$

where k_i is the number of neighbors of agent i.

In Fig. 5b, we show $\mathrm{IRF}\left(A, s^{(\mathrm{I})}, U\right)$ with an orange line, and $\mathrm{IRF}\left(A, s^{(\mathrm{I})}, A\right)$ with dark red.

7.3.3 Alvarez-Zuzek 2017

In Alvarez-Zuzek 2017 [3], each agent interacts with its neighbors by exchanging opinions about vaccination according to the M-model [86] with $M = 2$. Each agents' opinion ranges from -2, strongly against the vaccine, to $+2$, strongly in favor. If two agents share the same opinion (i.e., if their opinions have the same sign), they reinforce each other, while if their opinions are different, they interact assimilatively. Once an agent reaches $M = 2$, it will vaccinate. Its opinion will not change again unless the vaccine fails and the agent becomes infected; in this case, the agent will develop a maximally negative opinion $M = -2$ towards vaccines. The vaccines are not perfect in this model, but simply reduce susceptibility of each agent by a factor of ω, bringing it to $\beta = \beta_0\omega$.

The first main simplification we did was to simplify the opinion dynamic. We set the opinion of each agent to a value between 0 and 1 and the probability of each agent to vaccinate to be equal to the local average opinion in i's neighborhood. Then we assumed that each unvaccinated agent has a fixed opinion $b = 1/4$, each vaccinated healthy agent has a fixed maximally positive opinion $b = 1$, and each vaccinated and infected agent has a maximally negative opinion $b = 0$. In this simplified scenario, the vaccination probability of agent i is

$$P\left(V \mid k_i, k_i^{(\mathrm{U})}, k_i^{(\mathrm{VS})}\right) = \frac{1}{k_i}\left[\frac{1}{4}k_i^{(\mathrm{U})} + k_i^{(\mathrm{VS})}\right], \tag{26}$$

where k_i is the total number of neighbors of i, $k_i^{(\mathrm{U})}$ is the number of unvaccinated neighbors, regardless of the infection status, and $k_i^{(\mathrm{VS})}$ is the number of vaccinated and susceptible neighbors.

Since we want to be able to compare how this model responds to the fraction of infected neighbors, Fig. 5c, and the fraction of vaccinated neighbors, Fig. 5f, we need to make an assumption on the average number of vaccinated and infected neighbors for each value of $s^{(\mathrm{I})}$ in (c) and of $s^{(\mathrm{V})}$ in (f).

To approximate this distribution, we assume that all neighbors are connected to all other neighbors, i.e., that the network is a complete graph. In such a system, the probability that a single agent becomes infected, given that n other agents are already infected, is equal to

$$P\left(I \mid b\right) = 1 - \left(1 - \beta(b)\right)^n, \tag{27}$$

where $b \in \{U, V\}$ is the vaccination status, and the susceptibility is equal to $\beta(U) = \beta_0$ if the agent is unvaccinated and $\beta(V) = \beta_0\omega$ if the agent is vaccinated.

We initialize the system by setting the health status of all agents to susceptible and assign their vaccination status such that the global fraction of vaccinated agents is equal to $s^{(\mathrm{V})}$. Then, we iterate over $n = 1, \ldots$ and infect an agent in every step. The first agent to be infected is always unvaccinated. After that, we infect either an unvaccinated susceptible agent with probability

$$\frac{P(I \mid U)\, s_n^{(\mathrm{US})}}{P(I \mid U)\, s_n^{(\mathrm{US})} + P(I \mid V)\, s_n^{(\mathrm{VS})}}, \tag{28}$$

or a vaccinated susceptible agent with probability

$$\frac{P(I \mid V)\, s_n^{(\mathrm{VS})}}{P(I \mid U)\, s_n^{(\mathrm{US})} + P(I \mid V)\, s_n^{(\mathrm{VS})}} \tag{29}$$

where $s_n^{(\mathrm{VS})}$ and $s_n^{(\mathrm{US})}$ are the fractions of vaccinated susceptible and unvaccinated susceptible agents, respectively, in the n'th iteration. We then repeat until the desired share of agents is infected.

7.3.4 Ndeffo Mbah et al.

In Ndeffo Mbah et al. [104], there are two populations of agents. Agents in the first group decide whether to vaccinate based on the perceived risk of infections, and those in the second imitate the behavior of their surrounding agents. In Fig. 5e, we assume a simplified case in which the agent of interest belongs to the first group. We also assume that the agent knows what fraction of other agents was vaccinated in the previous wave, $s^{(\mathrm{V})}$, and how many of the unvaccinated got infected, $s^{(\mathrm{I})} = 1/3$.

Following the idea of the paper, we fix the cost for vaccination to C_V and the cost for infection to C_I. The perceived cost for infection is obtained by multiplying C_I by the perceived probability of becoming infected: $P_I\left(s^{(\mathrm{V})}\right) = 1 - (1-\beta)^{N\left(1-s^{(\mathrm{V})}\right)s^{(\mathrm{I})}}$. If the cost for vaccinating is higher than the perceived cost of infection, the agent will decide not to vaccinate, and vice versa. The IRF with respect to the observable $s_i = s^{(\mathrm{V})}$ is

$$\mathrm{IRF}\left(V, s^{(\mathrm{V})}, x_i\right) = \begin{cases} 1 & \text{if } C_V < C_I - C_I\,(1-\beta)^{N\left(1-s^{(\mathrm{V})}\right)s^{(\mathrm{I})}} \\ 0 & \text{otherwise.} \end{cases} \tag{30}$$