RESEARCH ARTICLE

A Tragic Coalition of the Rational and Irrational: A threat to collective responses to COVID-19

Marinus Ferreira[a], Marc Cheong[b], Colin Klein[c], and Mark Alfano[a]

[a]Macquarie University, Sydney, Australia [b]University of Melbourne, Melbourne, Australia
[c]Australian National University, Canberra, Australia

ARTICLE HISTORY
Compiled September 7, 2022

ABSTRACT
There is not as much resistance to COVID-19 mitigation as there seems, but there are structural features that make resistance seem worse than it is. Here we describe two ways that the problem *seeming* to be worse than it is can *make* it worse. First, visible hesitation to implement COVID-19 responses signals to the wider society that mitigation measures may not succeed, which undermines people's conditional willingness to join in on those efforts. Second, our evaluations of others' willingness to implement these measures are informed by our attempts to mind-read them. Yet attempts to mind-read groups often mislead us, because groups invariably act from diverse motives whereas mind-reading works best when identifying relatively stable and consistent motivations. This means that a small minority of people refusing to implement measures can have an outsized prominence that prompts mind-reading to diagnose widespread hesitation. These two factors form a feedback loop with each other: we see some people's hesitation, which prompts us to mind-read other people as being more uncertain about the responses than they actually are, which undermines our confidence in the responses, which in turn encourages others to mind-read this hesitation, which further undermines confidence.

KEYWORDS
COVID-19, pandemic, coordination, irrationality, mind-reading, defection cascades

1. Introduction

Vaccination and vaccine mandates, social distancing regimes, and restrictions on business and travel all form part of public health responses to the COVID-19 pandemic. These are all examples of *effortful cooperation*, where individuals perform actions that are personally costly, but that contribute to a collective behavior that hopefully has larger mutual benefit as an outcome.

Each of these public health measures have also been met with resistance that goes past the usual disagreements on best policy and becomes apparently unconditional, such that the resisters claim that such mitigations are inherently harmful, or malicious, or advanced with deceitful intent, making cooperation out of the question. This resistance poses a social dilemma, since effortful cooperation is vulnerable to individuals refusing to participate, which makes beneficial outcomes less likely. This paper offers a diagnosis of how effortful cooperation can fail even when most people are willing

CONTACT Author. Email: marinus.ferreira@mq.edu.au

to cooperate. We show how a small but visible minority of hesitant actors can undermine the general confidence that effortful cooperation will succeed, and can do so out of proportion to how many people actually do hesitate. By undermining the general confidence in success, success becomes impossible.

Such failures lead to what appears to be a *collectively irrational* failure to adopt mitigation measures: a large majority want such measures but the whole does not adopt them. The focus in academic and public discussions often centers around this collective irrationality.

Yet we hesitate to talk about irrationality *tout court*, whether individual or collective, for two reasons. First, what matters is that there are people who change their behavior because they diagnose enough people around them as unconditionally or near-unconditionally unwilling to cooperate. Irrationality only enters the picture because if you diagnose someone as irrational, it fatally undermines your confidence that you could recruit them into effortful cooperation. We go on to argue that having visible examples of apparent irrationality can push us to overdiagnose (near-)unconditional refusal to cooperate. We are concerned with *apparent* rationality—that is, diagnosing irrationality *rightly or wrongly.* Talk of rationality as such requires arguing for some standard of rationality and arguing that some collective behavior fails to meet it. We do not commit to any such standard, and do not need to. We are less interested in whether the collective response is actually irrational; we are interested in the surprising fact that groups can end up in places that most of their members feel is irrational. The mechanisms we describe here work whether the irrationality is genuine or not.

Second, by focusing on COVID-19 mitigation measures, we include many proposed measures without considering whether they are in fact rational or irrational at either an individual or collective level. We take it as read that at least some of these measures are things we should do, even taking into account the costs of doing so. But the uncertainty around effective measures means that there are actual proposals that turned out to be neither sensible nor useful. For instance, early in the pandemic, restrictions on movement and contact included bans on meeting even out-of-doors. We have since learnt that the risks of meeting out-of-doors are probably marginal. Following such a directive now does not seem obviously rational. Similarly, there have been serious suggestions that even if lockdowns were appropriate in many parts of the world, they were inappropriate for regions such as sub-Saharan Africa: the combination of younger populations and higher economic and welfare costs do not obviously favor lockdowns. (Broadbent 2020; see Allais and Venter 2020 for an alternative take). So we stress that we are concerned with cases where people may fail to adopt mitigation measures even when they mostly think doing so *would* be rational, conditional on other people adhering as well.

We examine the link between apparent irrationality and effortful cooperation by considering two related arguments. The first argument, drawn from work done on game-theoretic approaches to the COVID-19 pandemic, is the *counterproductive signals argument*:

Effective responses to the COVID-19 pandemic require effortful cooperation.

Effortful cooperation requires confidence that other people will also make the effort to cooperate.

If we see other people refuse to cooperate even when refusal appears to be irrational, we lose confidence that a sufficient number of people will also make the effort to cooperate.

∴ Effective responses to the COVID-19 pandemic are threatened by the appearance of irrationality.

The second argument, taking the conclusion of the first argument as one of its premises, is the *overdiagnosis of irrationality argument*:

Effective responses to the COVID-19 pandemic are made harder by the appearance of irrationality.
The difficulty in mind-reading groups increases the appearance of irrationality.
The more apparent irrationality there is, the more effective responses are threatened.
∴ Effective responses to the COVID-19 pandemic are made harder by the difficulty in mind-reading groups.

The first argument contributes to a recent literature that frames responses to the COVID-19 pandemic in terms of other familiar social dilemmas, in this case, *stag hunt* dilemmas. In stag hunt dilemmas, we can either settle for a low-effort outcome that we can secure ourselves, or aim at a high-effort outcome that requires effortful cooperation. For it to make sense to commit to effortful cooperation, we need confidence that enough other people will also do so. If we find it too hard to justify this confidence, then we are likely to defect from effortful cooperation. And the more people there are who defect for this reason, the more likely it is that someone else will defect as well. This becomes a *defection cascade* as people give up on the high-effort outcome and settle for the low-effort but less preferred outcome.

The overdiagnosis of irrationality argument introduces a new dimension to the problem. Since the likelihood that an individual will defect depends on their judgment about whether *others* will be hesitant, what matters in the first instance is not the actual likelihood of hesitation but the *apparent* likelihood of hesitation. Structurally the situation is similar to the a Keynesian beauty contest, in which people attempt to guess the modal guess as to the modal response.

When working through this problem, we do so through a process of ascribing mental states to other people, or *mind-reading*. We will show that mind-reading groups can go astray in systematic ways, and that these make more likely that we overdiagnose irrationality when looking at groups rather than individuals. If irrationality seems to be more prevalent than it actually is, we are more likely to defect from high-effort cooperation, and more likely to end up in inferior outcomes by our own lights.

Our approach is calibrated to explain situations where most people are conditionally willing to conform to COVID-19 mitigation measures, but there are reasons other than their own willingness that may make them hesitate to conform. We think this is a useful case to consider, because the failure of some interventions in the COVID-19 crisis is all the more striking given that public support for wide-ranging interventions is in fact very high. For instance, in Australia, one poll in May 2021 found that 73% of people supported mandatory vaccination (Smith, Attwell, and Evers, 2021) and another more fine-grained poll in September 2021 found that 62% supported mandatory vaccination for any kind of worker, and support went as high as 83% for mandates targeted at healthcare workers (Murphy, 2021). It is rare for contested government policy to receive that level of support, and more or less unheard-of for policies as onerous as mandatory vaccination. As a contrast, a poll in Australia showed only 60% support for costly interventions to mitigate climate change (Kassam and Léser, 2021). So, if we measure rationality by a conditional willingness to join effortful coordination, the question cannot be why responses to COVID-19 are less rational or popular than for other similar large-scale issues, because they are more rational and more popular. This makes us take a conditional willingness to adhere to large-scale responses to COVID-19 as our starting point.

2. Signalling in COVID-19 responses

We begin by examining the effects that apparent irrationality may have in prompting a *defection cascade*, which is where some individuals opt out of joining effortful cooperation, which in turn prompts more people to doubt the likely success of high-effort cooperation and defect in turn, creating a feedback loop until coordination collapses.

A defection cascade involves thresholds where individuals switch over from conditionally joining the effortful cooperation to conditionally defecting. In our example it is a threshold concerning the extent to which hesitancy has taken root in the wider population concerning the mitigation measure in question. We can and should see individuals as having different thresholds where they defect. The people who need the most confidence that others will cooperate before they join in themselves will be the first to defect. But their doing so makes it more likely that the next person is going to also defect, no matter how high or low their own threshold is. If this pattern continues to the point where not enough people remain to make the effortful cooperation succeed, then the community will fall from the high-effort equilibrium to the low-effort one, and the defection cascade will be complete.

Recent theoretic work by Quintana, Rosenstock, and Klein (2021) argues for an unexpected cause for defection cascades in public health contexts. Measures such as vaccinating and masking come with some (often nontrivial) personal cost. The benefits, however, are maximized only when a sufficiently large number of people take the measure. Recent experience with vaccination provides a familiar illustration: a vaccine need not be particularly effective at the individual level, so long as enough people are vaccinated to prevent widespread community transmission.

An individual deciding whether to take action often needs to know two things: the individual risks they bear, and the chance that sufficiently large numbers of other people will take the same action. This is a familiar sort of multi-equilibrium problem, and as with many such problems collectively salient information can play an important role (Schelling, 1960).

Quintana et al. (2021) argue that public health pronouncements can play this role. An official recommendation in favor of vaccination, for example, is seen by everyone, and everyone knows it is seen by everyone (and so on). A public health message thus carries both a first-order signal ("Vaccination is good") and a second-order signal ("People around you are likely to get vaccinated").

However, Quintana et al. (2021) note that this signalling can backfire. In particular, under conditions of uncertainty, public health officials may not know what the best course of action is. A natural instinct is to change messaging when new information comes in, or to accurately express the mixed scientific consensus on a proposed measure (consider in this regard mixed messaging around masking during much of 2020). In such cases, the first-order signal may be entirely accurate, but the effective second-order signal is that one cannot trust other people to coordinate on collective action—because in seeing mixed messages, they might well doubt the coordination power of the signal, as you do as well. This can happen even if everyone actually believes the truth of the accurate first-order message.

We describe this effect as a *counterproductive signal defection cascade*. A social situation may offer two equilibrium outcomes: a low-effort equilibrium that individuals can reach on their own, and a high-effort equilibrium that requires effortful cooperation but is a better outcome for everyone. We often use signals to coordinate our effortful cooperation. If there is such a signal with a first-order content that guides individuals to the high-effort equilibrium, but with a second-order content that some people are likely to hesitate in doing so, there is a risk that many people will lack sufficient confidence that effortful cooperation will succeed, and so would choose not to take part. The more people hesitate to adhere to the first-order content of a signal, the more salient its second-order content becomes. The more salient the second-order signal becomes, the more likely it is that the group will in fact opt for the low-effort rather than the high-effort equilibrium, because of their lack of confidence that the effortful cooperation will succeed.

3. Mind-reading groups

Counterproductive signal defection cascades are primarily an individual-level phenomenon, in the sense that it mostly involves trying to figure out what other individuals will do. When we evaluate the behaviors of individuals, one prominent and important way to do so is to first ascribe some states of mind to them—so-called ‘mind-reading’—and then evaluate the behavior in light of our ascription. Humans are extraordinarily good at mind-reading; even our failures occur against a backdrop of routine success. It is not surprising that humans are very good at mind-reading: we are thoroughly social beings who are deeply reliant on other people and our ability to understand them, and since human behavior varies widely, we need to have the ability to accurately judge across a respectively wide range. And we have that ability, as social beings so dependent on the judgements of others really ought to have (for a recent survey, see Spaulding 2018a).

As powerful and useful as the machinery of mind-reading is, we argue that it can lead us astray when we try to explain group behavior. To our knowledge there is as yet no systematic treatment of how groups can be the target of mind-reading. As noted by Spaulding (2018a), failures of mind-reading are an understudied area in general (for the view that there is not much to study, see Westra 2020), and what little exists on that topic deals with the central case of individuals ascribing mental states to other individuals. There is some work on how membership of a group may affect mind-reading (Spaulding, 2018b; Tullmann, 2019), which is a related but different question. Despite this gap in the literature, ascriptions of mental states to groups are ubiquitous. Very often these are in generic terms, where there is a profile that is generically applied to the members of a group, e.g. noting that people dislike being told what to do. Sometimes they are attempts to give a representative profile that is meant to correspond to the modal or median individual in the group (it is often unclear which), e.g. the journalistic practice of using *vox populi* or ’man on the street’ interviews as a way to describe public opinion. Sometimes they are attempts to describe some attitude distinctive of a group, even if that attitude is not a majority view but only more prevalent in that group than in comparative ones, such as saying that wine-drinkers are pretentious. All of these are attempts to mind-read a group by providing a profile that is meant to apply across the members in some informative way. Those attempts are of vital importance when individuals are judging things like the likelihood of their neighbors joining effortful cooperation.

Mind-reading can go astray on groups because, even if a behavior is universal across the members of a group (which is rare, and basically never the case for large groups), the same behavior can result from different sets of motivations. Following Ferreira

(2021), we can characterise an action as having both a *behavioral profile* and an *intentional profile*, where the former describes the bare movements displayed by an action, and the latter the motivations, excitations, sensitivities, etc., at work in the agent. The names of the two profiles are meant to evoke the traditional understanding of an action as a behavior that is performed with an intention. Here 'intention' is to be understood broadly to include all the occurrent psychological features at work when acting. These two profiles stand in a process/product relationship to each other, where the intentional profile is (part of) the process of acting through which the agent displays the behavioral profile. That is, the ascription of an intentional profile serves as an explanation of the behavior. It is this link that allows mind-reading to occur, as we relate the visible behavior to the states of the agent we understand to produce it. The link between intention and behavior is not one-to-one, since the same intention can lead to multiple different actions, such as when in shock we may either strike out or freeze (part of the so-called fight-flight-freeze-faint response), as Ryle (1949) described for multi-track dispositions, and the same behavior can result from many different intentions, as is the focus of a large literature on how individuals' economic behaviors underdetermine their preferences (Moscati, 2021).

However good we may be at mind-reading individuals, the social case is much harder. In the individual case, we need to allow for a number of *possible* different intentional profiles to ascribe; by contrast, in the social case, there will almost always be many different *actual* intentional profiles at play among the members of the group. Indeed, coordination can occur with remarkably little intentional overlap, as is illustrated by Lewis (1969)'s classic example of campers collecting firewood.

In Lewis' example, four people share a camp, and they need to regularly collect firewood. While each collects firewood individually, it is counterproductive for them to cover the same ground. They make an explicit agreement where one covers the area north of camp, one the east, another the south, and the last the west. This is a good arrangement that persists longer than the four campers stay together. One of them leaves, and a new camper takes their place. That new camper slots into the existing arrangement, since the other three continue covering their cardinal direction, and the new camper goes in the remaining direction. In this way the arrangement can persist, even if none of the original campers remain, but each time a new camper arrives, it is sensible for them to cover the ground now vacated by their predecessor. They can do so simply by seeing that this now-vacated area is not covered by the remaining campers.

We have here a diversity of intentional profiles. The original four campers have the profile 'cover the ground as specified by our agreement'. The replacing campers can have a variety of different profiles, especially since there are many descriptions they could have in mind that describe the same behavior, in line with Anscombe (1963) and her characterization of actions being intentional under one description but not under another. One description may be 'cover the ground not covered by the other campers', or any of the other profiles that identify the same ground, such as 'head out in a different direction to the other campers', or 'head into the area that has been picked for firewood the least', and so on. Lewis's point (in our language) is that this diversity of intentional profiles amounts to conforming to the same social regularity.

We can adopt this same point for our purposes of illustrating how the members of a group can act in relevantly similar ways for a variety of reasons. Indeed, social phenomena involve diverse actors in diverse situations acting for diverse reasons, and our explanations need to respect that (Little, 1991).

4. A coalition of the hesitant

Some examples of apparent irrationality regarding COVID-19 arise due to the difficulty of mind-reading groups with multiple intentional profiles. Again, we focus on cases of people who hesitate to adhere to COVID-19 mitigation measures despite being conditionally willing to adhere to them. The point of these examples is that there are many ways to hesitate to adhere to these measures that can easily look like unconditional refusal. When we see hesitancy, the argument goes, and we know that at least some people are outright refusers, then we are tempted to mind-read those who are merely hesitant as also being outright refusers. This means that even if the unconditionally hesitant are only a small part of the population, they may be enough to prompt defection by the conditionally hesitant.

The problem that the outright refusers pose is that there is no real prospect of getting them to join the effort. The fact that their refusal is (near-)unconditional makes it more likely that others think they are irrational, but the irrationality is not what is operative: instead, it is the hesitation to cooperate. That is just as well, because irrationality is hard to pin down but cooperation or hesitancy can be observed. The problem is that our ability to perceive cooperation or hesitancy is coarse-grained, in that we can only see the bare behavior of joining the effort or not, and must use mind-reading to infer someone's reasons for doing so. Much outright refusal is loudly and visibly unconditional. This means that we see hesitancy, the most salient diagnosis is in terms of outright refusal. The tendency to see hesitancy as aligned with outright refusal is of course defeasible, but defeasibility does not do away with the threat offered to cooperation.

Hence the presence of even low numbers of outright refusers can make effortful cooperation harder simply by making it more likely that if we see someone who is hesitant we expect them to be (near-)unconditionally hesitant. That small push is likely to turn at least some people from conditionally adhering to a COVID-19 mitigation measure to conditionally hesitating to do so. And the more people turn from adherence to hesitation, the more likely it is that the next person will also hesitate. This process threatens to cause a defection cascade.

The outright refusers and those who are in league with them form a *coalition* in the sense used by Collins (2019) in her distinction between collectives, coalitions, and combinations. Collectives are groups that have a shared decision-making procedure and where the members implement the choices that result from that procedure. Combinations are groups of people who just happen to be thrown together by circumstance. Coalitions fit between collectives and combinations in their degree of internal structure: they are groups that share ends but lack a shared decision-making procedure.

The problem comes in when the coalition in opposition to COVID-19 mitigation measures can recruit people who are in fact conditionally willing to adhere to these measures. The outright refusers are not likely to change enough conditional conformers over to their side, but what can happen is that even conditional conformers can display hesitancy about implementing a mitigation measure. What matters for succeeding at effortful cooperation is whether someone is willing to join the effort. Hesitation to do so, whether unconditional or conditional, can serve as a signal that the individual in question may not make the effort. So, for the outright refuser's goal to frustrate effortful cooperation (because they do not think that it has a worthwhile end, or that it is necessary, or whatever), it may very well be enough to get enough people who would otherwise adhere to hesitate to do so, since this may start a defection cascade. In any case, such recruitment will make the effortful cooperation harder. If the conditional conformers can be made to hesitate, they join in something like a tragic coalition with people whose views they do not share, and end up doing things that by their own lights they should not do, because, as discussed above, a significant majority of people are actually conditional conformers.

The rest of this section offers a grab-bag of examples of intentional profiles that are likely to produce hesitancy but that fall short of outright refusal. The point of all of these examples is that all of them involve people whom we would expect to conditionally cooperate with COVID-19 mitigation measures, even if perhaps they require more convincing to do so. As with everything else we have looked at in this paper, there is going to be a spectrum of willingness to cooperate, which ranges from some people who may only cooperate when it appears that implementing the measure is a *fait accompli*, to people who would very much like to cooperate but despair about the measures being successful.

4.1. Resentful conformity

There there are many people who ultimately join in effortful cooperation, but resent it and do so only under protest. That there is a sizable and vocal group like this is something that we see in past and present experience with most any large-scale public health measures, such as banning smoking in indoor spaces (Poland, 2000), mandating seat belts for car travel (Giubilini and Savulescu, 2019), fluoridization of water (Wrapson, 2005), and so on. Simply put, whenever people are told to do something on a sufficiently large scale, there are many people who reflexively kick against the pricks. We can expect this group, *resentful conformers* we may call them, to conform to vaccine mandates and similar measures, but not be happy to be forced to do so, and have some sympathy for people who outright refuse to conform. Accordingly, if we see a group consisting of outright refusers and resentful conformers, the extent of apparent support in favour of outright refusal is exaggerated. For example, when vaccine mandates have actually been enforced, the number of outright refusers is surprisingly tiny (Palosky, 2021).

4.2. Status quo bias

Another group of conditional conformers who may come to hesitate about COVID-19 mitigation measures are people sensitive to *status quo bias* (or omission bias). This is the propensity of individuals to disproportionately stick with the status quo (or a choice highly characteristic of the status quo) in the face of alternatives (Samuelson and Zeckhauser, 1988). There is already work that uses status quo bias as one way of explaining group behavior in response to the COVID-19 pandemic (Chappell, 2022). This is not new; status quo bias has long been used to characterize the issue of vaccine hesitancy (Asch, Baron, Hershey, Kunreuther, Meszaros, Ritov, and Spranca, 1994; Ritov and Baron, 1990).

Status quo bias is likely to be at work in current responses to the pandemic. One way is via the assumption that COVID-19 has the same risk profile as existing diseases (such as the common cold or the flu), and therefore it would be less dangerous to contract it compared to taking a new vaccine. Another is that for the cold and flu we had not resorted to costly measures to the extent that we are being asked to for COVID-19.

Another relevant example of status quo bias is people who are in general willing to vaccinate who hesitate to take mRNA and adenovirus-vector vaccination regimes given their novelty, instead holding out for other variants such as NovaVax that use traditional technologies. In other work we have found ample empirical evidence of statements on social media of individuals avowing one of these attitudes (Quintana, Cheong, Alfano, Reimann, and Klein, 2022). These attitudes amount to a (conditional) refusal to join in effortful cooperation around COVID-19 mitigation measures because of the extent to which they differ from perceptions of how previous public health efforts around infectious disease were made.

4.3. Affiliation effects

We can predict what an individual believes to a remarkable extent simply by looking at what their peers believe. A great deal of recent social epistemology has highlighted the extent to which we depend on the (perceived) views of our peers when we decide upon our own views, building upon a long-standing appreciation of the import of testimony to individual reasoning. This is frequently explained in terms of what Bicchieri (2016) calls an individual's *reference network*, being the people to whom they feel an affiliation and whom they look to when estimating the established views on a topic.

In the COVID-19 case, we find the import of affiliation effects illustrated in a number of ways. One is where political affiliation is a good predictor of willingness to vaccinate, to the extent that, given the greater risk of the disease to the unvaccinated, in the US supporters of the Republican party (who disproportionally hesitate to vaccinate) are dying from the illness in significantly higher numbers than non-Republicans (Gao and Radford, 2021). Another, smaller scale, example is where individuals in close-knit social networks such as a family group all hesitate to vaccinate, until one of the members admits that they have been vaccinated, and then other members of the group quickly follow. This phenomenon is well-studied in the case of hesitancy to take the flu vaccine (Bruine de Bruin, Galesic, Parker, and Vardavas, 2020).

Looking a bit closer at this second example, what has happened is that the members of the group find themselves in a situation where their reference network are all unvaccinated, and take that as a cue to themselves hesitate to vaccinate. When one of the members of their reference network announces that they have vaccinated, it makes vaccination a live option. Once vaccination is not foreclosed by the opinions of their peer network, all the positive reasons for vaccination can have an effect, and the reference network switches from being vaccine-hesitant to vaccinated. This is one example of people who would otherwise be willing to vaccinate in the absence of affiliation effects. These people are likely to be part of the coalition of the hesitant.

4.4. Prestige bias

Prestige bias is yet another factor driving hesitancy to adhere to COVID-19 mitigation measures. The thought behind prestige bias was developed by Henrich and Gil-White (2001) who "suggested that people use indirect cues of success (e.g., differential levels of attention paid to models by other social learners) as adaptive short-cuts to select models from whom to learn" (Jiménez and Mesoudi, 2019). In social philosophy, prestige bias is taken as another example of how reference networks play a role in determining the views of individuals. What makes prestige bias different from affiliation effects is that the high visibility of some individuals can lead them to have an outsized effect on the views of large groups people, rather than in the case of affiliation effects where the relationship between peers is much more symmetrical. Bicchieri (2016) gives many examples where highly visible individuals with high prestige, such as characters on popular television shows, can prompt social change as many people use them as role models.

When it comes to COVID-19, vaccine hesitators and deniers are liable to prestige bias by prioritising the views of celebrities and other high-status individuals over those of relevant authorities such as scientists and public health organisations. There is a widespread phenomenon where COVID-19 misinformation is caused by celebrities expounding views that are then readily echoed by their followers (Grimes, 2021). The spread of such misinformation lowers the likelihood that people subject to it will join effortful cooperation. The effects of prestige bias are notable on COVID-19 discourse on social media, where high-profile accounts account for the overwhelming majority of visible sentiment around COVID-19 and are good predictors for the views we see echoed among the population as a whole (Quintana et al., 2022).

5. How the two problems exacerbate each other

We have surveyed two ways in which people's ability to act rationally by their own lights can be hampered by not being able to judge whether their fellows will act in the way it requires. One way is that they can be unsure whether the the signal to effortfully cooperate will lead to people complying or to counterproductive resistance, and another way is that they can be systematically misled about how resistance to effortful cooperation is. What matters here is not just whether people hesitate, but whether they hesitate unconditionally or near-unconditionally. If someone hesitates but has a conditional willingness to effortfully cooperate, then that person is likely to effortfully cooperate once it becomes clear enough of their neighbors are also conditionally willing to do so.

These two problems make each other worse by forming a self-reinforcing feedback loop. The more we are misled to think that opposition to the required measures is more common than it is, the more we ourselves form part of the group who are uncertain about the measures. The existence of such a group is itself a signal that compliance with COVID-19 mitigation measures is uncertain. But, recall that on the Quintana et al. (2021) account many messages about COVID-19 mitigation measures give two competing signals: a first-order signal of what people should do to mitigate the effects of the pandemic, and a second-order signal that some people will fail to comply with the measures.

The usual mechanism is that the first-order signal highlights some countervailing feature of the measures and you can predict that that countervailing feature will prompt some people not to comply, such as the (responsible) reporting of AstraZeneca having a rare side-effect of blood clots highlights that risk, and there is a worry that some people will overreact to that risk and hesitate to take AstraZeneca as a result (Forman, Jit, and Mossialos, 2021). The difficulties with mind-reading groups worsens the risk that such second-order signals lead to defection cascades. Because a wider range of actual intentional profiles can prompt us to mind-read people as being unwilling to implement COVID-19 mitigation measures, we are likelier to think they will not cooperate with a mitigation measure.

That leads to the following exacerbated version of the defection cascade which powers the overdiagnosis of irrationality argument:

A social situation may offer two equilibrium outcomes: a low-effort equilibrium that individuals can reach on their own, and a high-effort equilibrium that requires effortful cooperation but is a better outcome for everyone. We often use signals to coordinate our effortful cooperation. If there is such a signal with the first-order content that guides individuals to the high-effort equilibrium, but with the second-order content that some people are likely to hesitate in doing so, there is a risk that many people will not have sufficient confidence that effortful cooperation will succeed, and so would choose not to take part. The more people hesitate to adhere to the first-order content of a signal, the more salient its second-order content becomes. The more salient the second-order signal becomes, the more likely it is that the group will in fact opt for the low-effort rather than the high-effort equilibrium, because of their lack of confidence that the effortful cooperation will succeed. Because mind-reading works on observed behavior, and the observed behaviors of people who hesitate to effortfully cooperate are similar, we are likely to overdiagnose the extent to which people will outright refuse effortful cooperation. That overdiagnosis increases the salience of the second-order content of the signal, which exacerbates the problem of the lack of confidence in the success of effortful cooperation.

So, the final outcome of this overdiagnosis of outright refusal is that the counterproductive signal defection cascade is made worse by it being harder to avoid. That counterproductive signal defection cascade depends on our predicting that some people in the population would not cooperate because of the second-order content of a signal of some COVID-19 mitigation measure. But the above discussion of mind-reading in groups has shown that we are likely to overestimate the amount of people who outright refuse to effortfully cooperate, as opposed to refusing despite a conditional willingness to adhere because they worry about whether enough other people will cooperate. This is because of the much larger size of the coalition who share the hesitancy of the outright refusers but not their motivations. And as the perceived outright refusal increases, the likelihood of the defection cascade does as well. Correspondingly, our ability to implement COVID-19 mitigation measures decreases.

That whole process can look like collective irrationality. Yet tragically so: for in fact the two sorts of defection cascades are triggered among those who are trying to reason carefully about what to do and are at least conditionally cooperative. Their attempts to rationally respond to this large-scale social problem has been frustrated by factors outside of their control, and they tragically end up in coalition with people who try to frustrate what they take to be the rational response.

6. Implications and recommendations

In this paper we have provided an analysis of how wider social factors can threaten even conscientious individuals' attempts to join effortful cooperation for COVID-19 mitigation measures. In conclusion, we now offer a very brief look at some of the practical implications.

Firstly, we have focused exclusively on how the social context of reasoning about mitigation measures can undermine conditional conformity. Readers may wonder whether we are thereby furnishing individuals with excuses to defect from a collective effort. But to insist on individual agency is to neglect the differences between acting in strategic as opposed to non-strategic situations. What makes a situation strategic is that the outcome of individual actions is partly determined by what all the agents do. This means that it is commonplace for individuals to find themselves somewhat trapped by the circumstances in what they consider to be sub-optimal outcomes, but where shifting to a preferable outcome requires more than just acting like you would in the preferable outcome (see Elster 1979). It simply is not in any individual's power to unilaterally change the behavior of those around them; they need mechanisms that spread change throughout the group. Defection cascades involve mechanisms of this kind, and this paper has given an explanation of how hesitancy may spread across a population. To insist that individuals ignore the social preconditions of their preferred outcome and push through their uncertainty about how their neighbors may act is to ask them to *unconditionally* pay the costs of effortful coordination, and that is a very different proposition.

Requiring unconditional cooperation in order to avoid defection cascades is difficult to defend. Firstly, when agents unconditionally cooperate they bear all the costs of doing so, but the benefits are still in doubt given that other agents do not automatically follow suit. Considering the matter in more detail, as a game-theoretic matter it has long been noted that unconditional preferences are only a way to avoid coordination problems, not to solve ones that have already arisen (Lewis, 1969). The thought is that by having an unconditional preference to cooperate, you warp you own response to a problem to such an extent that failure to conform to some collective solution is no longer an option, and that should make it easier for others to also join in, and sometimes making it more costly for them not to join in, because the unconditional cooperators are locked into their behavior. But for COVID-19 mitigation measures, the fact that I am vaccinating, wearing a mask, or engaging in social distancing does not in fact make it more costly for you to not do these things. So, this is one of those cases where unconditional preferences fail to change the costs for other people to join effortful cooperation. The idea then is that enough people unconditionally cooperating will removing doubt about the uptake of the mitigation measures. This idea is intelligible, but unhelpful: if I am in a position where it is doubtful that enough people are conditionally willing to cooperate, how much worse is it if we require the much more demanding standard of *unconditional* willingness? If we had enough people who were cooperating, we would not have had the problem to start off with; again, that is Lewis's point that unconditional preferences avoid coordination problems, not solve them. Neither excuse nor blame seems warranted in the face of the fact that effortful cooperation, a feature of a population taken together, is simply in a different category from individual actions like setting your own threshold for cooperating.

Instead, we suggest that the proper level at which to address defection cascades is the environment within which individuals reason (Levy, 2021a). Concerns about a coalition of the hesitant reflect a kind of pluralistic ignorance, where individuals mistakenly believe their neighbors have different views on COVID-19 than they do (Leviston, Stanley, and Walker, 2022). The implication, in line with existing work on social norms (especially Bicchieri 2016) is that the most important factor for encouraging effective social action is what individuals can see in their reference networks. We suggest that authorities can and should use their visibility in this domain, e.g. to (truthfully) frame outright refusal as rare and aberrant behavior against a backdrop of widespread cooperation. The easier it is to identify conditional conformers and the reasons they do so, the more secure the population will be against defection cascades, exactly because such familiarity guards against the uncertainty of the views of others that drives the defection cascades we have outlined.

A useful approach is to look at cases where government action has been more successful at sustaining population-level effortful cooperation. Attitudes towards the government and authorities vary substantially across populations, and these are reliably linked with attitudes towards COVID-19 mitigation measures (Pavlović et al., 2022).

The relevant attitudes are various, but there is a pronounced effect of the sense of national identification (Van Bavel et al., 2022) and especially that the amount of trust the population have in authorities, predominantly the government (Liu, Shahab, and Hoque, 2022). Both tend to promote effective group action. Given our discussion, these results should not be surprising: the features that predict successful cooperation are ones that would forestall trying to guess what the attitudes of the people around you are, precisely because they would encourage the belief that a tendency to cooperate is something of a given (see also Levy 2021a). These are relevant, because defection cascades arise in the face of uncertain higher-order evidence about what the people around you believe (see also Levy 2021b). If we do not need to rely on questionable higher-order evidence, then we can avoid the fraught game of guessing at the attitudes of the people around us, and with that avoid the defection cascades and tragic coalition described in this paper.

References

L. Allais and F. Venter. Lockdown or no lockdown: we face hard choices for complex times. *Mail & Guardian*, Apr. 2020. URL `https://mg.co.za/article/2020-04-13-lockdown-or-no-lockdown-we-face-hard-choices-for-complex-times/`.

G. Anscombe. *Intention*. Blackwell, Oxford, 2nd edition, 1963.

D. A. Asch, J. Baron, J. C. Hershey, H. Kunreuther, J. Meszaros, I. Ritov, and M. Spranca. Omission bias and pertussis vaccination. *Medical Decision Making*, 14(2):118–123, 1994. .

C. Bicchieri. *Norms in the Wild: How to Diagnose, Measure, and Change Social Norms*. Oxford University Press, 2016.

A. Broadbent. Lockdown is wrong for Africa. *Mail & Guardian*, Apr. 2020. URL `https://mg.co.za/article/2020-04-08-is-lockdown-wrong-for-africa/`.

W. Bruine de Bruin, M. Galesic, A. M. Parker, and R. Vardavas. The role of social circle perceptions in ”false consensus” about population statistics: Evidence from a national flu survey. *Medical decision making*, 40(2):235—241, 2020. .

R. Y. Chappell. Pandemic Ethics and Status Quo Risk. *Public Health Ethics*, 01 2022. . URL `https://doi.org/10.1093/phe/phab031`.

S. Collins. *Group Duties: Their Existence and Their Implications for Individuals*. Oxford University Press, Oxford, 2019.

J. Elster. *Ulysses and the Sirens: Studies in Rationality and Irrationality*. Cambridge University Press, 1979.

M. Ferreira. Inscrutable processes: Algorithms, agency, and divisions of deliberative labour. *Journal of Applied Philosophy*, 38(4):646–661, 2021. .

R. Forman, M. Jit, and E. Mossialos. Divergent vaccination policies could fuel mistrust and hesitancy. *The Lancet*, 397(10292):2333, 2021. .

J. Gao and B. J. Radford. Death by political party: The relationship between COVID-19 deaths and political party affiliation in the United States. *World medical & health policy*, 13(2):224–249, June 2021. .

A. Giubilini and J. Savulescu. Vaccination, Risks, and Freedom: The Seat Belt Analogy. *Public Health Ethics*, 12(3):237–249, 2019. .

D. R. Grimes. Medical disinformation and the unviable nature of covid-19 conspiracy theories. *PLOS ONE*, 16(3):1–17, 03 2021. .

J. Henrich and F. J. Gil-White. The evolution of prestige: freely conferred deference as a mechanism for enhancing the benefits of cultural transmission. *Evolution and Human Behavior*, 22(3):165–196, 2001. .

A. V. Jiménez and A. Mesoudi. Prestige-biased social learning: current evidence and outstanding questions. *Palgrave Communications*, 5(1), 2019. .

N. Kassam and H. Léser. Climate Poll 2021. Technical report, Lowy Institute, Sydney, Australia, May 2021. URL `https://www.lowyinstitute.org/publications/climatepoll-2021`.

Z. Leviston, S. K. Stanley, and I. Walker. Australians underestimate social compliance with coronavirus restrictions: findings from a national survey. *Australian and New Zealand Journal of Public Health*, 46(3):304–306, 2022. .

N. Levy. *Bad Beliefs: Why They Happen to Good People.* Oxford University Press, 2021a. ISBN 978-0-19-289532-5.

N. Levy. Echoes of covid misinformation. *Philosophical Psychology*, 2021b. . Advanced online publication. https://10.1080/09515089.2021.2009452.

D. Lewis. *Convention: A Philosophical Study.* Harvard University Press, Malden, MA, 1969.

D. Little. *Varieties of Social Explanation: An Introduction to the Philosophy of Social Science.* Westview Press, 1991.

J. Liu, Y. Shahab, and H. Hoque. Government Response Measures and Public Trust during the COVID-19 Pandemic: Evidence from Around the World. *British Journal of Management*, 33(2):571–602, 2022. .

I. Moscati. On the recent philosophy of decision theory. *Journal of Economic Methodology*, 28 (1):98–106, 2021. .

K. Murphy. Guardian Essential poll: majority of Australians support vaccine mandates. *The Guardian*, Sept. 2021. URL `https://www.theguardian.com/australia-news/2021/sep/14/guardian-essential-poll-majority-of-australians-support-vaccine-mandates`.

C. Palosky. 1 in 4 workers say their employer required them to get a covid-19 vaccine, up since june; 5% of unvaccinated adults say they left a job due to a vaccine requirement. Kaiser Family Foundation, 2021. URL `https://www.kff.org/coronavirus-covid-19/press-release/1-in-4-workers-say-their-employer-required-them-to-get-a-covid-19-vaccine-up-since-june-5-of-unvaccinated-adults-say-they-left-a-job-due-to-a-vaccine-requirement/`.

T. Pavlović, F. Azevedo, K. De, J. C. Riaño-Moreno, M. Maglić, T. Gkinopoulos, P. A. Donnelly-Kehoe, C. Payán-Gómez, G. Huang, J. Kantorowicz, M. D. Birtel, P. Schönegger, V. Capraro, H. Santamaría-García, M. Yucel, A. Ibanez, S. Rathje, E. Wetter, D. Stanojević, ..., and J. J. Van Bavel. Predicting attitudinal and behavioral responses to COVID-19 pandemic using machine learning. *PNAS Nexus*, page pgac093, July 2022. .

B. D. Poland. The 'considerate' smoker in public space: the micro-politics and political economy of 'doing the right thing'. *Health & Place*, 6(1):1–14, 2000. .

I. O. Quintana, S. Rosenstock, and C. Klein. The coordination dilemma for epidemiological modelers. *Biology and Philosophy*, 36(54), 2021.

I. O. Quintana, M. Cheong, M. Alfano, R. Reimann, and C. Klein. Automated clustering of covid-19 anti-vaccine discourse on twitter, 2022.

I. Ritov and J. Baron. Reluctance to vaccinate: Omission bias and ambiguity. *Journal of Behavioral Decision Making*, 3(4):263–277, 1990. .

G. Ryle. *The Concept of Mind.* University of Chicago Press, 1949.

W. Samuelson and R. Zeckhauser. Status quo bias in decision making. *Journal of Risk and Uncertainty*, 1(1):7–59, 1988. .

T. C. Schelling. *The strategy of conflict.* Harvard University Press, 1960.

D. T. Smith, K. Attwell, and U. Evers. Support for a covid-19 vaccine mandate in the face of safety concerns and political affiliations: An australian study. *Politics*, 2021. . Advanced online publication. https://doi.org/10.1177/02633957211009066.

S. Spaulding. *How we understand others: Philosophy and social cognition.* Routledge, 2018a.

S. Spaulding. Do you see what I see? How social differences influence mindreading. *Synthese*, 195(9):4009–4030, 2018b. .

K. Tullmann. The Problem of Other Minds. *Metaphilosophy*, 50(5):708–728, 2019. .

J. J. Van Bavel, A. Cichocka, V. Capraro, H. Sjåstad, J. B. Nezlek, T. Pavlović, M. Alfano, M. J. Gelfand, F. Azevedo, M. D. Birtel, A. Cislak, P. L. Lockwood, R. M. Ross, K. Abts, E. Agadullina, J. J. B. Aruta, S. N. Besharati, A. Bor, B. L. Choma, ..., and P. S. Boggio. National identity predicts public health support during a global pandemic. *Nature Communications*, 13(1):517, 2022. .

E. Westra. When is mindreading accurate? A commentary on Shannon Spaulding's How We Understand Others: Philosophy and Social Cognition. *Philosophical Psychology*, 33(6): 868–882, 2020. .

J. Wrapson. Artificial fluoridation of public water supplies in new zealand: 'magic bullet, 'rat poison, or communist plot? *Health and History*, 7(2):17–29, 2005.