Brit. J. Phil. Sci. **71** (2020), 1043–1068

Capricious Kinds

Jessica Laimann

ABSTRACT

According to Ian Hacking, some human kinds are subject to a peculiar type of classificatory instability: individuals change in reaction to being classified, which in turn leads to a revision of our understanding of the kind. Hacking's claim that these 'human interactive kinds' cannot be natural kinds has been vehemently criticized on the grounds that similar patterns of instability occur in paradigmatic examples of natural kinds. I argue that the dialectic of the extant debate misses the core conceptual problem of human interactive kinds. The problem is not that these kinds are particularly unstable but 'capricious'—their members behave in wayward, unexpected manners that defeats existing theoretical understanding. The reason for that, I argue, is that human interactive kinds are often 'hybrid kinds' consisting of a base kind and an associated status, which makes mechanisms that support patterns of change and stability systematically difficult to understand and predict.

1 Introduction

The question of whether the human world can be studied in the same way as the natural world has given rise to several heated controversies over the last two centuries. On the one side, proponents of the 'unity thesis' argue that

doi:10.1093/bjps/axy024

Advance Access published on March 14, 2018

investigation of the human world ought to be modelled closely on our scientific methods for the investigation of the natural world. On the other side, proponents of the 'difference thesis' defend the idea that the human world is importantly different from the natural world, and therefore requires methods fundamentally different from those of the natural sciences. Today, this highly polarized characterization looks somewhat outdated. For better or worse, grand claims about the nature of 'the natural' as opposed to 'the human' sciences have given way to a more nuanced investigation of specific scientific disciplines and approaches. Accordingly, the idea that the investigation of the human world requires a fundamentally different approach to that of the natural sciences has become a minority view in philosophy of science.

One of the last spokespersons of this view is Ian Hacking. For Hacking, the special status of the human sciences lies with the kinds they study: while the kinds that figure in the natural sciences are independent of (or, in Hacking's word, 'indifferent to') scientists' classificatory practices, some human kinds interact with the classifications scientists are using. Hacking terms these kinds human 'interactive' kinds and makes two controversial claims about them: (i) only human kinds are interactive kinds; (ii) human interactive kinds cannot be natural kinds. Both claims have been vehemently criticized—the first on the grounds that there seem to be non-human interactive kinds; the second on the grounds that, even if the phenomenon of interactivity could be limited to human kinds, this would not prevent them from being natural kinds. Despite finding Hacking's detailed case studies insightful, critics have converged on the conclusion that the general account of human interactive kinds that he extracts from them should be rejected.

This article aims to challenge this consensus. I argue that, although the critics correctly identify weaknesses in Hacking's argument, their focus on Hacking's suggestion that human interactive kinds are ontologically unstable fails to recognize the core conceptual problem of human interactive kinds. Accordingly, a shift in focus is due. I argue that we should stop understanding the question of whether human interactive kinds can be natural kinds as hinging on the issue of ontological stability. Instead, we should focus on the role of understanding mechanisms that support patterns of change and stability in our epistemic practices surrounding natural kinds. Pace Hacking's critics, considering human interactive kinds from this perspective potentially undermines their status as natural kinds, which has not been acknowledged in the extant discussion and merits further investigation.

In the following section, I recapitulate the extant discussion between Hacking and his critics. In Section Three, I point out how the dialectic of this discussion centres on the issue of ontological stability over time. I discuss two reasons why this way of framing the debate is misguided. First, it cannot account for the epistemic problems posed by human kinds that participate in

stabilizing, as opposed to destabilizing, feedback effects. Second, it is based on an oversimplified account of the scientific investigation and use of natural kinds. If these observations are correct, the assumption that human interactive kinds are problematic because their objects are unstable is wrong and has led the discussion astray. In Section Four, I develop an alternative understanding of human interactive kinds as hybrid kinds consisting of a base kind and an associated social status. I argue that such kinds pose specific difficulties for scientific understanding, which suggests some caution in thinking of them as natural kinds.

2 The Extant Discussion

2.1 Hacking's account of interactive kinds

Hacking's account of interactive kinds is motivated by a number of detailed case studies of psychiatric kinds like multiple personality disorder, child abuse, and schizophrenia (see Hacking [1986], [1988], [1991], [1992], [1995a]). Hacking notes that the studied phenomena develop over time in a very peculiar way that is unknown to the natural sciences. The objects of classification 'interact' with the classificatory schemes that are used to investigate them: classified individuals change, sometimes up to the point where the original classification is considered obsolete and thus revised. He calls these kinds 'interactive' (or 'looping') kinds. Phenomena studied in the natural sciences, by contrast, are unresponsive to our classificatory practices. Quarks, to use Hacking's familiar example, do not change in response to how we classify them.

We can understand the underlying process as a two-phase feedback loop. In the first phase, individuals react to the classifications that are (potentially) applied to them by changing their behaviour and characteristics. This phenomenon has been described in the sociological literature on criminal behaviour under the name 'labelling theory' (see, for instance, Schur [1971]). However, Hacking's account of interactive kinds features a second phase that has not been discussed in labelling theory. He suggests that the changes brought about by labelling can be so extensive as to render the original classification obsolete. Due to labelling effects, individuals might no longer correspond to the criteria or theoretical associations of the original classification. Upon noticing this development, those in charge of the classification (for instance, scientists or politicians) may decide that the mismatch is serious enough to necessitate a revision of the definition or theoretical understanding of the classification. Hence, in the second phase, the change in individuals' behaviour or characteristics feeds back into the understanding of the classification used to describe them.

Hacking's ([1999], pp. 113–4) discussion of the changing symptom profile of schizophrenia provides a good illustration of this process. He describes two iterations of the feedback loop, each of which features the two phases described above. According to Hacking, when the diagnosis of schizophrenia was first introduced, experts emphasized 'flat affect' and considered auditory hallucinations a minor problem that was not specific to schizophrenia. Auditory hallucinations being such an 'unproblematic' symptom, large numbers of people classified as schizophrenic expressed and reported them to their doctors. As a result, auditory hallucinations were found to be universal among schizophrenics when the classification was operationalized about thirty years later, and were therefore established as a major diagnostic criterion. This is the first iteration of the feedback effect. A second iteration occurred as schizophrenia became a decreasingly 'fashionable' diagnosis that individuals tried to avoid. Individuals stopped reporting auditory hallucinations; auditory hallucination ceased to be a widespread characteristic of people diagnosed with schizophrenia, and was successively de-emphasized as a diagnostic criterion.

Hacking makes two controversial claims about interactive kinds. He argues (i) that only human kinds are interactive kinds and (ii) that human interactive kinds are not natural kinds. Some clarifications are in order before we proceed to the criticism of Hacking's account. First, although Hacking often seems to refer to human kinds in general, he is not committed to saying that all human kinds are interactive. To avoid confusion, I will refer to those human kinds that are subject to the feedback effects described above as 'human interactive kinds'. Second, given the controversy about the concept of natural kinds, we need to know what concept is at issue in this discussion. Hacking's ideas about natural kinds are sketchy and—including kinds like mud (see his [1995b], p. 352)—unusually permissive.[1] Hacking's critics recognize this, but argue that there is a substantial question as to whether human interactive kinds can be natural kinds according to more orthodox understandings of natural kinds that include biological species as paradigmatic examples (see, for instance, Boyd [1991]; Dupré [1993]; Millikan [1999]). I put aside for now the larger debates about what natural kinds are and whether species qualify, and simply accept the critics' assumption that species are paradigmatic natural kinds. I will come back to the account of natural kinds underlying this debate in Section Three.

2.2 Classificatory feedback in non-human kinds

Hacking's claims have been subject to extensive criticism. Critics have invoked a variety of non-human kinds that allegedly participate in the same feedback

[1] In later work, Hacking ([2007]) distances himself from the notion of natural kinds altogether, arguing that the concept has outlived its usefulness.

effects as human interactive kind, including kinds of bacteria, marijuana plants, and livestock (Douglas [1986]; Bogen [1988]; Cooper [2004]). The most detailed case has been made with respect to domestic dogs (Khalidi [2010], pp. 345–6). According to Muhammad Khalidi, research suggests that the process by which the species domestic dog diverged from wolves consists of many iterations of the two-phase feedback effect described above. In the first phase, individuals classified as tame were selectively bred, producing increasingly tame individuals over time. In the second phase, upon recognizing that extant individuals do not conform to the existing classification of them, humans revised their classifications (for instance, from wolf to domestic dog, and later from domestic dog to particular dog breeds). These examples are not only used to reject Hacking's first claim that only human kind can be interactive, but are frequently taken to challenge his second claim that human interactive kinds cannot be natural kinds. As Cooper ([2004], pp. 74–7) points out, many of these examples qualify as natural kinds not only on Hacking's own, somewhat idiosyncratic account, but on many non-essentialist accounts of natural kinds that accommodate species as paradigmatic examples. Accordingly, it looks like the classificatory feedback effects that Hacking identifies as unique to human kinds in fact produce similar patterns of ontological instability in paradigmatic examples of natural kinds This would imply that both of Hacking's claims are false.

Hacking's staple response to this objection is to insist that the examples above do not qualify as interactive kinds on his view because the objects in question lack awareness of their classification (see, for instance, Hacking [1997], p. 15). Critics have pointed out a number of problems with this response. First of all, if awareness of one's classification is a necessary feature of interactive kinds, some of Hacking's own examples no longer qualify. Hacking ([1995b], p. 374) suggests that although young children and individuals with severe autism might be unaware of how they are classified, they might nevertheless participate in classificatory feedback that involves 'a larger human unit, for example, the family'. The idea seems to be that individuals who are unaware of how they are classified might nevertheless respond to the classification indirectly, for instance, by responding to family members or caretakers who are aware of how the individual is classified. This implies that awareness of one's classification is not a necessary feature of interactive kinds.

Second, it has been argued that even if Hacking would consistently restrict his account of interactive kinds to kinds whose members are aware of their classification, he has trouble explaining why these kinds cannot be natural kinds. While change in reaction to becoming aware of one's classification might be specific to humans, it is not clear how this makes human interactive kinds different from the examples of natural kinds discussed above. According to Cooper ([2004], p. 79), in order to make this claim, Hacking would have to

assume that classificatory feedback via awareness is of 'greater metaphysical significance' than the classificatory feedback we find in other kinds. Khalidi ([2010], p. 352) makes the same point with respect to feedback effects that are generated phylogenetically, via selective breeding. He argues that Hacking provides no reason why these phylogenetic feedback effects do not have the same philosophical implications as feedback effects that are created ontogenetically, via awareness.

In other words, both critics agree that even if Hacking stipulatively restricted the concept of interactive kinds to kinds whose members are aware of their classifications, he would still have to face two challenges. First, he would have to exclude some of the examples he previously described as interactive kinds from that category. Second, and more importantly, he would still owe a justification for the claim that human interactive kinds cannot be natural kinds. If Hacking wants to use the notion of interactivity to defend the idea of a fundamental difference between the human sciences and the natural sciences, an *ad hoc* emphasis on awareness will not do. Instead, so the critics suggest, he has to point to an ontological peculiarity of human interactive kinds that disqualifies them as natural kinds. Otherwise, his argument that human interactive kinds cannot be natural kinds fails. I will suggest that these objections, although correct, are somewhat beside the point: their focus on an ontological facet of Hacking's account (instability over time) obscures the main conceptual problems of human interactive kinds. To show this, we need to discuss the premises of the above criticism in more detail, beginning with the underlying account of natural kinds.

3 Natural Kinds and Ontological Instability

What, if anything, could prevent human interactive kinds from being natural kinds? The critics' comparison of human interactive kinds with biological kinds suggest that the difference—if there is one—has to be ontological. This assumption is reflected in Khalidi's question of whether human interactive kinds are 'real', as well as in Cooper's concern with whether classificatory feedback really marks 'a fundamental metaphysical distinction' between human interactive kinds and natural kinds. However, when we look at how both critics frame their investigation, a different aspect emerges. Cooper ([2004], p. 84) motivates her discussion with reference to the central epistemic role that natural kinds play in scientific inquiry:

> If human kinds are natural kinds then this suggests that accounts of laws, explanations, and the basis of sound inductive inferences, developed for the natural sciences, can be carried across into the human sciences. If human kinds are not natural kinds, then this will be a reason for thinking that distinct accounts will be required.

Similarly, Khalidi ([2010], p. 358) suggests that we should consider human interactive kinds as real, via adopting 'a weak realist view that considers as real any kind that plays an indispensable role in explaining phenomena, making successful predictions, and otherwise featuring in successful inductive inference'. Both remarks suggest that the guiding motivation of the debate is not purely metaphysical interest, but the question of whether human interactive kinds can fulfil the epistemic role of natural kinds.[2] The critics' concern with the status of human interactive kinds as natural kinds is effectively an epistemological and methodological one: if human interactive kinds are natural kinds, we do not need to come up with radically new approaches to understand them—their investigation can simply be modelled on the methods and epistemic practices used in the natural sciences. This hope stands in sharp contrast with some of Hacking's remarks. He suggests that any attempt at investigating human interactive kinds in the same way as natural kinds is destined to fail, and that more suitable approaches are yet to be invented (see, for instance, Hacking [1997]). Against this background, we can understand the rejection of Hacking's account as an attempt to reassure us that the phenomenon Hacking describes is not as epistemically troublesome as he makes it out to be. To evaluate Hacking's claims, we need to understand what could possibly hinder human interactive kinds from being scientifically investigated and epistemically used in the same way as natural kinds.

On many occasions, Hacking suggests that the problem with using human interactive kinds as natural kinds has to do with the fact that they are unstable. In Hacking's words, human interactive kinds are 'on the move' or 'moving targets' (see, for instance, Hacking [1999], Chapter 4, [2006]). This idea resonates with the example of schizophrenia discussed in Section Two. There, it seemed that by classifying individuals as schizophrenic, investigators unleashed a process in which the classified individuals change until they no longer fit the original classification. The resulting epistemic problem seems to be described most clearly with respect to the kind child abuse. Here, Hacking suggests that there might not be 'a stable object [. . .] to have knowledge about' (Hacking [1995a], p. 61). The idea seems to be that members of human interactive kinds constantly change in virtue of feedback effects, and we are not able to acquire knowledge and make inductive inferences about objects that constantly change over time. Accordingly, Hacking's critics have focused on instability as a potential problem for human interactive kinds' status as natural kinds. Khalidi ([2010], p. 342), for example, suggests that human interactive kinds seem to pose an epistemological problem because

[2] Cooper and Khalidi develop these accounts in more detail elsewhere (see Cooper [2005], Chapter 2, [2007], Chapter 4; Khalidi [2013]).

'after successive iterations of the looping effect, it seems that we may no longer be dealing with the same thing we started with'.

In other words, the debate is essentially about whether the members of human interactive kinds are unstable in a way that precludes them from functioning epistemically as natural kinds. Hacking seems to affirm this claim. His critics reject the claim on the grounds that similar patterns of instability are not considered a problem in the many examples of non-human kinds presented above. Neither side of this debate, however, seems to consider the association between ontological stability and the epistemic role of natural kinds worthy of further scrutiny. In the following, I discuss two reasons for questioning this assumption. First, it is based on an account of natural kinds as vectors for projections and generalizations that is oversimplified. Second, it cannot account for the epistemic problems posed by human kinds that participate in stabilizing, as opposed to destabilizing, feedback effects.

3.1 Understanding instability

In order to bring into focus the assumptions about the relation between ontological stability and the epistemic features of natural kinds that form the background of the above discussion, we need to specify what kind of instability is considered a potential threat to natural kind status, and why. For that purpose, we first need to specify what sort of change we are talking about. As described above, there are two sorts of change involved in the classificatory feedback that characterizes human interactive kinds. There can be changes to the members of a kind, for instance, when the extension of the kind changes (new members join, extant members lose membership or cease to exist), or when the characteristics of the individuals within that extension change (members acquire new properties or shed old ones). Alternatively, there can be a change in the theoretical beliefs associated with the kind, such as when we discover new properties of the members and adapt our theoretical understanding to accommodate these. Although participants in the debate occasionally talk of kinds themselves 'changing' or 'being unstable', this terminology should be avoided because it is ambiguous between these two quite different processes: the change of members is something that happens in the world, the change of theoretical understanding is something we deliberately bring about. What participants in the debate mean when they talk of a kind being 'unstable' is that the members of the kind change in ways that require us to alter our existing theoretical understanding of the kind.

Note that not just any type of change among members constitutes this sort of instability. Change is abundant in the natural world and scientists understand, explain, and predict the behaviour of a great variety of objects that change over time, such as reactive chemical compounds, or animals that

undergo metamorphosis. Take the kind water (H_2O). We know a lot about the properties of this kind, for example, that it has a melting point of 0°C and a boiling point of 100°C. However, we do not think that these properties are fixed or absolute, but know that they change depending on atmospheric pressure. Accordingly, natural kinds can have properties that are theorized as changing under specific circumstances, just as the melting point and boiling point of water are theorized as changing relative to atmospheric pressure. Therefore, what we mean when we say the natural kind water is stable is not that instances of water do not change under differing circumstances. We mean that, over time, instances of water do not change or develop new properties that are at odds with our existing scientific understanding of water. This suggests that we need to be more precise when asking whether 'instability' prevents a kind from functioning as a natural kind category. The problem with human interactive kinds is not merely that the classified objects change, but that they change in ways unforeseen by our extant theoretical understanding of the kind. This is not the case for chemical compounds like H_2O.

The case is different for biological kinds like species. Here, existing members of a kind are constantly replaced by new members with slightly different properties. As a result, the set of properties that characterizes members of a species can be transformed over time—instances of domestic dog today are characterized by very different properties than instances of domestic dog 200 years ago. Hence, instances of a species can, in a sense, change properties in a way that is at odds with our existing understanding of the species at any given point. When critics liken the instability of a human interactive kind like schizophrenia to the instability of biological kinds like domestic dog, what they have in mind is this instability over time of the set of properties associated with a kind. The rich biological literature on species like domestic dog suggests that biological kinds are quite capable of facilitating prediction, explanation, and inductive inference, and thus epistemically qualify as natural kinds. Since members of human interactive kinds seem to change over time in much the same way as biological kinds, Hacking's critics conclude that it is implausible to claim that the latter can have natural kind status whereas the former cannot. They anticipate that Hacking might respond by arguing that members of human interactive kinds change at a significantly higher rate than members of biological kinds, and cannot have natural kind status for that reason. However, Cooper and Khalidi dismiss this point fairly quickly (Cooper [2004], p. 79; Khalidi [2010], p. 350). They argue that even if it was evidently true that the members of human interactive kinds change faster than those of non-human kinds—which they doubt—this would not by itself explain why human interactive kinds cannot be natural kind categories. The difference is, after all, only one of degree.

But at this point, it seems like the critics' metaphysical concerns with natural kinds have gotten ahead of their underlying epistemic motivations. It might be plausible to argue that a gradual difference in the rate of change cannot establish a metaphysical difference between human interactive kinds and natural kinds. However, given the motivating epistemic concern with natural kinds, the dismissal seems somewhat hasty. From an epistemic perspective, the claim that human interactive kinds function as natural kinds because they change too quickly deserves serious consideration. After all, it seems perfectly reasonable to assume that a classification's ability to facilitate inductive inferences that allow us explain the behaviour of past instances and predict the behaviour of future ones depends crucially on how much its objects have changed in the meantime. A defender of Hacking could develop this point further by arguing that an epistemically significant threshold lies between the rates of change of members of biological kinds and those of human interactive kinds: while members of biological kinds change slowly enough for our scientific understanding to catch up, members of human interactive kinds outrun our efforts to theorize about them. Mallon ([2016], Chapter 7) explores this idea in some detail.[3] According to Mallon, whether we can have knowledge about a human interactive kind depends on whether scientists manage to increase the accuracy of their theories about members of the kind at a higher rate than the rate at which the members change. I call this the 'hare-and-tortoise' account of scientific understanding. Mallon ([2016], p. 166) illustrates this account in the case of biological species, arguing that scientists

> [. . .] can have knowledge of members of these changing kinds that allows us to engage in successful induction, prediction, explanation, and intervention because our capacity to gain accurate knowledge of these kinds can (sometimes) be far more rapid than the processes that underwrite biological change.

Certain aspects would need to be addressed further to develop this idea into a solid argument—for instance, how to operationalize rates of change and rates of theory improvement in a way that allows us to compare the two. But instead of doing that, I want to draw attention to the limitations of the accounts of natural kinds and scientific understanding that underpin this line of argument.

To begin with, the hare-and-tortoise account might suggest that there is an inverse relationship between the objects' rate of change on the one side, and our ability to develop scientific understanding of them as natural kinds on the

[3] Interestingly, Mallon ([2016], pp. 173–81) uses this proposal to defend rather than challenge the claim that human interactive kinds can function as natural kinds. He suggests that we should expect human interactive kinds to often develop at a slower rate than the theories we formulate to explain them, because stabilizing feedback tends to be more prevalent and powerful than destabilizing feedback.

other: the idler the objects of inquiry, the better they can be studied and function as natural kinds. However, there are reasons to think that change at a very slow pace poses problems of its own. Picking up Mallon's example of species, it would not be far-fetched to suggest that the slow rate at which most readily observable species evolve has hindered our understanding of evolution. If horses and birds had the generation time of bacteria, we might have arrived at a theory of evolution, and hence a better understanding of the natural kinds horse and bird, at a much earlier point in human history. Change at a very slow rate tends to escape our attention and if this happens, we fail to incorporate this aspect into our theoretical understanding of the kind. Admittedly, the relative stability of the members of many species has epistemic advantages: we can make a great number of predictions and inductive inferences about members of the kind, precisely because change occurs at a rate slow enough as to not interfere with them. However, our inductive inferences across wider time spans will be susceptible to error, and our explanations will lack information on phylogenetic history and evolutionary mechanisms. Overall, we would be inclined to say that, without these, our knowledge of the kinds in question is highly incomplete at best.

The example above shows that a slow rate of change of the members of a kind is by no means sufficient for the kind to facilitate scientific understanding. Other examples suggest that a relatively slow rate of change is not necessary for acquiring scientific understanding either. Consider bacteria. For some strains of bacteria, an individual can within thirty hours grow into a population in which every single base pair in the genome has mutated thirty times.[4] It seems unlikely that scientific theories about bacteria really approach accuracy at a faster rate than that. Fortunately, scientists working on these organisms do not start out from scratch, but can draw on theoretical resources from other areas. For example, much of the knowledge applicable to bacteria is derived from the study of species that change at a less breath-taking speed, such as fruit flies. Additionally, experimental setups can be used to limit possible causes of change and to ease the process of tracking members of a specific strain without having to identify each bacterium on the basis of shared characteristics, as was achieved by the development of the pure culture method in microbiology (see O'Malley [2014]).

These arguments suggest that the hare-and-tortoise account that motivates the focus on instability is overly simplistic. Scientists' ability to improve the accuracy of their theories does not simply stand in inverse relationship to the studied objects' rate of change, but depends on a host of factors, such as the possibility of making relevant observations, the ability to draw on existing understanding of underlying mechanisms, and the opportunity to study

[4] See (Pray [2008]).

objects under laboratory conditions. Accordingly, when deciding how well human interactive kinds can fulfil the epistemic role of natural kind categories, all these factors need to be taken into consideration. This point has not been explicitly addressed in the extant discussion on human interactive kinds, which focuses mainly on stability.

3.2 The problem of stabilizing feedback

The second problem with focusing on ontological instability as a crucial feature of natural kinds is that this view cannot account for the epistemic challenges posed by human classifications that are stabilized, rather than destabilized, by classificatory feedback. Hacking tends to focus on case studies where classificatory feedback makes individuals 'outgrow' existing classifications, such as the example of schizophrenia discussed above. Call this type of classificatory feedback 'destabilizing' feedback. However, there is a second type of classificatory feedback—'stabilizing' feedback—that achieves the contrary result: labelling effects reinforce properties associated with a classification, which is then interpreted as support for the existing classificatory practice. Standard examples in labelling theory describe such a process. They suggest, for instance, that the fact that someone has been labelled a criminal plays a role in their engaging in further criminal behaviour (see, for instance, Lemert [1951]; Becker [1963]; Chiricos *et al.* [2007]; Worrall and Morris [2011]). If the confirming labelling effects of a particular category are powerful enough, members of the category will generally conform to the properties associated with the category to a higher degree than they would have had, had they not been labelled. In response, those in charge of the classification might interpret the fact that individuals fit their labels so neatly as confirmation of the classificatory practice. In keeping with Hacking's metaphor, we might say that human kinds that are subject to stabilizing feedback are 'held in place' rather than 'sent on the move'.

For someone who believes that ontological instability is the main threat to human interactive kinds' status as natural kinds, stabilizing and destabilizing feedback effects need to be treated radically differently. While destabilizing feedback prevents human kinds from being natural kinds, stabilizing feedback would presumably make them more suitable candidates for natural kind status. After all, if natural kind categories need to refer to stable objects in order to facilitate induction, explanation, and prediction, and stabilizing feedback provides us with such stable objects, it should enable at least some human interactive kinds to function as natural kinds. Murphy ([2006], pp. 267–70) makes an argument along these lines. He suggests that if the norms, social pressures, stereotypes, or medical opinions that facilitate stabilizing feedback persist over time, the resulting patterns of behaviour that

characterize a human interactive kind might 'freeze in place', thus making the kind perfectly suitable for inductive inferences. Accordingly, a proponent of the view that ontological instability is the main threat to natural kind status would have to hold one of the following claims: (i) the concept of human interactive kinds includes only kinds that are subject to destabilizing feedback, or (ii) the concept of human interactive kinds also includes kinds that are subject to stabilizing feedback, but this does not commit us to saying that the latter cannot be natural kinds. While Hacking's position on the matter is not entirely clear, from an epistemological perspective, both of the above claims should be rejected.[5] The reason for this is that the epistemic challenges posed by stabilizing feedback can be substantial, and are in some respects more detrimental to the acquisition of scientific knowledge than the challenges associated with destabilizing feedback.

The debate on the causes of differences between men and women is a notorious case in point. As already noted in Mill's ([1984]) *The Subjection of Women*, the crux in this debate is that, for many observed behavioural or psychological differences between men and women, we have trouble identifying whether they are due to 'nature' or due to 'society'—in other words, whether it is due to natural, biological differences between men and women or due to differences in social upbringing and differential social constraints and opportunities. If the latter factors play a role (as we now have plenty of evidence to believe), it is very compelling to think of men and women as human interactive kinds that are subject to stabilizing feedback effects. We can imagine the underlying two-part feedback mechanisms operating in the following way: In the first part, individuals are born into a society that has certain preconceived ideas about men and women (for instance, that there are natural differences between them that not only determine their distinct morphological features, but also differences in character, abilities, and preferences). The society socializes individuals and arranges social institutions in accordance with these preconceived ideas. As a result, individuals classified as men or women continuously encounter differential social expectations and constraints and, over time, develop behaviour patterns, character traits, and abilities suitable to their circumstances—they come to fit their classification. In the second part, the fact that individuals classified as men or women squarely conform to these preconceived understandings is interpreted as evidence for the adequacy of the existing classificatory practice and its theoretical associations. It looks like men and women do naturally differ in character, ability, and preferences. This feedback mechanism is iterated as scientific testimony to the existence of such natural differences between men and

[5] Hence, I am not suggesting here that Hacking's critics are guilty of misinterpretation by wrongly attributing to him either (i) or (ii). At least with respect to (i), careful readers will find passages that support it, as well as passages that undermine it (see his [1999], p. 34 versus his [1995b], pp. 369–70).

women emerges. Scientific testimony strengthens the associated labelling effects, which is again, in turn, interpreted as confirmation of the classificatory practice and the theoretical understanding that underpins it. Due to these classificatory feedback effects, scientists came to firmly understand men and women as natural kinds that facilitate explanation and prediction not only of anatomical features, but also of a broad range of behavioural and psychological characteristics.

Assuming that this story is more or less accurate, we can see how stabilizing feedback effects not only obscured and facilitated the oppression of women, but also contributed to an erroneous understanding of the kinds men and women.[6] Many explanations facilitated by this understanding have been either false or substantially incomplete. Moreover, since the theoretical understanding suggested that differences between men and women are largely invariable across different societies, predictions and inductive inferences made on its basis were unreliable. In other words, the example above suggests that human kinds that are subject to stabilizing feedback can make for very poor natural kinds.

But more than that, there is reason to believe that human kinds that are subject to stabilizing feedback are, in some respects, worse candidates for natural kind status than human kinds that are subject to destabilizing feedback. Destabilizing feedback is, in some sense, transparent. The fact that classified phenomena resist and undermine our classificatory practices rubs our nose in the fact that the classifications we are using are based on an inadequate understanding of the phenomena in question. Stabilizing feedback, by contrast, is opaque. The apparent success of our classification can lull us into a false sense of security about the adequacy of the theoretical understanding that underpins the classificatory practice. If these observations are correct, and stabilizing feedback is at least as, and arguably more, epistemically challenging than destabilizing feedback, the assumption that human interactive kinds are problematic because their objects are unstable is wrong. Instead, the case of gender differences suggests that the problem is down to an inadequate understanding of the underlying determinants of change and stability in members of the kinds—only when we understand the mechanisms that support patterns of change and stability among the members of a kind are we in a position to provide accurate explanations and make inductive inferences across a variety of contexts.

3.3 Summary

Putting together the observations from the previous sections, the assumed connection between ontological stability and the epistemic features of natural kinds starts to look rather fragile. Section 3.1 suggests that the focus on ontological

[6] This is not to say that stabilizing feedback alone is responsible for the poor epistemic outcome. Other factors, such as bias on the part of an overwhelmingly male research community, have arguably played an important role (see, for instance, Longino [1990]).

stability reflects an overly simplistic hare-and-tortoise account of scientific inquiry and natural kinds. The case of stabilizing feedback in Section 3.2 corroborates these findings. It suggests that using ontological stability as a chief criterion for natural kind status may leave us with an epistemically thin and potentially misleading understanding of the kinds in question. Fortunately, it also indicates where a more nuanced understanding can be found: our epistemic practices surrounding natural kinds require knowledge of the causal processes that support patterns of change and stability in the classified objects. In order to be able to explain, predict, and make inductive inferences about the behaviour of members of a kind, we not only need to know that members typically display certain patterns, but also what produces them. In other words, natural kind categories should be understood not simply as vectors for projections and generalizations, but as analytic tools that incorporate assumptions about the causal mechanisms that constitute the kind.

These insights apply neatly to the example of domestic dog we started out with. Proponents of the hare-and-tortoise account suggest that domestic dog qualifies as a natural kind because change in the set of properties associated with this kind occurs at a pace slow enough for our understanding to 'catch up' and produce accurate explanation and predictions. The discussion above suggests that something different is going on. It suggests that domestic dog is a natural kind because we understand sufficiently well the evolutionary mechanisms by which members of the kind change their characteristics over phylogenetic time. Hence, while changes in the set of properties associated with this kind might, in one sense, overhaul our existing understanding of domestic dog—dogs in 200 years will probably look very different from dogs today—it is, in a different sense, perfectly in accord with our existing understanding. By contrast, if Hacking's description of the historical development of schizophrenia is correct, the reason we are taken aback by the instability of the set of properties associated with schizophrenia is that we have a wrong or incomplete understanding of the causal processes that support it. In other words, in trying to understand whether human interactive kinds can be natural kinds, we ought to stop putting so much emphasis on stability and instead ask if there is anything about these kinds that hampers our efforts to understand the underlying causal processes. In the following section, I argue that considering human interactive kinds from this perspective provides some reasons to be cautious about their status as natural kinds, thus rendering Hacking's account more convincing than his critics acknowledge.

4 Capricious Kinds

What, then, is the problem with human interactive kind, if not unusual instability? I suggest that the problem has to do with their peculiar ontological

structure. Human interactive kinds tend to have a dual nature: while we commonly think of human interactive kinds in terms of the properties that explicitly define the category, they can also be understood in terms of the social position that individuals occupy in virtue of being recognized as members of the category. In other words, human interactive kinds are often 'hybrid kinds'—they consist of what I call a 'base kind', constituted by the properties that define the category, and an associated 'status kind', constituted by the social position that individuals acquire *qua* being recognized and treated as members of the specific category.

The example of men and women from the previous section is useful to illustrate this idea. It is one of the few cases where the dual nature of a hybrid kind has been comprehensively conceptualized, in the form of the sex/gender distinction. Feminists have historically used the sex/gender distinction to tackle the idea that differences between men and women are biologically determined (see Mikkola [2017]). Roughly speaking, the distinction between sex and gender was meant to distinguish differences in biology ('sex') from differences that are due to culture and society ('gender'). Terminologically, this distinction is sometimes expressed by using 'male'/'female' to refer to sex categories, and 'men'/'women' to refer to gender categories, although I do not adhere to this terminology, but instead use 'men' and 'women' in the theoretically naïve sense that makes no such explicit distinction.[7] While there are many ways to spell out the idea of gender (for instance, in terms of gender identity, or socialized behaviour), the understanding that is relevant to my idea of a hybrid kind is best captured by the feminist slogan 'gender is the social meaning of sex'. This slogan expresses the idea that gender is a social position or role that individuals occupy in virtue of being recognized as members of a specific sex, an idea that has been developed in much detail by Haslanger ([2012]) and Ásta ([2011], [2013]). As a social position, gender is characterized by the norms, expectations, privileges, constraints, and opportunities that apply to individuals *qua* being recognized as members of a certain sex. In my terminology, sex is the base kinds, and gender (understood as a social position) the associated status kind. As Ásta ([2013]) argues in detail, the relationship between membership in the base kind and membership in the status kind is of a special and somewhat fragile nature—members of the base kind come to occupy the social position that characterizes the status kind only if they are recognized as members of the base kind, and individuals who are wrongly believed to be members of the base kind might nevertheless come to occupy the associated social position. Although this relationship does not guarantee complete coextension of the base kind and the status kind, the properties of the base kind and the properties of the status

[7] See (Saul [2006]) for an argument that ordinary speakers do not distinguish sex from gender.

kind are associated reliably enough to suggest that the terms 'men' and 'women' refer to hybrid kinds—they are commonly understood as, and often succeed in, distinguishing people on the basis of biological characteristics, yet they also unwittingly track an associated distinction in terms of social position.[8] On this account, the distinction between sex and gender can be understood as an attempt to conceptualize the hybrid nature of the human categories men and women, with 'sex' denoting the base kind and 'gender' the associated status kind.

While Haslanger and Ásta use this perspective primarily to develop a detailed metaphysical understanding of gender and other status kinds, I am more interested in what it tells us about the prospect of using human interactive kinds as natural kinds. I think the classificatory feedback effects described by Hacking can be understood as feedback effects between a base kind and the respective status kind. By being classified as members of a human category defined in terms of certain base properties, individuals come to occupy a specific social position (become members of the corresponding status kind) that is characterized by specific norms, expectations, constraints and opportunities, and that influences how others relate to them as well as how classified individuals relate to themselves. In virtue of these features, membership in the status kind can affect the characteristics of classified individuals, which may stabilize or destabilize our theoretical understanding of the base kind. In the remainder of the article, I argue that understanding human interactive kinds as hybrid kinds should make us wary about treating them as natural kinds. The reason for this is that hybrid kinds are susceptible to two problems that complicate their functioning as natural kinds: (i) biased conceptualization, which theorizes about the base kind whilst disregarding the status that is imposed onto members of the base kind; and (ii) difficulty conceptualizing, explaining and predicting the social status that is associated with a base kind.

4.1 Biased conceptualization

Biased conceptualization describes a phenomenon by which we theorize about and investigate the base of a hybrid kind while paying little attention to the associated status. The argument in Section 3.2 suggests men and women had been conceptualized in a biased manner before the distinction between sex and gender was introduced. Similarly, reconsider Hacking's paradigmatic example of schizophrenia. Schizophrenia is commonly understood either in terms of a specific symptom profile, or in terms of an underlying neurological condition

[8] Note that Haslanger and Ásta would probably disagree with this characterization—they suggest that 'men' and 'women' should better be understood as referring exclusively to the associated status kinds. See (Saul [2006]) for a discussion.

that is assumed to produce these specific symptoms (Murphy [2006]). Yet the category schizophrenia also picks out a status kind, which is a specific position in a network of social relations that individuals occupy in virtue of being classified as schizophrenic. Hacking's discussion details how people diagnosed as schizophrenic are singled out for particular interactions and treatments, and are subject to a number of specific expectations, opportunities, and constraints. In fact, Hacking makes quite clear that it is this network of social relations that mediates classificatory feedback in the kind schizophrenia.

How does biased conceptualization threaten the natural kind status of human interactive kinds? In order for a kind to function as a natural kind, we need to have a sound theoretical understanding of the causal processes that underpin the properties associated with it. We want to know not only that members of the category typically behave in certain ways, but also why they typically behave in these ways and under which circumstances we should expect them to behave differently. However, when we conceive of a human interactive kind solely in terms of the base kind, without considering the associated status, causal pathways associated with the status disappear out of sight. If these causal pathways have a significant influence on the properties of classified individuals, undetected biased conceptualization will prevent us from developing the causal understanding that is necessary to use human interactive kinds as natural kinds. In other words, although biased conceptualization does not necessarily affect all human interactive kinds, or necessarily preclude a proper understanding of all those kinds affected by it, it is an unacknowledged potential hindrance to using human interactive kinds as natural kinds and, as such, needs to be addressed in the debate.

The previous discussion suggests that the categories men and women have been severely affected by biased conceptualization. Here, the focus on a biological conceptualization concealed the role social positioning played in producing observed differences. Accordingly, scientists prioritized the search for biological determinants of the observed differences (such as brain size and shape, or hormones) over the search for social ones (such as socialization or social structural constraints). The same might have been true for schizophrenia, if Hacking's description is correct and changing medical beliefs did play a significant role in the changing symptom profile. In this case, it seems that conceptualizing schizophrenia as a cluster of symptoms or as a neurological disorder, without taking into account the associated status, obscured changes in medical beliefs about schizophrenia as a possible cause for the changing symptom profile.

In several places, Hacking remarks on our tendency to 'biologize' or 'geneticize' human interactive kinds (see, for instance, Hacking [1995b], p. 353, [2006]). Although these remarks resonate somewhat with my idea of biased conceptualization, they are misleading in that they suggest that biased

conceptualization always necessarily involves human kinds that are conceptualized as biological. This is not the case—kinds that are explicitly conceptualized as social can be hybrid kinds affected by biased conceptualization, too. Other passages in Hacking align with this idea. He cautions that the classification woman refugee is associated with social and material factors that affect the characteristics of women thus classified ([1999], pp. 10–11), and that our tendency to think of children who watch television as a 'species', might reify the kind child viewer of television via classificatory feedback effects ([1999], p. 27). Unfortunately, Hacking does not say anything more concrete about effects of biased conceptualization in each case. But we can illustrate the idea with the example of the kind unemployed. Here, the base kind, understood as being without paid work but available to work, is explicitly defined with respect to social institutions. Nevertheless, being unemployed is also associated with a status that, among other things, involves social stigma. If the social stigma of people classified as unemployed has a crucial influence on their properties, biased conceptualization that only considers the base kind could lead to gaps in our understanding. There is some evidence that this has taken place with respect to health disparities between employed and unemployed people. O'Donnell *et al.* ([2015]), for instance, found evidence that stigma negatively affects the psychological and physical health of unemployed people, but also note that there is very little existing research on this hypothesis. Former studies, they argue, instead focus on factors like financial strain, or lack of time structure, social contact, and activity—all factors typically associated with the base kind rather than the status of the hybrid kind unemployed. As O'Donnell *et al.* observe, this perspective not only provides a limited theoretical understanding of the existing health disparities, it also obscures potential interventions, such as changing public perceptions of unemployment or teaching skills for coping with stigmatization.

4.2 Studying social status

If biased conceptualization was the only potential problem with using human interactive kinds as natural kinds, the solution would be fairly straightforward: simply identify the associated status and understand what feedback effects it has on classified individuals. Unfortunately, there are reasons to believe that understanding these statuses and their feedback effects is anything but straightforward. As Kuorikoski and Pöyhönen ([2012], p. 191) point out, although much of social science is limited to describing patterns of social life, only an understanding of the underlying mechanisms allows scientists to make inferences about counterfactual scenarios and enables them to extrapolate findings to new contexts and identify effective interventions. Hence, in order to be able to explain and predict how the status associated with a classification

affects classified individuals, we need to understand not only the social and psychological mechanisms that mediate feedback effects, but also the mechanisms that stabilize and modify the status over time. There are several factors that potentially complicate this understanding.

Consider first the feedback-mediating mechanisms. Several philosophers have provided extensive discussions of these mechanism, often illustrated with examples supported by social scientific research (Murphy [2006]; Kuorikoski and Pöyhönen [2012]; Drabek [2014]; Mallon [2016]). Accordingly, I will not repeat their points here, but simply point to the diversity of causal pathways that this literature has identified. Mallon ([2016], pp. 68–89), for instance, distinguishes three main pathways by which classifications can lead classified individuals to change their behaviour: intentional change of behaviour, automatic change of behaviour, and environmental construction. Each of these, he suggests, can occur via several different causal pathways. Intentional change, for instance, can happen via change in salient possibilities for action, or via strategic or non-strategic reasoning. Environmental construction involves processes such as transmission of culture and institutions, or modifications of the material and spatial environment. In addition to that, Kuorikoski and Pöyhönen ([2012], pp. 196–7) point out that classificatory feedback can operate with or without the individual being aware of the classification. They discuss examples showing how classificatory feedback can happen without awareness, through processes such as the alteration of the practical reasoning of classified individuals or the modification of other people's expectations towards classified individuals.

This literature suggests that, although it is possible to identify and, to some extent, empirically investigate the social mechanisms that mediate classificatory feedback, these mechanisms are quite varied and complex. To complicate things further, different mechanisms may pull in different directions, thus amplifying or attenuating their respective effects. For example, on the intentional pathway, the effect of me being classified as a criminal might be that the classification troubles me, so that I resolve to make special efforts to act lawfully in the future. At the same time, my efforts to do so might be frustrated by the structural and material constraints that affect me as someone classified as a criminal. I might, for instance, no longer be eligible for a variety of jobs, which makes earning a living via illicit activities a more compelling option.

In addition to that, attempts at understanding and predicting classificatory feedback are complicated even further by the fact that the social meanings associated with classifications may vary both synchronically and diachronically. At any given time, a classification can mean different things in different contexts and interact with other classifications. Consider again the example of men and women. In this case, the conceptualization and investigation of the associated status is relatively advanced, arguably more so than in any other

human interactive kind. In the last few decades, a comprehensive literature that theorizes gender as a status kind has emerged (see, for instance, Oakley [1972]; MacKinnon [1989]), followed by systematic empirical investigation into the associated determinants of differences between men and women. Fields like social psychology, for instance, now provide ample empirical support for activists' and critical theorists' long-held claim that psychological differences between men and women cannot be explained purely in terms of biology, but require consideration of their differing treatment and positioning in society (see, for instance, Eccles [1987]; Eagly [1987]; West and Zimmerman [1987]; Spencer *et al.* [1999]). However, simultaneously with these developments, a discussion has emerged as to whether a unitary category of women's gender is a useful category at all, given how racial, cultural, and class differences influence the positioning and experiences of individuals classified as women (see Spelman [1988]; Crenshaw [1989]; Butler [1999]; Mikkola [2006]; Stoljar [2011]). At the centre of this discussion is the observation that the specific social position that an individual occupies in virtue of being classified as a woman varies greatly depending on a number of other factors. These include the background culture in which the classification is used, other classifications that are applied to the individual, as well as not classification-induced social and economic factors.[9] The debate suggests that it might not always be possible to identify a unitary status associated with a certain classification. Instead, in order to understand the causal processes that support a human interactive kind, one needs to understand how the classification affects individuals in different circumstances and in interaction with other classifications.

In addition to that, the status associated with a classification may change over time. Again, the problem is not that the social meanings of classifications change at all, but that they change over time in ways that are difficult to explain and predict. Why did the Stonewall riots in 1969 in New York lead to a gay liberation movement that radically changed the status kind associated with the category homosexual? Historians can discuss the merits of different hypothesis to explain this event and its impact, but they have little way to empirically decide between them. Events like the rise of the gay liberation movement are the result of complex social and political processes that possibly involved a unique constellation of a myriad of factors that cannot be reproduced or tested under laboratory conditions. As a result, social scientists cannot explain or predict changes of the meanings associated with human kinds with any certainty.

[9] By 'not classification-induced factors' I mean factors that do not depend on the individual being recognized as of a certain kind, although the factors might be causally associated with a certain kind. For instance, many people are poor because they are working class, but their being poor is not (or not primarily) due to being classified as working class.

These are both familiar points in the discussion of social scientific methodology, yet their relevance to the question of whether human interactive kinds can function as natural kinds has not been explicitly addressed in the extant literature. In particular, they suggest that status kinds themselves may often make poor candidates for natural kinds. If we cannot explain and predict the social meanings associated with human classifications, we are in no good position to explain their respective classificatory feedback effects, or to make reliable inferences about what feedback effects the classification is going to bring about under different circumstances. Hence, although the above discussion does not establish that human interactive kinds can never function as natural kind categories—there might be cases where we have a firm understanding of the associated status, and the mechanisms facilitating feedback are few and well studied—it does provide some reasons to be cautious.

5 Conclusion

In this article, I discussed Hacking's heavily criticized suggestion that human interactive kinds cannot be natural kinds. I suggested that there might be more to Hacking's claim than his critics acknowledge, albeit not for the reasons Hacking identifies. Hacking suggests that interactivity is primarily a phenomenon of instability of the set of properties associated with a kind. His critics rightly object that interactivity thus understood does not preclude human interactive kinds from being natural kinds. I argued that both sides miss the core threats to natural kind status because they presuppose an oversimplified understanding of the epistemic role of natural kinds. Natural kinds are not simply vectors for projections and generalizations, but analytic tools that incorporate assumptions about the causal mechanisms that constitute the kind. At the same time, human interactive kinds tend to have an ontological structure that compromises their ability to fulfil this epistemic role. They can often be understood as hybrid kinds, consisting of a base kind and an associated status kind, and are subject to several features that potentially threaten their status as natural kinds. These include the tendency towards biased conceptualization, the diversity and complexity of mechanisms mediating classificatory feedback, and most importantly, the fact that there is reason to think that status kinds themselves make poor candidates for natural kinds.

What are the methodological implications of my account? Recall that the discussion so far has been characterized by two methodological positions. According to Hacking, the phenomenon of human interactive kinds supports the difference thesis. For him, the fact that human interactive kinds cannot be natural kinds implies that we need radically new and different methods for

understanding these kinds. His critics, by contrast, seem to support the unity thesis. By insisting that human interactive kinds can be natural kinds, they suggest that investigating these kinds is just 'science as usual'—we do not need methods that radically differ from those of the natural sciences. My own account locates the truth somewhere in between these two positions. Although there might be cases in that we understand the associated status and its feedback effects well enough to use a human interactive kind as a natural kind, there is reason to believe that some human interactive kinds will be unsuitable as natural kinds. Yet this need not imply that investigating these kinds requires a radically new methodology. Coming back to the example of the kinds men and women, extant work in this area suggests that many researchers are perfectly well aware of the challenges and have found different ways of responding to them. After the crucial initial step of theoretically distinguishing the status kind gender from the base kind sex, feminist theorists have offered an understanding of gender as diverse and context-specific (see, for instance, Spelman [1988]; Butler [1999]), or suggested to understand gender along a specific politically relevant dimension (see, for instance, MacKinnon [1989]; Haslanger [2012]). These accounts of gender might not have (and are often not intended to have) the inductive power that we typically associate with natural kinds. But they might nevertheless provide an adequate understanding of how particular social mechanisms produce properties associated with men and women in specific contexts, or elucidate aspects of gender that are of central importance in emancipatory politics. In other words, contrary to Hacking's claim, the challenges of human interactive kinds need not demand a radically new scientific methodology. In many cases, a better engagement with the resources that are already on offer will do.

Acknowledgements

I am very grateful to Samir Okasha, Karim Thébault, John Dupré, and Alexander Bird for helpful discussions of earlier versions of this article. I would also like to thank Sam Humphrey, Claas Braun, and Sven Feldkord for useful editing advice. Finally, I would like to thank two anonymous referees for their detailed and valuable feedback. I am kindly supported by the Arts and Humanities Research Council South, West and Wales Doctoral Training Partnership.

Department of Philosophy
University of Bristol
Bristol, UK
jessica.laimann@bristol.ac.uk

References

Ásta Sveinsdóttir [2011]: 'The Metaphysics of Sex and Gender', in C. Witt (*ed.*), *Feminist Metaphysics*, New York: Springer, pp. 47–65.

Ásta Sveinsdóttir [2013]: 'The Social Construction of Human Kinds', *Hypatia*, **28**, pp. 716–32.

Becker, H. S. [1963]: *Outsiders: Studies in the Sociology of Deviance*. New York: Free Press.

Bogen, J. [1988]: 'Comments on "The Sociology of Knowledge about Child Abuse"', *Noûs*, **22**, p. 65.

Boyd, R. [1991]: 'Realism, anti-Foundationalism, and the Enthusiasm for Natural Kinds', *Philosophical Studies*, **61**, pp. 127–48.

Butler, J. [1999]: *Gender Trouble*, London: Routledge.

Chiricos, T., Barrick, K., Bales, W. and Bontrager, S. [2007]: 'The Labeling of Convicted Felons and Its Consequences for Recidivism', *Criminology*, **45**, pp. 547–81.

Cooper, R. [2004]: 'Why Hacking Is Wrong about Human Kinds', *British Journal for the Philosophy of Science*, **55**, pp. 73–85.

Cooper, R. [2005]: *Classifying Madness: A Philosophical Examination of the Diagnostic and Statistical Manual of Mental Disorders*, Dordrecht: Springer.

Cooper, R. [2007]: *Psychiatry and Philosophy of Science*, Stocksfield: Acumen Publishing.

Crenshaw, K. [1989]: 'Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory, and Antiracist Politics', *University of Chicago Legal Forum*, **1989**, available at <chicagounbound.uchicago.edu/uclf/vol1989/iss1/8>.

Douglas, M. [1986]: *How Institutions Think*, Syracuse, NY: Syracuse University Press.

Drabek, M. [2014]: *Classify and Label: The Unintended Marginalization of Social Groups*, New York: Lexington Books.

Dupré, J. [1993]: *The Disorder of Things*, Cambridge, MA: Harvard University Press.

Eagly, A. H. [1987]: *Sex Differences in Social Behavior: A Social-Role Interpretation*, Hillsdale, NJ: Lawrence Erlbaum.

Eccles, J. [1987]: 'Gender Roles and Women's Achievement-Related Decisions', *Psychology of Women Quarterly*, **11**, pp. 135–72.

Hacking, I. [1986]: 'Making up People', in T. C. Heller, M. Sosna and D. E. Wellbery (*eds*), *Reconstructing Individualism: Autonomy, Individuality, and the Self in Western Thought*, Stanford, CA: Stanford University Press, pp. 222–36.

Hacking, I. [1988]: 'The Sociology of Knowledge about Child Abuse', *Noûs*, **22**, pp. 53–63.

Hacking, I. [1991]: 'The Making and Molding of Child Abuse', *Critical Inquiry*, **17**, pp. 253–88.

Hacking, I. [1992]: 'World Making by Kind Making: Child-Abuse for Example', in M. Douglas and D. Hull (*eds*), *How Classification Works: Nelson Goodman among the Social Sciences*, Edinburgh: Edinburgh University Press, pp. 180–238.

Hacking, I. [1995a]: *Rewriting the Soul: Multiple Personality and the Sciences of Memory*, Princeton, NJ: Princeton University Press.

Hacking, I. [1995b]: 'The Looping Effects of Human Kinds', in D. Sperber, D. Premack and A. J. Premack (*eds*), *Causal Cognition: A Multidisciplinary Debate*, New York: Clarendon Press, pp. 351–94.

Hacking, I. [1997]: 'Taking Bad Arguments Seriously', *London Review of Books*, **19**, pp. 14–16.

Hacking, I. [1998]: *Mad Travelers: Reflections on the Reality of Transient Mental Illnesses*, Charlottesville, VA: University Press of Virginia.

Hacking, I. [1999]: *The Social Construction of What?* Cambridge, MA: Harvard University Press.

Hacking, I. [2006]: 'Kinds of People: Moving Targets', British Academy Lecture, available at <www.thebritishacademy.ac.uk/sites/default/files/hacking-draft.pdf>

Hacking, I. [2007]: 'Natural Kinds: Rosy Dawn, Scholastic Twilight', *Royal Institute of Philosophy Supplements*, **61**, pp. 203–39.

Haslanger, S. [2012]: *Resisting Reality: Social Construction and Social Critique*, New York: Oxford University Press.

Khalidi, M. A. [2010]: 'Interactive Kinds', *British Journal for the Philosophy of Science*, **61**, pp. 335–60.

Khalidi, M. A. [2013]: *Natural Categories and Human Kinds: Classification in the Natural and Social Sciences*, Cambridge: Cambridge University Press.

Kuorikoski, J. and Pöyhönen, S. [2012]: 'Looping Kinds and Social Mechanisms', *Sociological Theory*, **30**, pp. 187–205.

Lemert, E. [1951]: *Social Pathology*, New York: McGraw-Hill.

Longino, H. E. [1990]: *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*, Princeton, NJ: Princeton University Press.

MacKinnon, C. [1989]: *Toward a Feminist Theory of State*, Cambridge, MA: Harvard University Press.

Mallon, R. [2016]: *The Construction of Human Kinds*, Oxford: Oxford University Press.

Mikkola, M. [2006]: 'Elizabeth Spelman, Gender Realism, and Women', *Hypatia*, **21**, pp. 77–96.

Mikkola, M. [2017]: 'Feminist Perspectives on Sex and Gender', in E. N. Zalta (*ed.*), *The Stanford Encyclopedia of Philosophy*, available at <plato.stanford.edu/archives/fall2017/entries/feminism-gender/>.

Mill, J. S. [1984]: 'The Subjection of Women, Essays on Equality, Law, and Education', in J. Robson (*ed.*), *Collected Works of John Stuart Mill*, Volume 21, Toronto: Toronto University Press, pp. 259–348.

Millikan, R. [1999]: 'Historical Kinds and the "Special Sciences"', *Philosophical Studies*, **95**, pp. 45–65.

Murphy, D. [2006]: *Psychiatry in the Scientific Image*, Cambridge, MA: MIT Press.

Oakley, A. [1972]: *Sex, Gender, and Society*, London: Temple Smith.

O'Donnell, A. T., Corrigan, F. and Gallagher, S. [2015]: 'The Impact of Anticipated Stigma on Psychological and Physical Health Problems in the Unemployment Group', *Frontiers in Psychology*, **6**, p. 1263.

O'Malley, M. A. [2014]: *Philosophy of Microbiology*, Cambridge: Cambridge University Press.

Pray, L. [2008]: 'Antibiotic Resistance, Mutation Rates and MRSA', *Nature Education*, **1**, p. 30.

Saul, J. [2006]: 'Gender and Race: Philosophical Analysis and Social Kinds', *Proceedings of the Aristotelian Society*, **80**, pp. 119–43.

Schur, E. M. [1971]: *Labeling Deviant Behavior*, New York: Harper Row.

Spelman, E. [1988]: *Inessential Woman*, Boston: Beacon Press.

Spencer, S., Steele, C. and Quinn, D. [1999]: 'Stereotype Threat and Women's Math Performance', *Journal of Experimental Social Psychology*, **35**, pp. 4–28.

Stoljar, N. [2011]: 'Different Women: Gender and the Realism–Nominalism Debate', in C. Witt (*ed.*), *Feminist Metaphysics*, Dordrecht: Springer.

West, C. and Zimmerman, D. [1987]: 'Doing Gender', *Gender and Society*, **1**, pp. 125–51.

Worrall, J. L. and Morris, R. [2011]: 'Inmate Custody Levels and Prison Rule Violations', *The Prison Journal*, **91**, pp. 131–57.