European Journal for Philosophy of Science (2022) 12:63
https://doi.org/10.1007/s13194-022-00493-7

PAPER IN THE PHILOSOPHY OF THE SOCIAL SCIENCES AND HUMANITIES

What should scientists do about (harmful) interactive effects?

Marion Godman[1] · **Caterina Marchionni**[2]

Received: 9 June 2022 / Accepted: 26 September 2022 / Published online: 10 November 2022

Abstract
The phenomenon of interactive human kinds, namely kinds of people that undergo change in reaction to being studied or theorised about, matters not only for the reliability of scientific claims, but also for its wider, sometimes harmful effects at the group or societal level, such as contributing to negative stigmas or reinforcing existing inequalities. This paper focuses on the latter aspect of interactivity and argues that scientists studying interactive human kinds are responsible for foreseeing harmful effects of their research and for devising ways of mitigating them.

Keywords Interactivity · Human kinds · Predictability · Moral responsibility of scientists · Unintended consequences

1 Introduction

Scientific claims about people and their behaviour can sometimes interact and change them in ways that are significant, harmful, and difficult to predict. A well-known example is scientific claims about gender differences in cognitive abilities. Such claims may affect women's self-conception or limit their agency, thereby reinforcing such purported differences (Kourany, 2016). Another set of established examples comes from the domain of psychiatric diagnosis. It is well-documented that at least some diagnoses of psychiatric illnesses are associated with negative stigmas. As such

This article belongs to the Topical Collection: Reactivity in the Human Sciences
Guest Editors: Marion Godman, Caterina Marchionni, Julie Zahle

✉ Caterina Marchionni
caterina.marchionni@helsinki.fi

[1] Department of Political Science, Aarhus University, Aarhus, Denmark

[2] Practical Philosophy, University of Helsinki, Unioninkatu 40 A, 00014 Helsinki, Finland

they can make it (even more) difficult to 'live with' the disorder, because the negative aspects of the diagnosis have been reinforced by scientific claims. We group these types of effects under the single label of *interactive effects*: the phenomenon whereby science affects its object of study in ways that feedback on the epistemic status of the science itself. The result of human kinds (and categories) undergoing such interactive effects is that they become *interactive kinds* (Khalidi, 2010).

Such interactivity has been studied in different fields under different labels such as *self-fulfilling prophecies*, *looping effects, reactivity, reflexivity*, and *performativity*. A central concern has been about the ontological status of interactive kinds and what their epistemic consequences are, especially for the reliability of the scientific claims. Interactivity, however, has consequences that go beyond the narrow confines of the scientific enterprise. It can threaten the perceived epistemic authority of science, but more importantly it can express stigma, disrespect, and cause significant harm to some groups. In such cases, it seems legitimate to ask whether such results should be disseminated or sought for to begin with. Preventing harm by suspending research or dissemination is not without costs, however. It risks compromising cherished values such as academic freedom and scientific autonomy. It might also obstruct epistemic values including the idea of scientific progress, and can prevent the production of knowledge that may be put to positive societal use. Think of how some psychiatric diagnoses may feel liberating to long-time sufferers, or of the practical handle of injustices in education we may get by studying the effects of gender on different academic outcomes.

Compounding this dilemma, interactive effects seem to be rather uncertain and difficult to anticipate. Many commentators of interactive kinds have emphasised this aspect (Hacking, 1986; Khalidi, 2010; Laimann, 2020). Nor do people react to scientific claims in a uniform manner; they may react in a haphazard way that does not add up to any overall significant effect. Thus, from the point of view of responsible scientific conduct, interactive effects are morally troubling as scientific claims can lead to bad moral and societal outcomes; but interactive effects are also epistemically precarious as interactivity seems highly variable – to the extent that changes to interactive human kinds are often deemed unpredictable.

In this paper we argue that the unpredictability of interactivity has been overstated. Interactivity and its effects do vary significantly across contexts, but this variability is not tantamount to unpredictability. On the contrary, interactive mechanisms and their effects can be systematically studied and knowledge about them cautiously extrapolated. This knowledge not only allows scientists to sometimes anticipate interactivity but also to devise pathways to block it or alleviate its harmful effects. The possibility of interactivity is thus a significant issue for the responsible conduct of science. But, as we will argue, it is also one that science can typically manage.

Before moving further, some clarification is due about what we are not arguing in this paper. First, a lot of philosophical interest has been centred around the question of whether interactivity that involves people or occurs in the case of human kinds is distinctive from other forms of interactivity between science and its object of study (cf. Cooper, 2014; Khalidi, 2010). Here we do not take a position on this but will focus on cases where people are the ones undergoing interactivity. Second, and relatedly, our goal is *not* to show that unintended consequences of research that involves interactive

human kinds entails a *distinctive kind* of scientific responsibility. Rather we are concerned with showing that considerations of possible harmful effects of interactivity should be included in the responsible conduct of science more generally. Third, we do not believe that the effects of interactivity are always epistemically problematic or morally and socially harmful. There are certainly cases where reactive effects are beneficial or even emancipatory for individuals. These kinds of cases are discussed for example in Koskinen (2022) and van Basshuysen et al. (2021).

2 The (moral) trouble with scientifically induced interactivity

Interactions between science and its object of study can occur both during the research process when we produce data for a scientific result and because of the dissemination of such results. In a way the problem might seem to be more acute when it takes place during the research process itself: unintended reactions from, say participants in an ethnographic study, can contaminate the study results making it hard to distinguish what is merely an artefact of participants' reactions to being studied and what is in fact the behaviour which the researcher is interested in.

In the case of the dissemination of scientific claims about new taxonomic results, generalisations, and explanations of those generalisations, the problem is slightly different. Many social scientists have been troubled by how the publicity and dissemination of their findings might somehow "contaminate" their results (e.g., Merton, 1948). Some of the most prominent philosophers of science have also been aware of this epistemic challenge, though most of the focus of the most recent philosophical debates has been about whether interactive effects imply a different ontological status for human kinds. Thus, the problem they describe chiefly falls under the scope of what Hacking (2007) has famously called "the looping effects" of human kinds, where people change in response to a scientific classification and then science, in return, has to adjust its claims in response to the new facts about the human kind that was their initial target of scrutiny. In this paper we will be less concerned with the second part of the loop (how science should respond to the new facts of the human kinds) than with how the human kind changes in response to scientific claims. Following Muhammad Ali Khalidi, we refer to this as the process of human kinds becoming *interactive* (2010).

Consider one of Hacking's first descriptions of the phenomenon; namely interactivity in the case of multiple personality disorder (1995). Hacking describes how the condition's generalized association with child abuse was brought about by the interplay between a powerful suggestion from psychiatry about the existence of such an association and its subsequent reception by people classified with multiple personality disorder who began to recover traumatic memories of child abuse. The association was thereby stabilised. Notice that even in such examples of stabilisation, when a claim is stable over time through interactivity, the claim is not stabilised by the same *mechanism* as was assumed or predicted by the scientists. Jessica Laimann gives the example of women being raised and socialised in a culture where they are told how they

differ from men innately or biologically. As a result, the scientific community testifies to these gender differences, thereby reinforcing them via social feedback mechanisms (Laimann, 2020 p. 1055 f.). The scientific claim is stable over time, but now via a social feedback mechanism.

There are also important cases of destabilising feedback. Recently some social scientists discovered that the likelihood of becoming a target of work-related harassment increased if one does not oneself intervene in the harassment of others (and conversely) (Nielsen et al., 2021). Now suppose that upon learning this fact, the level of intervention against harassment increases in workplace organisations, but the level of harassment actually increases as a result of a form of conscious deterrence by bullies. This would be a case of destabilising feedback ultimately produced by the dissemination of a scientific claim that ends up changing the world in unintended ways.

In what sense should we find stabilising and destabilising interactive effects on human kinds troubling? We have already hinted that one set of worries is epistemic. We might be concerned with whether changes to people's experiences can be tracked in such a way that science can keep up with its object of study. When claims are stabilised (as in Hacking's example of the association between multiple personality disorder and child abuse above) or destabilised (such as when individuals resist claims made about them and as a result act contrary to the scientific expectations of them) the result is that the accuracy of scientific claims might not be particularly long-lived. In addition, we typically don't know *whether* there will be interactive effects or *what* they will be. We will turn to this epistemic issue below. For now, we want to concentrate on another set of effects that have received far less attention in the philosophical debate on interactive kinds: the moral and socially detrimental effects of interactivity and why they should be of concern to scientists. As we will argue below, these effects are central to choices about communication and dissemination of scientific claims.

By harmful interactivity, we mean harms or detrimental effects that interactivity has on the people or groups to which the scientific claims pertain – either directly because the claims are about them, or indirectly, because they affect other relevant actors and institutions (as in the example of Asperger's syndrome below).[1] For example, scientific claims may express things deemed stigmatising or disrespectful, even when the targeted group affirms the claim about themselves (Hellman, 2008). This seems especially concerning when it comes to historically vulnerable groups. A related worry with scientific claims that end up confirming stereotypes is that they reduce the space of agency and freedom of those that the claim concerns (Fine, 2010; Moreau, 2020, Ch. 3). These concerns would be central to several different normative ethical frameworks. From a Kantian or deontological framework, we would worry about the disrespect that such claims display. From a virtue theorist we would worry about

[1] What is more, it is not clear that interactive effects are always homogeneous (stabilising or destabilising) within a population. The same claim could also have heterogeneous effects and point in different directions within the group.

the departure from both moral (and epistemic) virtues expected of scientists. And from a broadly consequentialist perspective, one would worry about such claims producing more overall harm or disadvantage compared to the benefits.[2]

As an example, consider the diagnosis of Asperger's syndrome (Kuorikoski & Pöyhönen, 2012; Eyal et al., 2010). It was introduced allegedly with the aim of distinguishing a subgroup of the population with autism that had partly a distinctive set of symptoms and were less hampered by their condition than those that fell under the then standard autism diagnosis. The new diagnosis was overwhelmingly welcomed by the community who felt it fit with their experience. At the same time, the diagnosis had the unanticipated effect of increasing the stigma and stereotyping of those who were diagnosed with Autism but did not match the Asperger's diagnosis in the new taxonomy (Cooper, 2014, Ch. 5). As a result of such experiences the diagnosis was dropped in the fifth edition of the influential Diagnostic Statistical Model (DSM) in favour of a general diagnosis of Autism. The latter was explicitly formulated to lie on the continuum, thereby stressing the continuity between both ends of the spectra (for a discussion, see Solomon, 2017). As the case illustrates, harmful interactive effects may also concern a different group (people with autism) from those initially targeted by scientific claims (those diagnosed with Asperger's syndrome). For many, Asperger's was welcomed as a diagnosis that fit with the lived experience of the patient; the problem was that at least arguably it unjustly affected other individuals due to enhancing unwanted stereotypes, feelings of inferiority and so on.

What should we do in the face of such harmful effects of scientific classification? At first pass if such effects are significant, the responsible thing to do would seem to be to prevent interactivity to occur in the first place. The obvious way of doing so is to avoid dissemination of potentially negative interactive results. On closer reflection, however, this seems to be a high price to pay since preventing interactivity in this way can obstruct the advancement and uptake of scientific knowledge, including knowledge that might be of great help precisely to the people and groups it targets. It also goes against the public role of science, which centrally involves a duty of communicating and disseminating sound research.

This situation then gives rise to a dilemma between, on the one hand, preventing interactivity to minimise the risk of causing unintended harm but also foregoing the benefits of research dissemination, and, on the other, continuing to disseminate and apply scientific knowledge (not least for all the benefits that flow from this) while accepting interactivity and its possible harmful effects.[3]

[2] Nor does one have to take a particular metaethical stance about whether these effects represent real and objective moral effects or rather represent functions or expressions of what we humans happen to care about. Either way, we seem to have reason to be concerned about how scientific claims affect the groups of people it pertains to.

[3] This way of formulating the dilemma is inspired by discussions of dual-use dilemmas, where the same piece of research has the potential to be used for both good and bad purposes (Miller & Selgelid, 2008). Here however we do not take a stand on what ethical principles should be adopted when deciding which horn of the dilemma ought to be chosen (in general or in a particular case).

3 Is the interactivity dilemma one that should concern *scientists?*

The study of interactive human kinds then poses a dilemma, which requires that different kinds of harms and goods, some of which are epistemic while others are moral, are considered. Is such a morally infused dilemma also a dilemma for scientists, though? We might think that while publishers, media, and policymakers should consider it in relation to scientific dissemination, scientists should be more sanguine about interactive effects. After all, it could be said that science should be autonomous of such considerations: as unfortunate as unintended consequences of the dissemination of a particular scientific claim or discovery might be, they need not burden scientists *qua* scientists. The primary responsibility of science is to describe things as they are; the responsibility for how those claims affect people is the responsibility of the journalist, the populariser, or the policy maker.

This limited view of scientists' responsibility has come under attack in recent decades as part of the critique of the value-free ideal in science. Douglas (2003, 2009), in particular, has made the influential argument that scientists are not merely responsible for the immediate consequences for say research participants, but also for the wider effects of their claims (see also Resnik, 1998, Carrier, 2021). Such responsibility includes duties such as considering the likely effects of one's claims in advance, deciding whether to communicate potentially harmful results, and also whether to pursue certain lines of research in the first place (see also Kitcher, 2003, Kourany, 2016).

Exactly how much this responsibility or duty grows out of the special role of scientists in society is debatable. Such duties can either be thought of as stemming from a general moral responsibility towards foreseeable consequences of one's actions (Douglas, 2003) or a special duty or professional responsibility (e.g., Pettit & Goodin, 1986). Either way, the upshot is that, as they are among the most powerful epistemic authorities in society, scientists have some *prima facie* responsibility to consider and respond to the unintended consequences of their scientific activities. One type of such unintended consequences are precisely the interactive effects discussed in the last section.

This kind of duty is already encapsulated in contemporary research ethical frameworks and guidelines, where scientists, but also science communicators, are asked to consider the unintended consequences of their research, both for research participants and for society at large. This makes it natural to characterise the duty as stemming from a forward-looking responsibility that focuses not so much on blaming scientists for particular harms resulting from their claims, but rather on what they should do to prevent or alleviate potential harms (Young, 2011; Smiley, 2017). This is a responsibility that, for reasons that will become clear below, does not only arise from the causal contribution of science, but also from the special epistemic position that scientists are in, which allows them to foresee such harms and their likelihood, as well as devising ways of alleviating them. Indeed, we can think that such responsibility lies with the collective of scientists rather than exclusively with the individual scientists (or team thereof) whose claims are at stake. Equally, the

responsibility for remedying or alleviating harmful effects and the duties might lie with the scientific community as a whole (Douglas, 2014).

The focus on forward-looking responsibility (rather than on its backward-looking counterpart) also allows us to bypass questions concerning how to apportion and assign blame among the different agents involved. For example, it could be argued that scientists should not be the only ones to blame considering that those who undergo reactivity sometimes also have the freedom (and with that responsibility) to decide how to act in response to the claims made by scientists. Relatedly, it might be argued that scientists are not fully responsible and hence blameworthy because they operate within institutional structures that can significantly constrain their options (for example, by disincentivizing publication of negative results). Accordingly, part of the responsibility arguably lies in the institutions themselves.[4] Irrespective of how one answers these questions, we can nevertheless hold a responsibility for scientists to take unintended consequences into account in the conduct of research on interactive kinds.

4 Solving the dilemma: middle-range theorising to the rescue

Before confronting the dilemma, scientists face a different issue which is the sheer variability of interactive effects. Will the new results contribute to reinforcing the negative stereotype associated with the disease or will the diagnosis instead feel potentially liberating to sufferers removing them of responsibility? Will young women be conscious of lurking sexism and resist the next bunch of claims about the neurological differences underpinning mathematical abilities, perhaps enrolling in the sciences in ever greater numbers; or will they and the people around them keep relying on the scientific claims and established stereotypes? Different groups may also respond differently to the same claims (consider for example the differential reactions to the Asperger's diagnosis).

This has motivated many to believe that the changes people undergo in reaction to scientific claims made about them are unpredictable. Laimann proposes to call interactive kinds *capricious*: the changes such interactive kinds undergo are far from systematic. The members of such kinds, she writes, "behave in wayward, unexpected manners that defeats existing theoretical understanding." (2020, p. 1043.) Hacking too seems to doubt the feasibility of a general theory of interactive kinds going beyond the description of a series of interesting cases. In *Making Up People*, Hacking writes: "I do not believe there is a general story to be told about making up people. Each category has its own history" (1986, p.168).

If interactive effects were indeed utterly unpredictable, the scope of scientists' responsibility would be significantly narrowed. Unforeseeable consequences are generally regarded to fall outside scientists' responsibility (Douglas, 2014) and cannot be ascribed moral value (see also Bergenholtz & Busch 2016). If so, scientists should be concerned with the epistemic effects of interactivity alone, which, as we

[4] Thanks to Uwe Peters for pressing us on these points.

have seen, might be challenging, but need not entail any significant changes to current scientific practices or to our conception of responsible conduct.

The unpredictability of interactive effects has been overstated, however. Some of the causal relations and mechanisms underlying interactive kinds may be relatively stable and therefore relatively predictable (Northcott, 2022). Predictability is both a function of how the world is – the stability of the interaction mechanisms across time and context – as well as of our current knowledge of those mechanisms. Take the kind "domestic dog", for example: Human interactions with the kind have changed it significantly, making dogs tamer and more obedient, but the breeding mechanisms are well-known and highly stable, and changes due to our breeding interventions can for the most part be reliably predicted (Khalidi, 2010; Northcott, 2022). In other cases, however, the mechanisms implicated have limited scope (their working and outcomes vary across time and contexts) and knowledge of them does not easily generalize. As a result, it might be hard to foresee change. For example, the interactive effects of the introduction of a new diagnosis might be difficult to predict because they will typically vary depending on the target population, culture and time of diagnosis. This does not amount to utter unpredictability, however. Even if we are unable to quantify the magnitude of these effects and identify which groups will be affected, we might know enough to correctly anticipate the direction of change such as whether some patterns of stigmatisation are likely to occur. Such broad qualitative predictions (cf. Elliot-Graves, 2016) can be sufficient to guide assessments of the risks of interactivity.[5]

Another way to think about this is to draw on recent analyses of social kinds as equilibrium solutions to coordination games. The "games" people play often have several possible equilibria that translate into different social roles occupied by members of a particular human kind (Guala, 2016; Mallon, 2018). Which equilibria a population (or social role) a kind ends up in depends on contingent features of the situation. It might seem especially difficult to predict at what equilibrium the group will arrive when facing interactivity. Take a popular example of a coordination game: we might be able to predict that once people coordinate on driving on one side of the road, they will continue to do so; but it is relatively harder to predict in advance which side of the road they will converge on. This example is a relatively simple convention involving only two possible equilibria with pure strategies. For the interactive kinds we are concerned with here, things are likely to be much more complicated. Several equilibria are in principle possible for any given coordination problem. On the flip side, there are often only a few equilibria that are salient in a particular culture at a particular time. In other cases, only one equilibrium will be possible, and we might be able to foresee which that will be.

[5] *Ceteris paribus*, the more is known about the mechanisms and the contexts in which they operate, the better we can identify possible unintended consequences of interaction effects – though some such consequences may remain unpredictable because of the fragility of the underlying causal relations.

Fortunately, the kind of knowledge (especially human) scientists produce makes them especially equipped to *cope* with predicting equilibria or social roles. Douglas (2009) and Alexandrova (2020) have argued that scientists *qua* scientists are in a position to think about the blend of epistemic and moral consequences of their research and its dissemination more generally. They argue that it is scientists, in virtue of their training and network, that can both access and assess the relevant empirical knowledge. This would include knowledge of likely interactive effects, and the possibility (and responsibility) to produce such knowledge when it is not already available. Thus, the task of considering interactive effects is not one that should be considered foreign to the scientific enterprise. On the contrary.[6]

We should therefore take some solace in that it is precisely scientists that come to the table with an epistemic perspective on interactivity. The mechanisms of interaction between knowledge of kinds and changes to the kind itself can be and are studied in the same vein as are other social and psychological mechanisms (Kuorikoski & Pöyhönen, 2012). The resulting theories have the shape and scope of Mertonian "middle-range theories", which lie in between on the one hand general theories of social phenomena, "which are too remote from particular classes of social behaviour, organisation, and change to account for what is observed" (Merton, 1968, p. 38) and on the other, ungeneralizable empirical descriptions of particular cases. As such, while middle-range theories fall short of adding up to a general theory of interactivity, they are generalizable given certain assumptions. Such theories do not deliver the precise quantitative predictions that we might expect in other domains, but the qualitative forecast they can sometimes provide are sufficient to anticipate likely effects.

Examples of such middle-range theories of interactive mechanisms are already available. For one, a wealth of sociological studies has sought to explain the "autism epidemic" by looking at social mechanisms of how information spreads.[7] It has been argued that this so-called epidemic was partly a result of feedback mechanisms from psychiatric classification to the social reality of autism and vice versa (Eyal, 2010). The reclassification of autism into a neurological disorder instead of a psychological one, and, later, from an emotional disorder to a developmental one, contributed to decreasing the negative stigma surrounding people with autism and their behaviour. In turn, this made more people attend to the diagnoses or made it more salient. At a finer grain, sociologists have suggested that the reason why diagnoses were clustered geographically was because information about the availability of the diagnoses, and help that came with it, spread through social networks that tend to be denser among neighbours (Liu et al., 2010).

Certain social and psychological mechanism schemas can be deployed for explanations of interactive mechanisms and to anticipate (likely) reactions towards scientific

[6] It could be argued that an aspect of inductive risk is involved here: the more harmful we deem the possible effects of interactivity to, the higher the burden on scientists to gather the relevant knowledge and deploy it to identify ways of mitigating harms. Thanks to an anonymous referee for pointing this possibility to us.

[7] Between 1993 and 2003 a 657% increase in the rate of autism has been recorded in the US by the Department of Education (Lilienfeld & Arkowitz, 2012).

claims. For example, there is convergent evidence from many different parts of psychology that citing genetic or brain-based explanations becomes associated with the idea that a kind or trait is fixed, irreversible or outside our control. When psychiatric symptoms are attributed to the brain they are considered less within the patient's control (e.g., Deacon & Baird, 2009). Why would we think they are outside our control? Here different psychological mechanisms are proposed, but, importantly, they point in the same direction as concern the outcome of the interaction. Some think the attribution is explained by our tendencies toward dualistic thinking (Miresco & Kirmayer, 2006); others take it to be our tendency to essentialize genetic, hormonal, and neural explanations (Heine et al., 2017). It has also been shown that beliefs about whether or not one can affect actions through one's efforts or free will are important for achieving certain results, whereas beliefs about them being out of our control have the opposite effect on our motivation (Mueller & Dweck, 1998; Baumeister et al., 2009). Either way, this case of middle-range theorising points toward the mere biologizing of traits interacting with psychological mechanisms that lead to more conformity with respect to the kind or diagnosis in question.

5 Bypassing the dilemma by design

Knowledge of the psychological and social mechanisms behind interactivity and their likely effects can not only assist us in addressing the dilemma by providing better grounds to decide which risks we are more willing to take, it can sometimes be drawn upon to bypass the dilemma altogether (cf. Miller & Selgelid, 2008). That is, it can be deployed to envisage strategies aimed at removing or mitigating likely negative moral consequences without suffering the epistemic costs of blocking the possibility of interactivity altogether. Sometimes, mitigation strategies are ways of disrupting an anticipated equilibrium and creating a new one that we (whoever we take the right ethical evaluator to be: the democratic polis, the moral expert, the policy maker, the scientist, or most likely a combination of them) deem more desirable compared to the existing one.

Take the case of research on gender differences in cognitive abilities and recall that the worry is that such research contributes to uphold stereotypes that we deem undesirable. One way to address this worry is to ban all research into such cognitive differences. This in effect is Kourany's (2016) suggestion, which is based on a reasoned assessment of the harms and benefits of this kind of research, including its impact on women's self-conception and agency. There might nevertheless be other ways of alleviating Kourany's worry. For example, it has been shown that people tend to judge results expressed in generic language as more important and more normative (DeJesus et al., 2019). It has also been argued that social kinds like race and gender are more likely to be thought of in essentialist terms when claims about them are made employing generics (Langton et al., 2012, Leslie, 2017). Peters (2021) draws on this work to hypothesise that this is because people tend to read off descriptive norms from generics more than from more carefully qualified claims. When descriptive norms in turn conform to existing gender stereotypes then they have a stronger conforming effect. The association between generic claims and descriptive norms is then precisely the

kind of thing a responsible scientist should be worried about contributing to. And yet scientists, often assisted by popular media, tend to formulate their generalisations, for example those about gender, as generics even though the evidence does not necessarily support the broad scope. Peters' (2021) suggestion is to reform communication strategies taking care to avoid needlessly eliciting descriptive norms, for example by avoiding the use of generics. If so, mitigation strategies that avoid needless generics would be especially important for the contexts and kinds that have a significant normative pull.

Let us now turn to the field of psychiatric classification and taxonomy. Here a great concern is that for some psychiatric disorders the negative stigma may be so significant that a diagnosis will do more harm than good. This is particularly the case for a diagnosis like anti-social personality disorder, associated with characteristics such as deceptiveness and aggressiveness. Beliefs that people falling under such classifications are dangerous (even if they should not be blamed for their condition) can induce avoidance of the psychiatric profession and even worsen the disease or its symptoms. Cooper (2012, 2021) argues that this may be because diagnostic labels tend to affect people's self-conception through the kind of narrative that they tell about themselves. She observes that communities of patients that develop positive narratives about the disorder can eventually help offset its negative stigma. (Think of the idea of a "good psychopath" namely, someone who has the traits of personality disorder, but who has managed to fit in in a social environment where some psychopathic traits are acceptable (McNab & Dutton, 2014)).

Cooper's idea is that we can harness this tendency to alleviate the negative effects of using the diagnosis in scientific communication. There tends to be a limited number of scripts or narrative types that are culturally salient and accessible, which makes it possible in some cases to encourage the adoption of positive narratives. Cooper's practical suggestion is therefore to phrase diagnostic criteria in such a way that they can more easily be incorporated into a positive self-narrative.

How do we identify fitting frames, narratives, and other mitigation strategies? In addition to the kind of middle-range theorising we just discussed, in the stage of communication one could also consult those who are most likely to be affected by the diagnosis or by some other scientific claim. That might not only give insights about which reactions are more likely, but also about what can be done to mitigate harmful effects of interactivity in the first place. More radically, mitigating the harms of interactivity might require involving the groups (or representative thereof) who are most likely to suffer negative reactive effects earlier on in the research process. Participatory research methods not only have the advantage of democratising science when sensitive and contestable value judgments are involved (see e.g., Alexandrova & Fabian, 2022), they can also help us better understand the likely experiences of those that will be affected by science and gather the kind of insights that can help us devise helpful mitigation strategies.[8]

[8] This could help distribute some of the responsibility onto those who will be mostly affected by the claims. Whether such a distribution is an advantage or not is debatable.

It should be noted that mitigating harmful interactivity is not the same as insulating research from reactivity altogether. There are at least three relevant complications with mitigation strategies.

First, we can imagine proposed strategies might induce more harm or disrespect than the ones they were intended to avoid. Take the different ways of mitigating interactivity such as introducing positive narratives of anti-social personality disorder. It could be that they end up producing their own harmful interactive effects, by, for example, rendering unacceptable aggressive behaviours more acceptable. Hence, in deciding whether to implement a given mitigation strategy, we should certainly consider whether the cure is not worse than the disease; in other words, that our countermeasures really will not make things worse overall. Example, people with personality disorders may feel emboldened to violence because the introduction of a new positive self-image associated with some scientifically transmitted narratives has somehow encouraged this. Therefore, the middle-range theories relied on should not be seen as single or static entities, but rather as theories that must continuously be assessed and developed such that they sharpen the ability to predict the reactive effects – not only of particular scientific claims, but also for alternative mitigation strategies.

The second complication has to do with the broader societal role of science as an epistemic authority. When science contributes to stabilising or destabilising results as a consequence of its dissemination, it can render its claims *self*-fulfilling or *self*-undermining. Apart from the moral or social consequences of particular claims, we might worry that insofar as science inadvertently contributes to changing the phenomena it is supposed to merely describe, explain or predict, its epistemic authority is progressively undermined (Lowe, 2021). For self-undermining interactivity, this should be obvious. But this can be the case also for self-fulfilling science, when the stability of its claims is due to different *mechanisms* than those assumed or predicted by the science.[9] As discussed above, this is the case for gender roles stabilised in part by the scientific claims themselves and not by the mechanisms originally predicted by science having to do with say neurobiological differences between men and women (see also van Basshuysen, 2022). The self-fulfilling or undermining nature of interactive kinds thus seems to undercut the idea that science is in the business of delivering a superior form of knowledge as compared to other sources of information. Indeed, the capacity of science to generate harmful interactivity is in part due to the special epistemic authority and trust already invested in it. This is perhaps a central concern for *institutions* that must manage the reputation of, and trust invested in, science. It is however hard to see how this type of concern can guide scientists' particular strategies to mitigate the harmful effects of their research.

Finally, mitigation strategies might not always be readily available. In some cases, there might not be a strategy that is practically or ethically feasible, or we might not be in the epistemic position to construct effective ones. In addition, on their own scientists might be able to affect the kind of changes needed to mitigate

[9] Yet another case is one in which the claim and the mechanism postulated is correct and yet it is self-stabilising. Here there is no epistemically worrisome consequence of interactivity, but we might still wonder whether science should or should not contribute to the stabilization of the phenomenon.

harmful effects. The example of positive narratives and psychiatric disorders makes clear that a lot of the institutional action needed to steer change in the right direction happens outside science narrowly conceived. But the fact that other actors and institutions are equally, and possibly often more, responsible and possibly even better placed to remedy such harms, does not mean that scientists are thereby completely off the hook.

6 Conclusion

We have argued that scientists have the responsibility to consider the possibility that their classifications, models and generalisations can change people and their behaviours in unintended and harmful ways. Moreover, we have argued that scientists are in a good position to take this responsibility by building and sharing middle-range theories of interactivity that help anticipate the effects of claims in particular contexts. Examples of such theories are those that delineate social and psychological features that predict how scientists might harm or disrespect those that fall under a diagnosis – or those that do not. Building on this capacity scientists might also devise mitigation strategies that allow the communication and dissemination of their results while preventing or alleviating harmful effects of interactivity.

Designing such mitigation strategies is therefore part of the responsible conduct of science when it deals with interactive human kinds. Clearly scientists with particular results might not have all the expertise at their disposal to be able to anticipate reactive effects or design a mitigation strategy. Hence the responsibility for interactivity does not so much entail that the same scientist should study both the substance of the claims about the kinds as well as their interactive effects. Rather, as it is currently the case, these might be a division of labour within the scientific community at large that not least can inform the research ethical frameworks and guidelines of research dissemination. As we have suggested throughout it makes sense to take the responsibility for and knowledge of interactivity as a collective scientific enterprise (cf. also Bergenholtz & Busch, 2016). Our claim is that at the very least scientists have some responsibility for interactivity; this does not mean that others have no responsibility or even more responsibility. Nor does this say much about how responsible conduct should be realized, although it would be natural to include it in research ethics guidelines for the human sciences and in research ethics education. Indeed, in some cases the institutional changes that need to be implemented to mitigate harmful effects of interactivity does require a wider range of institutional actors than the scientific community.

Acknowledgements The authors wish to thank participants at the Reactivity, Prediction and Intervention workshop in Helsinki (2021) and at the Philosophy of social science workshop in Milan (2022) as well as two anonymous reviewers for perceptive and encouraging comments.

Funding Open Access funding provided by University of Helsinki including Helsinki University Central Hospital. Marion Godman's research behind this article was funded by the Independent Research Fund in

Denmark (DFF 9062-00049). Caterina Marchionni gratefully acknowledges the support of the Academy of Finland.

Declarations

Conflict of Interest We hereby declare that there are no conflicting financial or non-financial interests and that the authorscontributed equally to this article.

Ethics statement Procedures for ethical approval and informed consent were not relevant to this philosophical research.

References

Alexandrova, A. (2020). Can the science of well-being be objective? *The British Journal for the Philosophy of Science, 69*(2), 421–445.

Alexandrova, A., & Fabian, M. (2022). Democratising measurement: Or why thick concepts call for coproduction. *European Journal for Philosophy of Science, 12*(1), 1–23.

Baumeister, R. F., Masicampo, E. J., & DeWall, C. N. (2009). Prosocial benefits of feeling free: Disbelief in free will increases aggression and reduces helpfulness. *Personality and Social Psychology Bulletin, 35*(2), 260–268.

Bergenholtz, C., & Busch, J. (2016). Self-fulfillment of social science theories: Cooling the fire. *Philosophy of the Social Sciences, 46*(1), 24–43.

Carrier, M. (2021). How to conceive of science for the benefit of society: Prospects of responsible research and innovation. *Synthese, 198*, S4749–S4768.

Cooper, R. (2012). Is psychiatric classification a good thing? In K. Kendler, & J. Parnas (Eds.), *Philosophical issues in psychiatry II: Nosology* (pp. 61–70). Oxford University Press.

Cooper, R. (2014). *Diagnosing the diagnostic and statistical manual of mental disorders.* Routledge.

Cooper. (2021). *Designing human kinds for better living*. Presented at the 2nd Reactivity workshop. Online Workshop.

Deacon, B. J., & Baird, G. L. (2009). The chemical imbalance explanation of depression: Reducing blame at what cost? *Journal of Social and Clinical Psychology, 28*(4), 415–435.

DeJesus, J. M., Callanan, M. A., Solis, G., & Gelman, S. A. (2019). Generic language in scientific communication. *Proceedings of the National Academy of Sciences*, *116*(37), 18370–18377.

Douglas, H. (2003). The moral responsibilities of scientists (tensions between autonomy and responsibility). *American Philosophical Quarterly, 40*(1), 59–68.

Douglas, H. (2009). *Science, policy and the value-free ideal*. University of Pittsburgh Press.

Douglas, H. (2014). The moral terrain of science. *Erkenntnis, 79*(5), 961–979.

Eyal, G., Hart, B., Onculer, E., Oren, N., & Rossi, N. (2010). *The autism matrix*. Polity.

Elliott-Graves, A. (2016). The problem of prediction in invasion biology. *Biology and Philosophy, 31*, 373–393.

Fine, C. (2010). *Delusions of Gender: How our minds, society, and neurosexism create difference*. WW Norton & Company.

Guala, F. (2016). *Understanding institutions: the science and philosophy of living together*. Princeton University Press.

Hacking, I. (1986). Making up people. In T. C. Heller, M. Sosna, & D. E. Wellbery (Eds.), *Reconstructing Individualism: Autonomy, Individuality, and the Self in Western Thought.* Stanford University Press.
Hacking, I. (1995). *Rewriting the soul: Multiple Personality and the sciences of memory*. Princeton University Press.
Hacking, I. (2007). Kinds of people: Moving targets. In *Proceedings-British Academy* (Vol. 151, p. 285). Oxford University Press Inc.
Heine, S. J., Dar-Nimrod, I., Cheung, B. Y., & Proulx, T. (2017). Essentially biased: Why people are fatalistic about genes. In *Advances in experimental social psychology* (Vol. 55, pp. 137–192). Academic Press.
Hellman, D. (2008). *When is discrimination wrong?* Harvard University Press.
Khalidi, M. A. (2010). Interactive kinds. *The British Journal for The Philosophy of Science, 61*(2), 335–360.
Kitcher, P. (2003). *Science, Truth, and Democracy*. Oxford University Press.
Koskinen, I. (2022). Reactivity as a tool in emancipatory activist research. *European Journal for Philosophy of Science*. https://doi.org/10.1007/s13194-022-00487-5
Kourany, J. (2016). Should some knowledge be forbidden? The case of cognitive differences research. *Philosophy of Science, 83*(5), 779–790.
Kuorikoski, J., & Pöyhönen, S. (2012). Looping kinds and social mechanisms. *Sociological Theory, 30*(3), 187–205.
Laimann, J. (2020). Capricious kinds. *British Journal for the Philosophy of Science, 71*(3), 1043–1068.
Langton, et al. (2012). Language and race. In G. Russell, D. Graff, Fara, et al. (Eds.), *The Routledge companion to philosophy of language* (pp. 753–767). Routledge.
Leslie, S. J. (2017). The original sin of cognitive: fear, prejudice, and generalization. *Journal of Philosophy, 114*(8), 393–421.
Lilienfeld, S. O., & Arkowitz, H. (2012). Is there really an autism epidemic? *Scientific American Special Editions, 17*(4s), 58–61.
Liu, K. Y., King, M., & Bearman, P. S. (2010). Social influence and the autism epidemic. *American Journal of Sociology, 115*(5), 1387–1434.
Lowe, C. (2021). *Self-fulfilling science*. De Gruyter.
McNab, A., & Dutton, K. (2014). *The good psychopath's guide to success*. Penguin Random House.
Mallon, R. (2018). Constructing race: racialization, causal effects, or both? *Philosophical Studies, 175*(5), 1039–1056.
Merton, R. K. (1948). The self-fulfilling prophecy. *The Antioch Review, 8*(2), 193–210.
Merton, R. K. (1968). *Social theory and social structure* (2nd ed.). Free Press.
Miller, S., & Selgelid, M. J. (2008). *Ethical and philosophical consideration of the dual-use dilemma in the biological sciences*. Springer Netherlands.
Miresco, M. J., & Kirmayer, L. J. (2006). The persistence of mind-brain dualism in psychiatric reasoning about clinical scenarios. *American Journal of Psychiatry, 163*(5), 913–918.
Moreau, S. (2020). *Faces of inequality: a theory of wrongful discrimination, Chapter 3*. Oxford University Press.
Mueller, C. M., & Dweck, C. S. (1998). Praise for intelligence can undermine children's motivation and performance. *Journal of Personality and Social Psychology, 75*(1), 33.
Nielsen, M. B., Rosander, M., Blomberg, S., & Einarsen, S. V. (2021). Killing two birds with one stone: how intervening when witnessing bullying at the workplace may help both target and the acting observer. *International Archives of Occupational and Environmental Health, 94*(2), 261–273.
Northcott, R. (2022) Reflexivity and fragility. *European Journal for the Philosophy of Science, 12*, 43.
Peters, U. (2021). Science communication and the problematic impact of descriptive norms. Forthcoming in the *British Journal for the Philosophy of Science*.
Pettit, P., & Goodin, R. (1986). "The possibility of special duties.". *Canadian Journal of Philosophy, 16*(4), 651–676.
Smiley, M. (2017). Collective responsibility. *The Stanford Encyclopaedia of Philosophy* (Summer 2017 Edition), Edward N. Zalta (ed.). https://plato.stanford.edu/archives/sum2017/entries/collective-responsibility/
Resnick, D. B. (1998). *The ethics of science. An introduction*. Routledge.
Solomon, M. (2017). On the appearance and disappearance of Asperger's syndrome. *Philosophical issues in psychiatry IV: Psychiatric nosology*, 176–186.

van Basshuysen, P., White, L., Khosrowi, D., & Frisch, M. (2021). Three ways in which pandemic models may perform a pandemic. *Erasmus Journal for Philosophy and Economics, 14*(1), 110–127. https://doi.org/10.23941/ejpe.v14i1.582

van Basshuysen, P. (2022). *Austinian model performativity*. *Philosophy of Science*. Forthcoming.

Young, I. M. (2011). *Responsibility for justice*. Oxford University Press.