




## Kinds behaving badly: intentional action and interactive kinds

Sophie R. Allen<sup>1</sup> 

Received: 29 September 2017 / Accepted: 3 July 2018 / Published online: 9 August 2018  
© The Author(s) 2018

### Abstract

This paper investigates interactive kinds, a class of kinds suggested by Ian Hacking for which classification generates a feedback loop between the classifiers and what is classified, and argues that human interactive kinds should be distinguished from non-human ones. First, I challenge the claim that there is nothing ontologically special about interactive kinds in virtue of their members being *classified* as such. To do so, I reject Cooper's counterexample to Hacking's thesis that kind descriptions are necessary for intentional action, arguing that classification (although not necessarily linguistic description) is required for intentional action. Having considered ways to characterise the metaphysics of interactive kinds and the semantics of kind terms, I argue that the fact that classification facilitates intentional action makes human interactive kinds ontologically distinctive because of the anomalous nature of the change which the kind-classification brings about. I then briefly examine further distinguishing features of human interactive kinds.

**Keywords** Interactive kinds · Natural kinds · Intentional action · Homeostatic property clusters · Psychiatric kinds · Social ontology

Ian Hacking has drawn attention to a group of kinds which appear to have a very interesting feature. In the case of such kinds, there is a feedback loop between classifiers and what they are classifying: kind membership brings about changes in the individual members of the kind or their environment, and these changes bring about changes in the kind classification itself; and then the process starts again. In virtue of being members of a kind, the members of a kind alter what it is to be that kind (Hacking 1986,

---

✉ Sophie R. Allen  
s.r.allen@keele.ac.uk

<sup>1</sup> School of Politics, Philosophy, International Relations and Environment, Keele University, Chancellor's Building, Keele ST5 5BG, UK

1988, 1995, 1999, 2002, 2006, 2007a). In earlier work, Hacking called these *human kinds*, since most of the paradigm examples involve kinds whose members are human beings or which are associated with human social groups. Such kinds include *multiple personality disorder*, *autistic spectrum disorder (ASD)*,<sup>1</sup> *attention deficit disorders (ADD and ADHD)*, *bipolar disorder*, *senior citizen*, *child abuser*, *problem drinker*, *psychopath*, *partisan*, *illegal migrant*, *terrorist*, *democracy*, *rogue state*, *failed state*, and so on. Thus, initially Hacking drew a distinction between human kinds on the one hand and natural ones on the other.<sup>2</sup>

However, there is reason to broaden the category of kinds which involve a feedback loop between classifier and classified, and so in later work Hacking characterised the distinction as one between *interactive* kinds and those which are *indifferent* to their classification.<sup>3</sup> For example, Muhammad Ali Khalidi (2010) argues that feedback loops are more pervasive than Hacking thinks and do not only occur among human kinds.<sup>4</sup> For instance, classifying a breed as a *pedigree dog* brings about changes in the properties which individuals of the kind instantiate through selective breeding, inbreeding, selection for showing on the basis of temperament and so on, and these changes may bring about a change in the classification as a result. A dog breed which develops a physical disability due to inbreeding may result in the classification of the kind being changed, the range of properties characteristic of the breed being broadened for example, so that the problematic mutation can be bred out. Similarly, when bacteria which cause harmful diseases in humans are classified as such, the environment in which they find themselves changes because humans put a large amount of time and effort into eradicating them or reducing their effects; in turn, this permits mutations of the bacteria to survive which would not have done so in the environment prior to the classification. As the bacteria mutates, the classification is changed to include novel strains. Khalidi argues that Hacking has not done enough to establish that human kinds have a special status and that there is no philosophically significant difference between human kinds and other interactive kinds.

Although I agree with Khalidi that interactive kinds are more pervasive than Hacking originally thought, the aim of this paper is to challenge Khalidi's objection that there is nothing distinctive about human interactive kinds.<sup>5</sup> I will argue that, given the

<sup>1</sup> I will retain the label 'autistic spectrum disorder' rather than the newer 'autistic spectrum condition' simply because I think that the former is more readily recognisable by readers. The fact that the name has recently been changed—largely due to the actions of kind members—is an example of interactivity.

<sup>2</sup> For this reason, some authors treat Hacking as providing a demarcation principle between the social and natural sciences, with the social sciences being those which deal in human kinds (Drabek 2010). Since there are good reasons to think that the classes of human kinds and interactive kinds are not coextensive with each other (nor with the class of social science kinds), I will not be concerned with this aspect of Hacking's project in what follows.

<sup>3</sup> In more recent work (2006, 2007a, b), Hacking has gone further and rejected the notion of interactive kinds because of reservations he has about the existence of natural kinds. He still, however, accepts the existence of looping effects between a classification and what is classified.

<sup>4</sup> This claim is also made by Douglas (1986, pp.100–102), Bogen (1988) and Ereshefsky (2004); it is also specifically acknowledged by Hacking (1999, p. 106) who claims (in answer to Douglas) that he offers the 'simple-minded' reply that microbes interact with us and our classification of them 'but not because they know what they are doing'. I aim to expand upon this reply.

<sup>5</sup> The term 'human interactive kind' may turn out to be a misnomer since the distinction might not turn on whether the kinds are essentially associated with humans, but I will stick with the terminology for the

existence of interactive kinds, human interactive kinds are distinct from non-human ones because there is a wider variety of ways in which humans can generate feedback. In order to establish this conclusion, I will first discuss another key objection to Hacking due to Rachel Cooper (2004) who argues that Hacking fails to show that *classification* of the kinds in question is responsible for the changes in their individual members, and those around them, who generate the feedback; members of the kind would behave as they do were they not classified as such.<sup>6</sup> Contrary to this, I argue that classification sometimes has an essential role to play in facilitating novel intentional actions and, moreover, that a subject acting intentionally has a greater range of possible effects than a member of a non-human kind.

In the course of this discussion, I will presuppose the existence of some interactive kinds, although I will not be concerned with defending the distinction between interactive and indifferent kinds. My argument will be effective even if all paradigmatically ‘natural’ kinds turn out to be interactive and there are no indifferent kinds or, on the other hand, if interactivity is an artefact of our epistemic position with respect to kinds which are really indifferent. I will, however, have more to say about this latter possibility towards the end of this paper. Since I agree with Khalidi’s assessment (2010, 2015) that Hacking has not done enough to motivate treating human interactive kinds differently from other interactive kinds, I will not attempt to locate a reason for the distinction in Hacking’s work but will engage in a general discussion about what differences there may be which is only sometimes based upon what Hacking has to say.

First, I will examine Cooper’s objection to Hacking, specifically her argument against his claim that a kind description is required for intentional action. I will argue that even if she is right that a linguistic description is not required, intentional action still essentially requires some form of classification and so her objection leaves Hacking’s position *prima facie* intact. Second, I will briefly consider ways in which one might characterise the ontology and semantics of interactive kinds with a view to clarifying the status of human interactive kinds. Third, I will investigate whether and why the fact that kind classification facilitates intentional action distinguishes human interactive kinds from the rest. I will argue that the subject’s awareness of the kind that they are, and the awareness of others around them, can produce changes which generate a feedback loop. These changes are unique to human kinds when they arise because kind members or those around them are mistaken about the properties which they or other members instantiate, and when kind members fake the presence of properties which they do not instantiate (or pretend not to have properties which they do instantiate); such changes cannot be brought about by membership of a non-human interactive kind. Finally, I will consider some other ways in which human interactive kinds might be distinct from non-human ones and note some of the implications of my conclusion.

---

Footnote 5 continued

purposes of this paper. I will also leave it open whether any non-human animals are able to act in such a way as to meet the requirements to be classified on the ‘human’ side of the distinction.

<sup>6</sup> Cooper also explores other reasons for Hacking to say that human kinds are distinct from natural ones and fails to find them, but I will not discuss these here since my argument can succeed without them.

## 1 The role of classification in intentional action: Cooper's objection to Hacking

Rachel Cooper argues that Hacking fails to establish that 'our classificatory practices can affect human kinds in ways that they cannot affect natural [indifferent] kinds' (2004, p. 77). In particular, she picks on his argument that classification is necessary for intentional action, which might serve to establish that classification instigates feedback which is peculiar to human kinds. We can summarise Hacking's argument as follows<sup>7</sup>:

- (1) All intentional action requires descriptions.
- (2) Kind classifications provide novel descriptions (which would not exist without the kind classification).
- (3) [Suppressed premise] Some of the descriptions in (2) are the same as those required by some intentional actions.

SO: Some intentional actions could not occur without kind-classifications.

To disarm Hacking's argument, Cooper challenges his first premise (2004, pp. 80–82), which he supports with Elisabeth Anscombe's (1957) claim that intentional action is only possible by an agent acting under a description; that is, an action is only intentional if the agent in question can answer a 'why?' question by giving a reason for acting. For instance, I am typing at my keyboard because I intend to write a philosophy paper, I have a reason to do so; the action I intend to perform is the writing of the philosophy paper, rather than the same behaviour which could be redescribed in a variety of different ways, such as *making a tapping noise*, or *polishing the keys of my keyboard*, and so on. The latter are forms of *voluntary behaviour*, but they are not intentional actions of mine (Anscombe 1957, pp. 11–12). If this is so, then the existence of the relevant description is essential to the existence of the intentional action: no description, no intentional action. Thus, Hacking asserts (1999, p. 104), the existence of a classifying description for a kind K opens up novel opportunities for the members of K and those around them to act intentionally in ways which were not available to them prior to the classification.

Cooper argues that Anscombe's characterization of an action's being intentional only under a description is idiosyncratic and that Anscombe's position would be better expressed in terms of her own gloss on 'under a description' as a modern form of '*qua*'. My current behaviour is an intentional action *qua* writing a philosophy paper and not *qua* making a tapping noise. Moreover, Cooper adds, only within the tradition of ordinary language philosophy or given some verificationist assumptions would we assume that an action could not take place except if there were some description under which it fell (and under which the agent recognises it as falling). We do not need language in order to intend to act in a certain way (2004, p. 82). In order to illustrate this, Cooper gives the example of Ug, a man sitting in a cave in the Lower Palaeolithic era before the development of human language. Ug has flints which he is banging together. Despite his lack of language, there is a fact of the matter about whether Ug is banging the flints together in order to make a musical noise, in order to sharpen them, or in order to make a fire; his action is intentional *qua* music making rather than *qua*

<sup>7</sup> The summary of the argument is my own, based on what Hacking has said in various places, especially 1999, pp. 103–105. I have added premise (3).

fire-starting. Thus, we have intentional action although Ug would not be able to answer a question about why he was doing what he was. The description is not necessary to the intentional action, Cooper concludes, and thus Hacking is not entitled to say that the descriptions introduced with kind classifications are necessary for the existence of certain types of action.

However, although Cooper has given a plausible counterexample to the claim that an intentional action cannot occur without the agent having a description for it, I do not think that her thought experiment does much damage to the thesis that some intentional actions could not occur without kind-classifications. While she has illustrated the short-comings of the claim that the appropriate descriptive language is required for intentional action, she has not presented a case in which there is intentional action in the absence of any form of discrimination or classification. The general claim that some intentional actions require kind classification—and thus would not occur without it—can be supported more directly by amending Hacking’s argument to the following:

- (1\*) All intentional action requires classification.
- (2\*) At least some of the classification required by intentional action is kind classification.

SO: Some intentional action could not occur without kind classification.

To see why premises (1\*) is plausible, let us return to Ug in the cave with his flints. If Ug intends to act in such a way as to make music and not to make fire, he must be able to understand and represent the differences between these two outcomes; despite not possessing language, he is able to discriminate the type of action he wishes to perform and, furthermore, to recognise a failure in this respect. If his flint-crashing has the result of setting fire to the bear skin upon which he is sitting, that will be an unintended outcome for him, not only for the hypothetical linguistically able onlookers watching from the future. But to recognise the success (or failure) of his intended action, and to attach values to the different outcomes, Ug must be engaging in some sort of classification of the world in relation to what he could do and does do in it. We do not have a case of intentional action in the absence of classification, just classification in the absence of language.

On the other hand, if we think that classification of outcomes has not taken place in Ug’s case—that is, that he cannot conceive of a difference between the future in which he makes musical banging sounds and the future in which he makes a fire (nor can he distinguish the making of music from any other outcome, such as a continued silence)—then we are unlikely to say that an intentional action is being performed by Ug rather than just voluntary behaviour. Someone cannot be said to be intentionally making music if they have *no* conception of what music is.<sup>8</sup> While it is plausible to think that we do not need the words to describe an action in order to perform it intentionally, the agent does require the discriminative ability to represent one outcome as being preferable to others which might co-occur with it (that is, other ways in which the same behaviour could be described were the agent linguistically able), or which may occur instead.

<sup>8</sup> That is not to say that a would-be musician’s conception of music need be as sophisticated as our own might be. Ug simply needs to be aiming towards producing a sound or rhythm or musical effect, rather than another action entirely such as making a fire.

Although I think that it is plausible for pre-linguistic Ug to act intentionally, those philosophers who think that classification or representation *require* language will see no room to respond in the way I have: pre-linguistic Ug cannot be classifying or representing anything at all. However, if one thinks this, I would urge that the correct interpretation of Cooper's counterexample should be that Ug is not acting intentionally either. While we, as third party, linguistically able observers can interpret his behaviour *as if* it is intentional, we can only do so because we have a classification system with which to taxonomise his behaviour as being *qua* one aspect rather than another. If Ug does not have the cognitive capacity to understand and to differentiate between those aspects then it is hard to see how he could care about them or how his behaviour could be directed towards one fine-grained outcome rather than another. Third person ascription of an intention to Ug is not sufficient for his acting intentionally, just as ascribing intentional states to a snail or to a malfunctioning computer network is not sufficient for them having genuine reason to act.<sup>9</sup>

If linguistic ability is not required for rudimentary classification and representation (and yet the latter are required to act intentionally), then it may be possible for non-human animals to engage in intentional action if they have the requisite awareness of the outcome toward which they are aiming. A cat might sit by a mouse's hole because she wants to eat, rather than because she wants to block the doorway, or an elephant might push her calf forward because she wants to introduce him to the herd, rather than because she wants to help him to walk (or to inadvertently push him over). Despite her talk of actions as being 'under a description', Anscombe seems to think that animals are capable of intentional action, especially since in her discussion of '*qua*' she gives examples of animals acting *qua* one way rather than another. For instance, she suggests that a bird may land on a twig *qua* getting food, rather than *qua* landing in bird lime, making the former intentional and the latter not (1979, p. 221).

It is now plausible to think, in the spirit of Anscombe's account, that intentional action requires that the agent conceive of the behaviour which they are aiming to produce as being desirable or undesirable *qua* one fine-grained outcome rather than another (even when those outcomes are aspects or ways of describing one particular event<sup>10</sup>), but that intentional action does not require that the agent or anyone else be able to describe that outcome unless one already thinks that intentional states require

<sup>9</sup> Here I am presupposing some kind of realism about intentional states. There are those such as Dennett (1987) and Davidson (1970) who would—for different reasons—claim that having an intentional state is nothing over and above that state being ascribed to a subject; that is, that there is no real individuable property of the mind or brain which that intentional state ascription picks out. In those cases, opinion might be split: we might say that Ug does have reason to act if there is reason to attribute an intentional state of *wanting to make music* to him by third parties, in which case he is acting intentionally in the absence of awareness of his own classification; on the other, one might think that the absence of Ug's own awareness or linguistic ability is sufficient reason for not ascribing a reason to act to him, in which case he is not acting intentionally but engaging in voluntary behaviour. In either case, however, the classification or description is still required for the intentional action, although in the former case Ug's awareness is not required.

<sup>10</sup> When a person acts, the resulting effect is an event which can be described both as an intentional action and as voluntary behaviour. I will not digress to consider how one might characterise such particular events (or instances of voluntary or involuntary behaviour as Anscombe would call them). One could opt for a Davidsonian account of event identity such that particular events are identical if and only if they are spatio-temporally co-located (1985), but any ontological account in which the taxonomy of event descriptions or properties is more fine-grained than the ontology of particular events would do.

language. Thus, although Cooper is correct that intentional action does not require linguistic description, she is wrong that classification is not necessary for intentional action. The plausibility of the first premise (1\*) of the amended version of Hacking's argument is not affected by her counterexample.

Before we leave this discussion, three further points should be made about Hacking's account which will have a bearing upon how we interpret the significance of the relationship between classification and intentional action for the interactivity of kinds. The first concerns whether the kind-member's awareness plays a necessary role in the instigation of feedback loops; the second concerns the role of individual kind members in instigating feedback; and the third concerns the role of instances of intentional action and voluntary behaviour as the physical cause of further physical effects.

First, Hacking does not want to claim that the difference between interactive kinds and the rest is entirely determined by the kind-member's own subjective awareness of his or her classification, or even that such first person awareness is necessary for classificatory feedback to occur. Hacking maintains that there are cases where the feedback loop occurs because the classification alters the way in which third parties interact with members of a kind *K* in virtue of their being members of that kind, or that classification may result in structures being put in place which alter the way in which members of *K* fit into the social groups of which they are a part (1999, p. 103). For instance, special schools may be set up and the members of a kind isolated or integrated more as a result of their diagnosis, thus instigating changes in the properties which members of that kind possess. This aspect of interactive kinds is important because in some cases the subject's own awareness of their classification, or the ability to understand a fine-grained taxonomy of their behaviour, may be lacking. For example, the intentions of a non-verbal child with ASD may not change as a result of his being diagnosed with ASD because he may not understand that he has the diagnosis; rather, what may alter is the behaviour of people towards him and the environment in which he finds himself in the form of the institutions which have been put in place because such a diagnosis exists. The presence of such institutions and the behaviour of others towards members of a kind may result in changes in the individuals of the kind and changes to the classification as a result. The awareness of the individuals of a kind that they are of that kind is not necessary for a feedback loop to occur.

Second, although classification may change how individuals act and the properties they possess, and this forms the basis of the feedback loop, the change in classification is not usually brought about by changes in one individual kind member's actions, but is the result of a collection of changes in kinds members, those around them or in the environment. As Hacking notes, 'the interaction takes place in the larger matrix of institutions and practices surrounding this classification' (1999, p. 103). Thus, although the discussion of later sections will be principally concerned with how classification influences and facilitates novel actions by individuals, this is a way of explaining how the collective changes come about which alter kind classification itself.<sup>11</sup> It is unlikely that an individual acting alone would bring about changes in classification.

---

<sup>11</sup> I will not take a stand on whether the collective changes brought about by classification can be reduced to changes in the way individuals act and localised changes in the environment. I am grateful to an anonymous reviewer for encouraging me to make the point of this paragraph more clearly.

The third point to note is that saying that intentional action requires the subject to conceive of themselves as acting under a description, or as aiming to produce a certain kind of act, does not rule out someone's performing the very behaviour which could be *described* as their performing an intentional action in the absence of such awareness. For example, Ug can bang his rocks together in the absence of an ability to conceive of a preferred outcome or to represent his behaviour as one action rather than another. However, if Ug has no thoughts about what he is doing, Ug is simply performing a bodily movement. Furthermore, such behaviour can occur even in the absence of any third party's ability to recognise Ug as performing one action rather than another or to classify him or his actions belonging to a specific kind. The particular event of rocks banging together—considered in a coarse-grained way—can occur in the absence of its being classifiable at the time as a kind of action. The possibility of such situations is extremely relevant to this discussion however, since the unintentional *banging the rocks together* appears to have the same physical effects (and perhaps the same place in the causal network of properties in general) as the intentional action of *making music* and so it is not obvious what role *having an intention to make music* plays. I will consider this question in Sect. 3.

So far, the conclusions of this section raise more questions than they answer. If intentional action is not possible without some variety of kind classification of the action, either linguistic or non-linguistic, then what are the implications for interactive kinds? Why is the ability to act intentionally—and the existence of the taxonomy which that requires—important to the nature of the feedback loop which results? And does this make human kinds distinct from other interactive kinds? Members of interactive kinds who are aware of their classification instantiate an additional property in virtue of that fact: each self-aware member of K has a belief that he or she is a member of K,<sup>12</sup> while others—both members and non-members of K—have beliefs about the K-membership of K-individuals. But it is not obvious why those beliefs about kind membership should generate a feedback loop which contribute to changing what being K is. To clarify this issue, it will be helpful to have a better grasp of the ontology and semantics of interactive kinds, and it is to this that I will turn in the next section.

## 2 Some remarks on the ontology and semantics of interactive kinds

To better understand how different types of feedback occur, it will be useful to distinguish between kinds, their individual members<sup>13</sup> and the classificatory practice which isolates, or attempts to isolate, kinds for the purposes of classification, generalisation and prediction. If one is a realist about kinds, the kinds and their members are part of the ontological account of what the world objectively contains, while the classificatory practice involves understanding and describing the kinds and so the entities involved in

<sup>12</sup> I am presuming here that a kind member's awareness of being a member of that kind is manifested in their having a belief that they are of that kind, rather than in their having a non-conceptual or phenomenally conscious state of being of that kind. Neither Cooper nor Khalidi think that awareness on the part of kind members will sufficiently distinguish human interactive kinds from others; on this point I disagree.

<sup>13</sup> I will assume that the individuals which are kind members exist whether or not they are actually classified as members of the kind (or the kind even exists).



this practice are linguistic or conceptual. How these ontological and semantic aspects relate to each other is crucial to the understanding of human interactive kinds.

The discussion of this section is primarily intended to provide a framework within which to clarify the implications of human kind members (or those around them) being aware of their classification, and having beliefs about it,<sup>14</sup> for the instigation of novel actions; and thereby to explain how human interactive kinds can be causally efficacious in a manner which is not available to non-human interactive kinds. Justifying a particular philosophical account of interactive kinds over others would go beyond the scope of this paper although I will be forced to talk in quite specific terms. So I aim to make my chosen framework as neutral as possible, avoiding commitment to specific philosophical positions about the ontology or semantics of interactive kinds which would require such additional justification and restrict the applicability of what I have to say. Furthermore, I will attempt to show why, in most cases, accepting a different ontological or semantic account instead of one of those I adopt would not alter the conclusions I reach; cases in which the conclusions would diverge were the ontology to differ will be clearly signposted.

## 2.1 Ontology

I will not offer an opinion on whether or not interactive kinds are natural kinds because the answer to that question depends in part upon what one considers natural kinds to be and that debate is tangential to the issues at hand. Nevertheless, a helpful starting point for a discussion about the ontology of interactive kinds is with the more frequently discussed ontology of natural kinds.<sup>15</sup> In this debate, if one is a realist about natural kinds—that is, one considers natural kinds to exist objectively, independently of our theorising about them—then a key question concerns whether or not all or some natural kinds have essences. By this, one might mean that there is at least one property which all kind members have essentially or, to put the point in a slightly different way,<sup>16</sup> that the possession of a certain property, or of certain properties, is necessary and sufficient for being a member of that kind.

Given what has been said about interactive kinds so far, they do not *prima facie* seem to be good candidates for having essences: first, because their defining feature is that the properties associated with the kind change over time in virtue of kind members being classified as such; and second, because the criteria for membership of several potentially interactive kinds do not require that members of those kinds share all properties associated with the kind at any one time but only a subset of them. Thus, individual kind members need not share properties either over time or

<sup>14</sup> I am presuming here that a kind member's awareness of being a member of that kind is manifested by their having a belief that they are of that kind, rather than in their having a non-conceptual or phenomenally conscious state of being of that kind. Neither Cooper nor Khalidi think that awareness on the part of kind members will sufficiently distinguish human interactive kinds from others; on this point I disagree.

<sup>15</sup> A broad range of philosophical positions could be suggested to account for the metaphysics and semantics of interactive kinds. For further discussion in relation to natural kinds, see Beebe and Sabbarton-Leary (2010), Bird and Hawley (2011).

<sup>16</sup> The difference between these formulations in terms of necessity and essence is not relevant for what follows and so I will treat them as being interchangeable. See Fine (1994).

synchronously (although they may do so). The former feature of interactive kinds is uncontroversial, since changes in properties of the kind over time will lead to changes in which properties kind members instantiate, but the latter may need further explanation and examples.

Cases in which individuals do not need to instantiate all the properties associated with a kind in order to count as kind members are widespread among psychiatric kinds and social kinds, and have also been recommended for the taxonomy of species (viruses, for example<sup>17</sup>). Many psychiatric conditions have polythetic diagnostic criteria, such that an individual must display at least a certain number of features from a list in order to be classified as suffering from the condition. No one feature is required for kind membership,<sup>18</sup> or (in some cases) one feature is required but is defined sufficiently broadly that it is variably realised in the individuals who are affected. This allows for a high level of heterogeneity among kind members. A slightly different case is presented by social kinds which classify a largely heterogeneous collection of individuals on the basis that they satisfy the criterion of having one specific property (such as being a *lone parent*, *illegal migrant* or *homosexual*). Despite membership of these kinds being determined by kind members instantiating a specific property, these social kinds may still, ontologically speaking, be characterised as being a cluster of properties which tend to be caused by or co-instantiated with the primary characteristic property. This feature is what makes them useful for prediction and explanation. But we would not expect each member of the kind to instantiate all the properties of the cluster (in fact, as in the case of the kind *problem drinker* which will be discussed below, some of these associated properties might be incompatible with each other).

Given these considerations, it seems that those who are committed to essentialism about natural kinds will either be forced to say that interactive kinds are not natural, to take a more sophisticated account of essence from the one I have given, or to say that interactivity is an epistemic, rather than an ontological, feature. As part of this latter claim, essentialists may also challenge the apparent heterogeneity of the psychiatric and social kinds discussed above. For instance, the essentialist might criticise polythetic membership criteria and argue that they are simply indicative of the limitations of our understanding about the nature of certain kinds: we do not know which properties are essential to having schizophrenia, and so we rely upon lists of criteria which attempt to capture all and only individuals with schizophrenia; ontologically speaking, there may be a property which all kind members share. Because of this concern, I will explore an ontological account of interactive kinds which is compatible with essentialism (but also compatible with its denial), and will explore the implications of this concern later in the paper.

A potential alternative to essentialism<sup>19</sup>—which is nevertheless compatible with some kinds having essences—is to treat kinds as clusters of properties, a move which

<sup>17</sup> See van Regenmortel (1992), p. 263.

<sup>18</sup> There are many such examples in DSM-4 and DSM-5 (APA 1994 and APA 2013). For instance, major depressive disorder; bipolar disorder, autistic spectrum disorders (see footnote 36), schizophrenia, borderline personality disorder (to name just a few).

<sup>19</sup> I consider this to be an alternative only in the sense that it is compatible with denying essentialism. As noted, it is also compatible with kinds having essences. I will consider the implications of such kinds having essences in Sect. 5.

permits us to characterise interactive kinds as a species of dynamic kinds (that is, kinds which change over time) and as kinds which have heterogeneous individual members. There are several ways to do this, and which of these ways is the best one is currently a matter of some debate, but the differences between them will not be relevant to the discussion of interactive kinds and so readers may substitute their preferred theory into the discussion which follows this section. One, fairly popular, way in which we can characterise dynamic kinds is in terms of Richard Boyd's account of kinds as *homeostatic property clusters* (HPCs) that construes kinds as clusters of properties which are co-instantiated in virtue of an internal or external causal mechanism (1989, 1991, 1999, 2010).<sup>20</sup> Biological species are presented as paradigm cases of HPC kinds<sup>21</sup>: their members instantiate a cluster of properties in virtue of their environment which serves to ensure the likelihood that properties of the cluster will be co-instantiated and also proves inhospitable to property clusters which diverge from the norm (through genetic mutation for example). However, a change in the environment can lead to a change in the dominant properties of the cluster and so over time the properties which characterise the kind may change. One form of change occurs when members of a kind K gain or lose a property, while another occurs when the extension of kind K alters as different individuals start to instantiate the properties associated with K. Thus, the HPC conception of kinds permits us to give an account of dynamic kinds and we can extend this account to include interactive kinds.<sup>22</sup>

On an HPC account, or on other property cluster theories of kinds, the feedback loop of an interactive kind is generated because the classification of a property cluster as being kind K brings about changes in the properties of K members (or the properties of other entities in the environment), alterations which in turn bring about changes to what counts as a K. For instance, in the case of the tuberculosis bacillus, its environment changes as a result of its being isolated as the cause of a deadly disease in

<sup>20</sup> Alternatives to Boyd's account include Khalidi's suggestion (2015) that the properties associated with a natural kind are causally structured in certain ways, which he proposes to improve upon Boyd's account. How distinct these ontological views are depends upon how different Khalidi's causally structured properties are from Boyd's internal causal mechanisms, which is a matter that need not concern us here. I will retain Boyd's terminology while allowing that an internal causal mechanism which sustains the properties of a kind in a cluster may consist in those properties being causally (or otherwise) related. Hauswald (2016, pp. 213–217) suggests a similar causal property cluster view of dynamic kinds which would be consistent with what I have to say. For other supporters of HPCs or similar accounts, see Murphy (2006), Kuorikoski and Pöyhönen (2012), Bird (2015), Kistler (2016). Finally, one might think that it is sufficient to maintain that kinds are stable property clusters, and to require no further explanation of the cohesion of such clusters. This view is suggested by Häggqvist (2005) and Slater (2015), however this account is less useful in the current context where changes to the properties in a cluster are at issue, since it is not explicit about what the mechanism for changes in the properties in a cluster might be.

<sup>21</sup> See Boyd (*passim*) and Wilson et al. (2007). This claim is contested by Ereshefsky and Reydon (2015) who point to the use of the Phylo-Phenetic Species concept in microbiology (2015, pp. 973–974), but their counterexample need not delay us here.

<sup>22</sup> Ereshefsky and Reydon (2015) argue that Boyd's account fails as an 'overarching account of scientific kinds' (2015, p. 972) and suggest an alternative. In the current context, their criticisms are not relevant since they are directed at HPC kinds as a *general* account of scientific kinds. My claim here is that HPCs are useful for giving an ontological account of some varieties of kinds, namely those which are dynamic or interactive, in a way which is consistent with realism about such kinds. I am also remaining neutral about whether an alternative property cluster account of kinds would be more plausible, as noted above. It may be possible to rework what I have to say about interactive human kinds in terms of Ereshefsky and Reydon's positive theory, but I will not attempt to do so here.

humans, which in turn causes the bacillus to mutate (or, rather, permits the survival of certain mutations which would have not have survived previously); this change in turn requires that the classification be altered and that the environment be further modified in order to treat or to prevent the disease. In the case of individuals who are aware that they are classified as *problem drinkers*, their awareness of their classification adds a property to the cluster which the individual kind members instantiate, and other people's awareness of a kind member's predicament constitutes a change in the problem drinker's environment. These changes may prompt novel intentional action on the part of the drinkers themselves: for example, causing the individual drinkers to seek help for their drinking and to avoid alcohol (hence leading to the situation where the extension of the kind *problem drinker* includes those who do not drink alcohol); they may begin to hide their drinking; or they may start to drink more heavily as a result. Meanwhile, the awareness of those around them of their kind membership may also bring about changes: other people might begin to interact differently with a person whom they believe to be a problem drinker, they may explain his or her actions differently, be less likely to accept first-person ascriptions of motivations or reasons which the problem drinkers make,<sup>23</sup> they may try to stop the drinking or stage interventions, and they may set up rehabilitation facilities and encourage (or force) problem drinkers to attend them. All these have the potential to change what counts as a problem drinker by changing the criteria for membership of the kind or which individuals satisfy those criteria.

This HPC account of kinds is not sufficient for naturalism, in fact Boyd himself thinks that the naturalness of HPC kinds is discipline relative (2010), but it is consistent with naturalist realism according to which the existence of certain property clusters is objectively determined. On the HPC view, kinds can fail to be natural for two reasons: either the properties within a cluster are natural but the mechanism by which they are clustered is non-natural, or *both* the mechanism and the properties within the clusters are non-natural. This allows for two 'strengths' of non-natural or socially constructed kinds.<sup>24</sup> Furthermore, the HPC account is also neutral about whether any of the properties in an HPC are necessary or sufficient for kind membership, so it is consistent with some kinds having essences.<sup>25</sup>

I think that the HPC account, as it has been presented so far, conceals some presuppositions about the unified nature of the feedback mechanisms and the homogeneity of the properties involved in the property clusters which may, when addressed, substantially complicate the theory. But despite these shortcomings—which will be revisited later—the HPC account is an interesting and potentially useful characterisation, since it allows for the properties associated with a kind to change and at least postulates a

<sup>23</sup> For instance, a claim such as 'I only bought a can of beer in the shop because I needed some change for the parking meter' is less likely to be accepted at face value if one knows the speaker is a problem drinker.

<sup>24</sup> Although it may turn out that some human interactive kinds are socially founded (or constructed), and thus distinct from non-human interactive kinds on that basis, I am interested in this paper in exploring whether the distinction remains even if the properties and mechanisms associated with human kinds are 'natural', that is, not socially constructed. See Haslanger (1995), Haslanger and Saul (2006).

<sup>25</sup> One might go further here and maintain that the homeostatic (or other) mechanism which sustains the stability of the property cluster of a kind (and which permits certain changes in it) is itself an *essence*, even if the mechanism is external to the property cluster (Griffiths 1997, p. 188; Häggqvist and Wikforss 2017, p. 14). If this more liberal conception of essences is acceptable, then HPC theory is a form of essentialism about kinds and not in competition with it. I will not explore this question further here.

mechanism by which this change comes about. In light of this, I will adopt it for use in this discussion. Nevertheless, in this paper I will not defend the HPC over other accounts, since I do not intend to maintain that the HPC account is the only, or even the best, account of the metaphysics of interactive kinds: one could probably explain them just as well with an alternative account of property clusters, or perhaps with an essentialist account (although that would be harder). I am not advocating the adoption of HPC for all kinds, nor am I even claiming that it is the correct account of interactive kinds; rather, it is useful heuristic device with which to get to grips with the metaphysics of interactive kinds. Thus, the employment of HPCs is a less controversial move than it may initially seem.

## 2.2 Semantics

There are, broadly-speaking, two families of theories in contention to give an account of the semantics of natural kind terms: the causal theory of reference<sup>26</sup> and descriptivism. The popularity of versions of the former is based, in part, on the fact that the causal theory explains how we can use kind terms to pick out kinds, even though we do not have the requisite information about the properties of that kind; reference is fixed by ostension through paradigmatic examples or by non-essential descriptions of the kind, and the empirical investigation will reveal more about the properties which that kind has. At first glance, this account might seem well-suited to the semantics of interactive kinds: for instance, much of what we currently believe about the kind *ASD* might turn out not to be essential to people with *ASD*, but we want to be able to pick out people who have it regardless of this; similarly for psychiatric diagnoses such as *anorexia* or *bipolar disorder*. We can refer to non-human interactive kinds in a similar manner: *these* animals are *pedigree dogs*, or *these* microbes are the *tuberculosis bacillus*, even though the properties which we would attribute to them may turn out not to be essential to being a member of those respective kinds.

However, there is a stumbling block to the adoption of causal theories of reference to explain the semantics of interactive kinds, since causal theories are often thought to entail essentialism about kinds and, as was noted in the previous section, interactive kinds do not obviously have essences.<sup>27</sup> The question of whether causal theories of reference entail a non-trivial form of essentialism about kinds is fraught, and I do not have the space to do justice to it here.<sup>28</sup> Nevertheless, if the entailment holds, the causal theory of reference will be less attractive as an account of interactive kind terms than it would otherwise be.

<sup>26</sup> I will use the terms ‘theory’ here because it has become common to describe the causal account of reference as a ‘theory’, even though Kripke did not think of what he had presented as a theory (1980, p. 93).

<sup>27</sup> As also noted above, one could make essentialism work with interactive kinds, by adopting a four-dimensional account of the identity of essences or by treating the mechanism which sustains an HPC kind as an essence. Furthermore, HPC kinds are consistent with kinds having essences. Nevertheless, while we know so little about interactive kinds, it would be unwise to adopt a theory of the semantics of natural kind terms which entails that they have essences.

<sup>28</sup> There is extensive literature on this debate. For instance, in favour of the entailment see Kripke (1980), especially pp. 128–140, Haukioja (2015), Häggqvist and Wikforss (2017); those arguing against the entailment include Mellor (1977), Salmon (1981), and Mackie (2006, chapter 10).

Moreover, there is another important factor which we need to take into account at this point, since we are interested in the way in which beliefs about kind membership can influence intentional action, rationalisation and explanation. Even if the causal theory of reference is true for interactive kinds and, by using a kind term ‘K’, we refer directly to members of the extension of that kind term which is determined externally to the speaker, this externally determined kind K (the nature of which speakers may be ignorant) may only play a partial role in influencing the behaviour of kind-members and others who are aware of their classification. What matters to the way in which people intend to act or to the way in which they explain the actions of others is what they think that members of K are like, whether or not Ks actually satisfy the descriptions which people associate with being K. As the discussion of Sect. 1 emphasised, it is these descriptions from the individual agent’s perspective (or the perspective of others who believe she is K) which are crucial to facilitating intentional action and explanation.<sup>29</sup> But the agent’s perspective (and that of those around her) is fallible and may be incomplete. So when a kind member says or believes ‘I am a member of K’, this claim is ambiguous: ‘K’ may directly refer to whatever kind K actually is (or to its essence if it has one); or the speaker may also understand her claim to mean that she satisfies some or all of the descriptions which are associated with Ks, whether or not entities in the extension of the term ‘K’ actually satisfy these descriptions and even though each individual kind member may not satisfy all of them.

This discussion suggests that even if some form of causal theory of reference is correct for kind terms, there is also a need for a descriptivist account of the semantics of interactive kind terms when we are considering the influence of classification on action and action explanation, since the descriptions which are associated with a kind are relevant to action and its explanation. This descriptivism need not endorse the a priori applicability of specific descriptions to a kind (a claim which would lead to objections of the variety which led to the causal theory in the first place<sup>30</sup>) and can either be treated as running alongside the causal theory of reference for kind terms (leading to a version of semantic dualism or two dimensional semantics for kind terms) or as an alternative to the causal theory.<sup>31</sup> As with the metaphysical account of interactive kinds, I am eager to be as non-committal as possible in order that my conclusions have the broadest possible scope.<sup>32</sup>

<sup>29</sup> Although, as was noted in Sect. 1, it is an ability to classify rather than a linguistic ability to describe which is required for intentional action, I will talk in terms of descriptions in this discussion of the semantics of kind terms.

<sup>30</sup> See Kripke (1980).

<sup>31</sup> In her discussions of social kinds, Haslanger (1995) (also in Haslanger and Saul (2006)) makes a potentially useful distinction between the *manifest* concept of a kind (the concept one takes oneself to be applying), the *operative* concept of that kind (the concept which one actually applies in practice) and the *target* concept (the concept of the kind) (2006, pp. 98–99). She uses an externalist account of meaning based on the Kripke–Putnam causal theory discussed above to justify the claim that there is a target concept, or ‘something to be right about’ (2006, p. 110). This strategy in effect combines elements of descriptivism with the causal theory.

<sup>32</sup> Given these observations, the supporter of the causal theory of reference might want to reframe the phenomenon of interactivity in a different way: being classified as a kind K brings about changes in action and explanation in virtue of what agents believe about the nature of K (about which they may be mistaken), not just because they or others are Ks.

### 2.3 Ontology and semantics combined

So far, I have defended an ontological account of interactive kinds and a semantic account of interactive kind terms which I intend to employ: HPC on the basis that it is a useful heuristic device, and descriptivism because it fits best with the account of intentional action given in Sect. 1 even if one favours the causal theory of reference.

When we postulate or isolate a kind  $K$ , we aim to individuate entities according to whether they possess a specific set of properties or whether a set of descriptions applies to them. Following Putnam's account of the semantics of natural kinds (1970, 1975a), this set of descriptions can be called the *stereotype* of  $K$  ( $S_k$ ); the meaning or concept associated with  $K$  is or is based upon this set of descriptions. Satisfying these descriptions is not essential a priori to being a member of  $K$ . From an ontological point of view, as we have noted above, being of a specific kind  $K$  involves instantiating a set of properties associated with being  $K$ —let us say the set  $P_k$ —but specific individuals may fail to instantiate some of those properties and yet still count as being members of  $K$ .

How do the ontological and the semantic stories match up? It would be convenient, but overly optimistic, to presume that they do so perfectly such that all the descriptions in  $S_k$  at a specific time pick out properties in  $P_k$  at that time. We could call this perfect match *the ideal scenario*. However, there are at least three reasons to think that the ideal scenario is not likely to be the case: first, we cannot presume that we have perfect epistemic access to the properties in  $P_k$  which make up the HPC; second, because interactive kinds are a species of dynamic kinds, which properties are in  $P_k$  can change over time, so the descriptions in  $S_k$  might cease to apply until the stereotype is amended to take this into account; third, in the case of interactive kinds, the descriptions of  $S_k$  can also change over time as a result of feedback, so the properties of  $P_k$  at a certain time might not be those picked out at a later time.<sup>33</sup> Nevertheless, for simplicity, I propose to ignore the complications just outlined because we can still say some interesting things about intentional interactive kinds while presupposing that the ideal scenario does obtain; that is, that for kind  $K$  at a particular time, there is a one-to-one correspondence between the descriptions in  $S_k$  and the properties in  $P_k$ . Such idealisation is justified because if the relationship between  $S_k$  and  $P_k$  is more complicated, then this will only make the problems which I raise more acute.

<sup>33</sup> The first problem is common to any kind, especially at earlier stages of empirical investigation; the second is specific to dynamic kinds of which interactive kinds are a subset; whereas the third is specific to interactive kinds. The latter two correspond to the two directions of influence which make up the feedback loop in interactive kinds: the first in which changes in the kind lead to changes in the classification, and the second where changes in the classification bring about changes in the kind. Given that, by definition, interactive kinds change over time, it is sometimes useful to index  $S_k$  and  $P_k$  to times in order to capture the alterations in the descriptions of  $S_k$  and the properties in  $P_k$ . However, if we identify  $P_k$  with a dynamic or interactive kind, we might want to note a difference between the persistence conditions of stereotypes such as  $S_k$  and property clusters such as  $P_k$ : if  $S_k$  at time  $t_1$  contains different descriptions to  $S_k$  at  $t_2$ , then  $S_k$  at  $t_1$  is not identical to  $S_k$  at  $t_2$ ; however, the property cluster  $P_k$  which the former picks out ( $P_k$  at  $t_1$ ) is the same kind as  $P_k$  at  $t_2$ , that picked out by  $S_k$  at  $t_2$ . The kind perdures while the concept or stereotype associated with it is replaced.

### 3 How does classification affect intentional action?

#### 3.1 Acting

We are now in a better position to understand more precisely the role which classification, and the subject's awareness of it, plays in the evolution of interactive kinds. The awareness of a kind-member that she is of that kind  $K$ —that is, her having the belief that she is a member of  $K$ —involves her believing that some or all of the descriptions in  $S_k$  apply to her. In some cases, the descriptions themselves may be novel ones which did not exist prior to the postulation of  $K$ . For instance, Melanie Yergeau (2013) suggests that the contrasting descriptions 'lacking a theory of mind' and 'having a theory of mind' did not exist as psychological descriptions prior to the postulation of Autistic Spectrum Disorders.

Thus, first, and most simply, when a kind is postulated, the range of possibilities for intentional action by the members of that kind is suddenly increased. When one is classified as a member of that kind, and is aware that one is classified that way, a range of novel descriptions is made available which can play a role in rational deliberation and formation of intentions to act, not simply by providing descriptions of actions themselves, but by characterising the beliefs, desires and other psychological states which make such actions attractive to the agent (Hacking 1999, p. 104). The awareness and understanding of an individual member of a kind that she is a member of that kind can change the effects which that individual can have by broadening the range of actions which she can undertake, as well as changing the way in which she rationally explains her behaviour. Furthermore, from a third person perspective, the additional descriptions generated by the postulation of a kind can result in two kinds of change: it can alter the way in which the behaviour of individuals of that kind is described and explained; and it can lead to novel intentional actions on the part of people who are not members of  $K$ , either towards members of  $K$ , towards non-members, or towards the environment in which the members of  $K$  find themselves.

But, even given the close association between the descriptions in the stereotype  $S_k$  and the cluster of properties  $P_k$  instantiated by members of the kind, these features alone are not sufficient to make the feedback generated by members of these interactive kinds a special case since, as was noted above, the members of a kind can exhibit the effects of the properties they instantiate in the form of unintentional voluntary behaviour. The properties do not need to be described in order to be causally efficacious. The kind classification and the individual's awareness of it seem to be irrelevant to the physical effects which a kind member's behaviour has, and thus upon her impact on the environment. Because the same causal outcome can occur in the absence of the classification, the intention to act which has been facilitated by the classification seems redundant.

However, this complaint misses out some variants of possible action which are brought about specifically by the subject's ability to form an intention to act, or to describe themselves and their actions in a fine-grained way, or by the ability of those around them to describe and explain their actions in novel terms. I will consider two



features of human interactive kinds which distinguish the way in which they produce feedback from that of other interactive kinds.

### 3.2 Mistakes

First, as noted in Sect. 2.1, membership of a kind  $K$  does not require that an individual member of  $K$  instantiate every property in  $P_k$ . Even if we assume the ideal scenario where there is a perfect match between descriptions in  $S_k$  and properties in  $P_k$ , there may be descriptions in  $S_k$  which an individual member does not satisfy.

Second, we should note that the attribution of properties, including the attribution of psychological states, is not infallible. Although we usually presume that a subject has first person authority about their own intentional states, that does not entail first person infallibility or incorrigibility in all cases. While we are more likely to be incorrect about the intentional states which someone has when trying to attribute them in the third person, an individual might sometimes be incorrect about her own intentional states too. There are some sensation states for which we might more readily claim infallibility, such as ‘I am in pain’ or ‘There appears to be something red in front of me’, as well as simple beliefs and desires ‘I want a drink of water’, ‘I believe that water is wet’ and so on. However, the intentional states associated with interactive kinds are often precisely those states which we might think of as being more difficult to self-attribute, such as those associated with psychiatric disorders including delusional states, psychoses, addictive behaviours or other pathologies.<sup>34</sup> Thus, we might expect that self-attribution is less trustworthy in such cases.<sup>35</sup> Moreover, this expectation may sometimes lead to further cases of mistaken attribution: it may be assumed that the members of a kind have diminished first-person authority even if this is not the case (Yergeau 2010, 2013).

I will argue that there is a key difference between the consequences of mistakes in property attribution to human kinds and non-human ones. Mistakes in the latter are unremarkable and widespread: we might think that a surface is brown rather than green, or that the mass of a sample is 5 kg rather than 4 kg, but these are the sorts of mistake which can usually be corrected easily by later investigation. However, in the case of interactive human kinds, such misattributions are harder to rectify and they

<sup>34</sup> The phenomenon of anosognosia, or lack of insight, into a neurological or psychiatric condition is widespread, affecting individuals with bipolar disorder, schizophrenia, alcoholism (Walvoort et al. 2016), anorexia nervosa (Lasègue 1873; Pryor et al. 1995; Vandereycken 2006), ASD, as well as in patients who have suffered strokes (Stone et al. 1993; Ramachandran and Rogers-Ramachandran 1996) or brain injury (Lebrun 1987; Prigatano and Schacter 1991; Vuilleumier 2004). This differs from denial (which also may play a role in the misattribution of properties associated with a kind) in that anosognosia is associated with neurological or sensory deficits. Some discussion in Pryor et al. (1995) suggests that denial of anorexia is intentional, rather than mistaken, and thus belongs in 3.3 (1995, p. 301). A different kind of mistake in self-attribution arises from difficulties with action monitoring by people with Obsessive Compulsive Disorder (Gehring et al. 2000).

<sup>35</sup> The extent to which delusional individuals or people with schizophrenia should be regarded as having a deficit in first person authority is a matter of some discussion. (See Mundt 1996; Sass 1994; Ikuta 2004.) To complicate matters, Kumazaki (2015) argues that persecutory delusions are characterised by their sufferers denying the first-person authority of others. Thus, the belief that oneself, or someone else, is a member of such a kind may influence how one regards the kind member’s *evaluation* of first person ascriptions (even when the kind member is oneself).

expand both the range of actions which an individual is able to perform and the ways in which her actions are explained.

The difficulty arises from the fact that membership of  $K$  does not require that an individual member of  $K$  instantiate every property in  $P_K$ , so we might misattribute properties to individual members of  $K$  on the basis that having such properties is stereotypical of members of their kind  $K$ . Thus we might interpret an individual member of  $K$  as acting as if she has property  $Q$  when in fact she does not have  $Q$ , and this will inform our understanding of what the possession of property  $Q$  (alongside other properties) causes members of  $K$  to do. Furthermore, given the complex nature of the property clusters associated with the kinds under discussion, it is plausible to think that similar errors could occur in the first person case too. An individual who is aware that she is a member of  $K$  believes that some or all of the descriptions in the stereotype  $S_K$  apply to her, even though some of these descriptions do not pick out properties which she has. She might plan and rationalise actions as if she is  $Q$ , when she is not. For instance, someone with ASD might consider herself to be resistant to changes in her daily routine because this is a characteristic commonly associated with autism even though she does not instantiate this property. This misapprehension might affect her deliberation, belief formation and the intentional action which results, even though the characteristic is not present because she is not resistant to change in her daily routine in the first place.<sup>36</sup> Thus, she may seek to maintain her routine, not because changing it would actually cause her anxiety, discomfort or distress, but because she believes that changing it would cause her anxiety, discomfort or distress.

This phenomenon leads to there being a crucial difference between a mistake in property attribution in the case of human interactive kinds and mistakes in other cases. A table which is mistakenly attributed the property of being 5 kg rather than being 4 kg will not cause anything *qua* a 5 kg table since the requisite causal property is not instantiated; it will continue to have effects *qua* 4 kg table, and the mistake will (most probably) be discovered easily. However, consider three individuals, Alice, Beth and Clara, all of whom are members of human kind  $K$  and are aware of this: Alice thinks that she instantiates a property  $Q$  which is stereotypical of  $K$  although she does not; Beth does not instantiate  $Q$ , but also does not believe that she instantiates  $Q$ ; while Clara both instantiates  $Q$  and believes that she does so. Even though Alice does not instantiate  $Q$ , she incorporates her *being Q* into her rational deliberations: *the belief that she is Q* affects her sense of who she is and her choice of actions; her having the property of *falsely believing that she is Q* has effects. Moreover, in decisions which

<sup>36</sup> As noted in 2.1, the absence of a particular characteristic does not imply that the individual in question does not have ASD. The diagnostic criteria of DSM-5 require an individual to exhibit deficits in two broad categories: social and communication deficits (of which the individual must satisfy all three criteria) and restrictive, repetitive patterns of behaviour (in which the individual must satisfy only two of four criteria); but *within* these subcategories, the diagnostic criteria are not specific about the nature of these impairments, nor are the examples given individually necessary for having autism. The DSM-5 criteria are slightly more restrictive than the polythetic criteria of DSM-4 which required that individuals satisfy six criteria from the ‘Triad of Impairments’ (from Wing and Gould 1979), including at least two forms of social impairment and one each from lists of communication impairments and repetitive behaviour impairments. Both these criteria permit different kind members to instantiate different features from the cluster identified with ASD. (For differences in the implications of the diagnostic categorisation of DSM-4 and DSM-5, see McPartland et al. 2012.)

rest upon being  $Q$ , Alice will act differently from Beth who neither possesses  $Q$  nor believes that she does, because *believing that she is  $Q$*  plays no part in Beth's mental life. However, we should not expect Alice to act in precisely the same way as Clara, the only one of the three who does instantiate  $Q$ , since Clara has a property which Alice lacks. It is plausible to think that (at least) sometimes Alice's actions will neither be those of someone with  $Q$  (who is aware of it), nor someone who lacks  $Q$  (and is aware of it); what she does will be anomalous with respect to the stereotypical properties associated with  $K$ . However, her actions will be judged as if they are not anomalous and may result in the classification of what makes something  $K$  being changed as a result. If behaviour like Alice's becomes widespread in similarly mistaken members of the kind  $K$ , this will be incorporated into the stereotype of  $K$  and the properties associated with  $K$ s will be adjusted accordingly. However, the additional property in question need not be *being mistaken about being  $Q$*  (although it might be<sup>37</sup>), but another which is postulated to give a positive reason for the individuals' actions.

### 3.3 Fakes

In addition to acting on the misapprehension of having certain properties, members of a kind  $K$  may deliberately act as if they do not instantiate certain properties which they do in fact possess, or else act as if they do possess properties which they do not instantiate. In some cases, this may be a matter of localised deceit in the sense that the deceit only affects a subset of actions: in social situations, such as when they are at school, girls with ASD are now known to mask their symptoms by intentionally engaging in activities which obscure some characteristic properties of their ASD.<sup>38</sup> Furthermore, some people with ASD are now explicitly taught how to camouflage ASD traits and to engage in behaviour which mimics neurotypical social interaction, although there is no expectation that the resulting actions will perfectly mimic those of someone who does not have ASD since they involve using instructions and higher level reasoning to replace innate abilities to socially interact (NAS 2010).<sup>39</sup> Similarly, people who become aware that they are members of a persecuted minority group in a society will behave differently from those who are members of that group and not aware, perhaps concealing their identity as a matter of self-preservation, or hiding some traits which would reveal themselves as members of that group. In both cases, the individuals' intentional actions to produce behaviour which deceives others can result in their having effects which are at odds with the properties of the respective kinds under which they fall; such anomalous effects can in turn result in feedback which

<sup>37</sup> Some of the property attribution mistakes noted as examples in footnote 34 would now count as being part of the stereotype of the kinds in question. But this does not remove the problem of mistaken attribution as there may be difficulties concerning mistakes about other properties in  $P_K$  and also whether a stereotypical property is instantiated by a specific individual kind member or not.

<sup>38</sup> Although masking is not exclusive to girls, there is now good empirical evidence to suggest that it is far more prevalent among girls with ASD compared to boys. Several theses have been advanced about why this is the case. See Attwood et al (2006), Attwood (2007), Lai et al. (2011, 2012, 2015, 2017), Head et al. (2014), Hiller et al. (2014), Lehnhardt et al. (2016), Parish-Morris et al. (2017).

<sup>39</sup> For instance, engaging in eye contact is taught as a set of instructions: look at the forehead or nose of the person you are talking to, count to five, look away, count to five, return to look at forehead again, etc.

changes the classification. For instance, in the case of ASD, the possibility of symptoms being masked by an individual has been added to the diagnostic criteria in the most recent edition of DSM-5 (APA 2013), although key differences are acknowledged between the masking behaviour and successful neurotypical social interaction (Hull et al. 2017).<sup>40</sup> An attempt to pretend that one has a property which one does not have does not always, or even usually, have exactly the same outcome as intentional action which is not based upon faked beliefs, desires or abilities.

Thus, more generally, we would not always expect the actions of K-individuals who are attempting act as if they possess property R (when they do not) to be the same as someone who does possess R, nor to be the same as an individual of their kind who does not possess R and is not faking. There is an anomaly between the properties of K ( $P_k$ ) and the actions which the faking individuals belonging to K produce, and such actions can affect the individuals' environment and those around them, eventually producing feedback which results in additions or adjustments to the stereotype of K. Also, as in the previous case of mistaken property attribution in Sect. 3.2, there is no causal mystery to the anomalous nature of the action with respect to the properties of K: the properties which any particular individual has are not exhausted by the properties of any one human kind K to which he or she belongs and these other properties may interact with whichever properties of K she instantiates. Thus, while the action is anomalous with respect to  $P_k$ , the kind properties of K, it is not anomalous with respect to the all properties of the individual in question.

This faking is not only a feature of human kinds, but can also be seen in other social kinds. For example, a common, and arguably politically beneficial, course of action for the government of a country which is classified as a *rogue state* is to start behaving as a rogue state is stereotypically supposed to do, whether or not that country actually has the properties of being a rogue state when it is classified.<sup>41</sup> In this case, an entity which has been classified as a certain kind but does not fit the stereotype for that kind acts as if it does fit that stereotype, with the likely result that the actions of the government and institutions of that state will not fit with either those of a rogue state or those of a non-rogue state. However, since the state has been classified as 'rogue', the effects of the state and its institutions might be taken to be paradigmatic of a rogue state, and so the classification might be altered to fit the behaviour of the wrongly labelled state.

The phenomena of mistaken property attribution and faked property possession are two ways in which individuals of human interactive kinds can intentionally act in novel ways in virtue of their classification which go beyond the simple addition to the range of intentional action which the novel descriptions of the stereotype facilitate. Furthermore, the phenomenon of mistaken property attribution can occur in the third person too because explanations available to third parties are expanded by the availability of novel descriptions under which behaviour can be categorised, but one can

---

<sup>40</sup> The recognition of masking and its subsequent inclusion as a diagnostic criterion led to ASD no longer being regarded as a condition which primarily affects men and boys, as it had been not long previously (see Baron-Cohen 2002, p. 251 which puts the male–female ratio at 10–1), and has resulted in the reclassification of women and girls diagnosed with other conditions as having ASD.

<sup>41</sup> I am grateful to Rebecca Richards for this example. Stephen Mumford also suggested that the kind *criminal* undergoes similar changes when applied to people who have not committed crimes.

make mistakes about which properties from a cluster an individual instantiates. Moreover, non-human interactive kinds do not suffer from analogous problems, since the anomalous effects associated with faked or mistaken properties result from a *mismatch* between the properties which an individual of a kind thinks that she has, or wishes to present herself as having, and those which she actually possesses. (Or, in the third party case, from the mistaken attribution of properties to explain or predict the actions of individuals of a kind.) In the case of human interactive kinds, such a mismatch can produce causal outcomes which are anomalous with respect to the cluster of properties associated with the kind. However, in the case of the pedigree dog or the tuberculosis bacillus, a misattribution of a property Q to an individual which does not possess Q will not result in the individual's behaving in a different way from an individual which simply does not possess Q.

### 3.4 Is there really a third way?

So far, there is an assumption embedded in my argument which should be made explicit. My explanation of the behaviour of interactive human kinds has assumed that an individual member of a kind K can lack (at least) one of the set of properties  $P_k$  associated with the kind—let us call this property R—and yet have the resources to attempt act as if she does possess R, either because she mistakenly thinks that she has R, or because she knows she does not have R but would like others to think that she does. I am presuming that these scenarios can make the behaviour of the individual different from those in which she has R *and* those in which she lacks R but is neither mistaken about it nor trying to fake the presence of R. In other words, I am presupposing that there is a 'third way' for action associated with a human kind to take, and this underlies my claim that classification as a human kind facilitates a greater range of actions (and thereby causal outcomes) than non-human ones. If I cannot sustain this claim, human interactive kinds would behave no differently from non-human interactive kinds.

Is this assumption plausible? I have two strategies to support such a presupposition, one conceptual and one empirical, although in this paper there is not space to provide a rigorous defence. First, I think that there is good reason to think that the assumption is plausible because our understanding of human action is rich enough to encompass cases of lying, delusional behaviour, or otherwise behaving as if one has a property (intentional or non-intentional) which one does not in fact possess, and to do so in such a way that what the agent does is different from when she acts sincerely. So I am not postulating these features simply to support my argument in an ad hoc way. We do not expect individuals who are lying, or who are acting, always to do so in such a way that they produce behaviour which is indistinguishable from the actions which would be produced by someone who genuinely instantiates the properties which they are trying to emulate. They may sometimes do so, but the fact that there are exceptions is enough to sustain my argument. Even good liars, such as poker players, have their 'tells'.<sup>42</sup> Furthermore, were an account of psychological explanation and causa-

<sup>42</sup> The distinctions between the behaviour of someone who is faking and someone who is not may be hard to spot, but that does not undermine the claim that there is a difference. See Ekman (1993), Etcoff et al. (2000), Ekman and O'Sullivan (2006).

tion to fail to discriminate between deceptive and sincere actions (and the different psychological states which produce them), we should regard that as a weakness of the theory. Although philosophers disagree about the ontological mechanisms which ground human action, and so the reason why my assumption holds is up for grabs, there are good reasons to maintain it independently of giving an account of interactive human kinds.

Furthermore, if there are differences in outcomes when the subject is intentionally trying to deceive compared to her acting sincerely, we would expect more marked differences in outcomes when an individual is mistaken about the beliefs which she has, or has mistaken beliefs about her own properties, in virtue of belonging to a kind. In the case of faking, there is an intentional attempt to conceal the deficit, but in the mistaken case there is none. Intuitively, at least, one would think that actions based on mistaken attributions are even less likely to emulate actions performed if the property were actually present than if the individual were attempting to deceive.

Fortunately, we do not have to rely entirely upon folk-psychological intuitions in these cases. In addition to these general arguments that human agents are fairly inefficient at emulating the presence of beliefs which they do not possess or properties which they do not have, there is some empirical evidence that the results of such actions fall short of what would occur if the property were present, or generate additional causal outcomes which do not occur when the behaviour is sincere. General work on this phenomenon has been done by Paul Ekman by analysing the ways in which the behaviour of deceptive individuals differs from that of non-deceivers, both linguistically and non-linguistically. (I say ‘behaviour’ here quite deliberately since the deceptive individuals do not intend to produce such an outcome.) Furthermore, there is more specific evidence from research into the behaviour of subjects belonging to various human interactive kinds. For instance, Hull et al. (2017) detail the consequences of masking for individuals with ASD, including exhaustion, increased anxiety, and ‘a need to withdraw from social interaction to re-set’ (2017, p. 2520), while Cage et al. (2013) notes the reduced ability to imitate social interactions and Bargiela et al. (2016) reports that the effort required for such masking goes beyond that which is required in neurotypical behaviour management and manipulation. Furthermore, embedded in the diagnostic criteria which describe such phenomena is the claim that most (if not all) masking and camouflaging behaviour lacks the depth and complexity of neurotypical social interaction and that instances of this can be spotted if one is trained in what to look for. Similarly, individuals with anorexia nervosa who deny that their weight loss is unhealthy show different responses to treatment to those who admit it (Pryor et al. 1995), with the deniers engaging with restrictive eating for a shorter period of time and being less prone to bulimia.<sup>43</sup> However, the deniers in such cases do not behave in the same way as those who lack the properties associated with having the eating disorder. Other empirical examples in which the outcomes brought about by faking are neither the usual lawlike outcomes of the property which is being faked

---

<sup>43</sup> It is obviously difficult to establish a causal link in such a case between the denial and the different actions, although most other factors in the individuals’ backgrounds were the same between deniers and admitters. Also, as noted in footnote 34, the denial might be thought to be accidental rather than deliberate, although Pryor et al. tend towards the latter interpretation. Either interpretation would be relevant to the discussion at hand.

nor paradigmatic of the property's absence have been reported in alcoholism, bipolar disorder, and schizophrenia (Rogers and Bender 2018).

There is more to be said in defence of the assumption that, because human agents can fake and make mistakes, classification as a human kind facilitates a wider range of novel effects relative to the kind than that available to non-human kinds. However, I hope to have at least made this assumption plausible pending further research into the matter.

### 3.5 Non-ideal scenarios

It is now time to justify the comment, made in Sect. 2.3, that for the purposes of my argument, there is no harm in presupposing the most simple relationship between semantics and ontology in which there is a one-to-one correspondence between the descriptions of  $S_k$  and the properties of  $P_k$ .

This assumption is benign because a mismatched relationship between  $S_k$  and  $P_k$  (in which the descriptions of  $S_k$  do not all pick out  $P_k$  properties or  $P_k$  properties exist for which there is no description in  $S_k$ ) would make my case stronger. In such cases, it would be more likely for the classification of individuals as members of interactive kind  $K$  to produce novel effects which brought about changes in the classification and thereby altered the set of descriptions contained in  $S_k$ . The reason for this may now be obvious to the reader, since the features which bring about the differences between human and non-human interactive kinds are based upon mismatches between the properties which an individual of  $K$  or those around her believe that she has (the properties which the descriptions in  $S_k$  pick out), or the properties which she wants to present herself as having, and the subset of  $P_k$  properties which that same  $K$ -individual instantiates. The anomalies in physical effects which prompt the classificatory feedback loop distinctive to human interactive kinds arise because the properties of  $S_k$  with which an individual describes herself (or with which others might describe her) are not always those which she actually instantiates; and when such anomalies occur, we may alter the descriptions which we think  $S_k$  contains accordingly. The more descriptions which are mistakenly included in  $S_k$ , the more likely that mistaken or faked behaviour will occur because there is a greater mismatch between the stereotype and the properties which members of the kind customarily have, and while the stereotype plays the primary role in the explanation and descriptions of action, the properties which are actually instantiated (both from  $P_k$  and the other properties which the individual possesses) play the causal role in producing the action. If what someone believes about being  $K$  is very different from what being  $K$  is actually like, then greater anomalies of effects result. Moreover, if this mismatch is a general problem across individuals of the kind and not just a case of individual kind members failing to instantiate some of the properties in  $P_k$ , then the anomalies of behaviour in relation to the stereotype will be greater and the likelihood that the classification will be altered will increase.

## 4 Some complications

In the case of human interactive kinds, many or all of the descriptions in the stereotype are aiming to pick out intentional states. However, one might think that the ontological account within which I have been working takes an overly simplistic view of the attribution of intentional properties, the nature of such properties, their relationship to the individuals which instantiate them, and their role in causation and explanation.

The arguments of Sect. 3 only require a naturalistic, atomistic and realist account of intentional states or properties—they are on a par with physical properties from an ontological point of view—and thus one need not think that there are any constitutive differences between intentional properties and the rest in order to draw a distinction between human and non-human interactive kinds. However, if there are such differences, then these may further differentiate human interactive kinds from other kinds, both with respect to their ontological status in relation to other kinds and to the epistemic and methodological implications of theorizing and experimenting in terms of such kinds.<sup>44</sup>

A second feature which potentially distinguishes intentional properties from others concerns the way in which they relate to each other, which may also serve to alter their relations to non-intentional properties and to change the nature of feedback loop which is generated when kind members are aware of their classification. For instance, one might think that intentional properties are not related causally but according to rational connections, where the concept of rationality involved may either be quite rigidly aligned to logical inference between belief contents or very broadly characterised to cover many different ways in which thoughts are linked. If such differences are irreducible, they may not simply distinguish human *interactive* kinds but be characteristic of some, or all, human kinds.

However, if one takes a fairly broad conception of rationality, there are some interesting consequences for interactive kinds because there are cases in which the classification of an individual as being of a kind involves the claim that his beliefs, desires and other intentional properties do not relate to each other in the ‘usual’ way. For instance, it is an integral part of the classification of several psychiatric disorders that the way in which the members of a kind think—in the sense of ‘the way in which their intentional states relate to each other’—might be different from that of the general population *in virtue of their being a member of that kind*. Conditions such as bipolar disorder, schizophrenia, psychosis, Alzheimer’s Disease, and more localised delusions (such as the Capgras, Cotard or Fregoli delusions) are such that we do not expect someone who has them to manifest the same ‘rational’ connections between their intentional states as the general population who are not members of the kind. Arguably, subjects with autistic spectrum disorders are also considered to have different patterns of psychological processing from those who could be classified (in comparison) as neurotypical. These differences could be global, or local: the condi-

---

<sup>44</sup> For example, intentional properties might differ from physical ones due to the former being irreducibly teleological (Millikan 1984, 1993; Papineau 1993; Jacob 1997), by involving a phenomenal aspect (there being something that it is like to possess them) (Horgan and Tienson 2002; Farkas 2008), or by having different spatial properties from non-intentional ones (McGinn 1995, 2004).



tion may be such that this difference in inferential links does not occur all the time in individual members of the kind, or in all areas of their thought.

The fact that being a member of a specific kind *K* involves some or all of one's properties being related to each other in a characteristically *K*-like way has enormous implications for the way in which being classified as a *K* affects how an individual may act and the way in which others act towards him or explain his actions.<sup>45</sup> Being classified as a specific kind with this feature alters how individuals are expected to think and can influence the testimonial authority which those individuals are accorded in the sense that their own rationalisations and belief reports may be considered to be 'different' from those of non-kind members. Furthermore, in the cases in which the difference in thought processes are either not constant or not global, the individual kind-member or others around her may presuppose that her intentional properties are related to each other in ways in which they are not at that time, leading to actions being misdescribed and the resulting anomalies being mistakenly attributed to the individual's being a member of the kind *K*. Thus anomalies of action may be construed as being paradigmatic of the kind in question, rather than as being the result of the mistaken assumption that the individual is thinking in a certain way, and thereby lead to the mistaken adjustment of the descriptions included in the stereotype which characterises the kind. Classificatory feedback will occur.

Finally, either of these two distinguishing features of intentional properties and the way they relate to each other may make the properties of interactive kinds relate to each other holistically, rather than atomistically as I have hitherto presumed. Such a difference will affect the way in which different individuals of a kind are able to act according to which properties of the HPC they have or lack. The presence of an additional kind property *S* in an individual may globally affect all the other properties associated with the kind which that individual instantiates, and consequently make a greater contribution to changing the range of intentional action than would be expected, compared to the individuals members of *K* who lack *S*. Such holism would better explain the examples of Sects. 3.2 and 3.3, since it would be harder for the *K*-individual pretending to possess *S* (for instance) to correctly anticipate the effect which possession of *S* would actually have. However, I do not think that the examples of Sects. 3.2 and 3.3 rely upon holism in order to work.

In this section, I have very briefly summarised the ways in which dualism about intentional properties or about the relations between them could further serve to differentiate human interactive kinds from non-human ones. Some of these features might be exploited to explain why the cases discussed in Sect. 3 make human interactive kinds different to the rest, although I do not think that they are required to do so. Having noted these additional potential differences, I will leave them as areas for further research and move on to consider some remaining issues.

---

<sup>45</sup> For instance, Yergeau (2013) highlights how the actions of individuals with autism are interpreted in specific ways in virtue of their having autism.

## 5 Concluding comments

I have argued that classification can prompt the individuals who are classified (or those around them) to produce novel effects, instigating a feedback loop whereby these effects lead to a change in their own properties and to a change the classification of the kind in question. Stereotypes are adjusted in order to capture the novel or unexpected actions of individuals of the kind or the changes in properties which those actions bring about. Thus being classified as a kind *K* can change what *being K* consists in. Furthermore, these effects are distinctive in human interactive kinds due to the anomalies which appear when a kind-member (or those around him) is mistaken about the properties associated with the kind which he possesses or he wishes to fake the presence or absence of those properties. This range of potential effects does not (and cannot) occur in the case of non-human interactive kinds such as pedigree dogs or biological pathogens, since they lack the faculties required for intentional action and for beliefs about their own kind-membership. Thus, there is a distinction between human interactive kinds and non-human interactive kinds. These phenomena will occur even if one does not accept that intentional properties or the connections between them are essentially distinct from broadly-speaking physical ones, as long as one accepts that humans (for whatever reason) can engage in deceptive, delusional or mistaken behaviour, and that they do not always do so successfully.

If I am right, we should reinstate the distinction between human interactive kinds and those whose members cannot engage in intentional action. The implications of this distinction are wide-ranging, since we would have to take account of it in general theorising, experimental design and statistical analysis in a broad range of medical and social sciences in which human subjects are studied. Furthermore, if it turns out that animals can engage in the requisite discrimination and deliberation to engage in intentional action which changes kinds in this way—albeit in a non-linguistic form—then we should drop the term ‘human’ and choose something more fitting instead. However, although it is plausible to think that some species may be able to discriminate in order to engage in intentional action, it is less clear how many species would count as being able to engage in deception or be genuinely mistaken about the properties they have such that they act upon the basis of this mistaken information. Only those species which do will qualify as belonging to ‘human’ interactive kinds.

Finally, one might wonder about the extent to which feedback can alter classification, especially if one is drawn to essentialism: is the change in classification which feedback produces due to an epistemic problem which can be rectified once the stereotype contains certain descriptions which pick out essential properties of the kind? Or could feedback generate opened-ended change in what counts as the kind in question? If there are any essences at all, I suspect that both cases could occur. In some cases, the feedback loop will halt upon the discovery of an essential property of the kind in question. Were we to discover an essential physical basis for a disease such as bipolar disorder (say) as we have for diseases such as Huntington’s disease or AIDS, there would be a fact of the matter about whether or not an individual had bipolar disorder or not, and the kind of feedback discussed in this paper would not be able to change the classification in a significant way. (At most, one might imagine that the genetic, environmental or other physical basis of the disorder might itself mutate in response

to classification, resulting in a non-human interactive kind.) Whereas, in other cases for which there is no essential basis, it is plausible to think that the feedback might never stop and that what counts as the kind, and whichever individuals best satisfy the descriptions of  $S_k$ , will continue to develop in light of the kind-members' actions. However, given that some of the changes in human kind classification are brought about through mistakes or deception, we might also wonder what (if anything, in the absence of an essence) determines whether particular changes in classification brought about by classificatory feedback are legitimate.

One might speculate however about whether there could be cases in which we have both an essential basis *and* classificatory feedback continues to occur, subtly changing what counts as the kind despite the existence of a core set of essential properties.<sup>46</sup> For instance, the sufferers of Pathological Demand Avoidance are good candidates for never being captured by a stable cluster of properties, even if their condition were grounded by an essential biological basis. The stereotype might continue to evolve with respect to the non-essential properties of the condition, even though it would remain 'centred' around certain essential properties. This question about the extent of interactivity merits further investigation, but I will postpone further discussion for another occasion.<sup>47</sup>

**Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

## References

- American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders (DSM-4)* (4th ed.). Washington, D.C.: APA.
- American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders (DSM-5)* (5th ed.). Washington, D.C.: APA.
- Anscombe, G. E. M. (1957). *Intention*. Cambridge, MA: Harvard University Press.
- Anscombe, G. E. M. (1979). Under a description. *Noûs*, 13, 219–233.
- Attwood, Tony. (2007). *The complete guide to Asperger's Syndrome*. London: Jessica Kingsley Publishers.
- Attwood, T., Grandin, T., Bolick, T., Faherty, C., Iland, L., Myers, J. M., et al. (2006). *Asperger's and girls*. Texas: Future Horizons.
- Bargiela, S., Steward, R., & Mandy, W. (2016). The experiences of late-diagnosed women with autism spectrum conditions: An investigation of the female autism phenotype. *Journal of Autism and Developmental Disorders*, 46, 3281–3294.
- Baron-Cohen, S. (2002). The extreme male brain theory of autism. *Trends in Cognitive Science*, 6, 248–254.
- Beebe, H., & Sabbarton-Leary, N. (Eds.). (2010). *The semantics and metaphysics of natural kinds*. London: Routledge.
- Bird, A. (2015). The metaphysics of natural kinds. *Synthese*, 195, 1–30.

<sup>46</sup> Tsou (2007) has noted a problem with Hacking's own distinction between interactive and indifferent kinds, in that the latter are defined in contrast to the former. Thus, it would be inconsistent to say that a kind is both interactive and indifferent (although Hacking occasionally does say this in discussing how interactive kinds might turn out to have a biological basis, for instance (1999, pp. 108–24)). However, one does not need to make the distinction in this way and could talk about kinds being interactive and yet being centred on a core set of essential properties.

<sup>47</sup> I would like to thank audiences in Madrid, Manchester and Keele for useful comments.

- Bird, A., & Hawley, K. (2011). What are natural kinds? *Philosophical Perspectives: Metaphysics*, 25(1), 205–221.
- Bogen, J. (1988). Comments. *Noûs*, 22, 65–66.
- Boyd, R. (1989). What realism implies and what it does not. *Dialectica*, 43, 5–29.
- Boyd, R. (1991). Realism, anti-foundationalism, and the enthusiasm for natural kinds. *Philosophical Studies*, 61, 127–148.
- Boyd, R. (1999). Homeostasis, species, and higher taxa. In R. A. Wilson (ed.), pp. 141–185.
- Boyd, R. (2010). Realism, natural kinds and philosophical methods. In Beebe and Sabbarton-Leary (eds.), pp. 212–234.
- Cage, E., Pellicano, E., Shah, P., & Bird, G. (2013). Reputation management: Evidence for ability but reduced propensity in autism. *Autism Research*, 6, 433–442.
- Chalmers, D. (Ed.). (2002). *Philosophy of mind: Classical and contemporary readings*. Oxford: Oxford University Press.
- Cooper, R. (2004). Why Hacking is wrong about human kinds. *British Journal for the Philosophy of Science*, 55, 73–85.
- Davidson, D. (1970). Mental events. In Davidson (1980), pp. 207–225.
- Davidson, D. (1980). *Essays on actions and events* (2nd ed., 2001). Oxford: Clarendon Press.
- Davidson, D. (1985). Reply to Quine on events. In Lepore, E., & McLaughlin, B. P. (Eds.). Reprinted in Davidson (2001) *Essays on Actions and Events: Philosophical Essays Volume 1* (2nd ed.), pp. 305–312.
- Dennett, D. (1987). *The intentional stance*. Cambridge, MA: MIT Press.
- Douglas, M. (1986). *How institutions think*. Syracuse: Syracuse University Press.
- Drabek, M. L. (2010). Interactive classification and practice in the social sciences: Expanding Ian Hacking's treatment of interactive kinds. *Poroi*, 6, 62–80.
- Ekman, P. (1993). Facial expression and emotion. *American Psychologist*, 48, 384–392.
- Ekman, P., & O'Sullivan, M. (2006). From flawed self-assessment to blatant whoppers: The utility of voluntary and involuntary behavior in detecting deception. *Behavioral Sciences & The Law*, 24, 673–686.
- Ereshefsky, M. (2004). Bridging the gap between human kinds and biological kinds. *Philosophy of Science*, 71, 912–921.
- Ereshefsky, M., & Reydon, T. A. C. (2015). Scientific Kinds. *Philosophical Studies*, 172, 969–986.
- Etcoff, N. L., Ekman, P., Mage, J. J., & Frank, M. G. (2000). Lie detection and language comprehension. *Nature*, 405, 139.
- Farkas, K. (2008). Phenomenal intentionality without compromise. *The Monist*, 91, 273–293.
- Fine, K. (1994). Essence and modality. *Philosophical Perspectives*, 8, 1–16.
- Gerhing, W. J., Himle, J., & Nisenson, L. G. (2000). Action-monitoring dysfunction in obsessive-compulsive disorder. *Psychological Science*, 11, 1–6.
- Griffiths, P. (1997). *What Emotions Really Are*. Chicago: Chicago University Press.
- Hacking, I. (1986). Making up people. In T. Heller, M. Sosna & D. Wellberry (Eds.), pp. 222–236.
- Hacking, I. (1988). The sociology of knowledge about child abuse. *Noûs*, 22, 53–63.
- Hacking, I. (1995). The looping effects of human kinds. In D. Sperber, D. Premack, & A. Premack (Eds.), pp. 351–94.
- Hacking, I. (1999). *The social construction of what?*. Cambridge, MA: Harvard University Press.
- Hacking, I. (2002). *Historical ontology*. Cambridge, MA: Harvard University Press.
- Hacking, I. (2006). *Types De Gens, Des Cibles Mouvantes*. [http://www.college-de-france.fr/media/ian-hacking/UPL4141595277092170127\\_6\\_\\_\\_Types\\_de\\_gens\\_des\\_cibles\\_mouvantes.pdf](http://www.college-de-france.fr/media/ian-hacking/UPL4141595277092170127_6___Types_de_gens_des_cibles_mouvantes.pdf). Accessed 23 Sept 2017.
- Hacking, I. (2007a). Kinds of people: Moving targets. *Proceedings of the British Academy*, 151, 285–318.
- Hacking, I. (2007b). Natural kinds: Rosy dawn, scholastic twilight. *Royal Institute of Philosophy Supplement*, 61, 203–239.
- Hägqvist, S. (2005). Kinds, projectibility, and explanation. *Croatian Journal of Philosophy*, 5, 71–87.
- Hägqvist, S., & Wikforss, Å. (2017). Natural kinds and natural kind terms: Myth and reality. *British Journal for the Philosophy of Science*. <https://doi.org/10.1093/bjps/axw041>.
- Haslanger, S. (1995). Ontology and social construction. *Philosophical Topics*, 23, 95–125.
- Haslanger, S., & Saul, J. (2006). Philosophical analysis and social kinds. *Proceedings of the Aristotelian Society*, 106, 89–118.
- Haukioja, J. (2015). On deriving essentialism from the theory of reference. *Philosophical Studies*, 172, 2141–2151.
- Hauswald, R. (2016). The ontology of interactive kinds. *Journal of Social Ontology*, 2, 203–221.

- Head, A. M., McGillivray, J. A., & Stokes, M. A. (2014). Gender differences in emotionality and sociability in children with autism spectrum disorders. *Molecular Autism*, 5, 19.
- Heller, T., Sosna, M., & Wellberry, D. (Eds.). (1986). *Reconstructing individualism*. Stanford, CA: Stanford University Press.
- Hiller, R. M., Young, R. L., & Weber, N. (2014). Sex differences in autism spectrum disorder based on DSM-5 criteria: Evidence from clinician and teacher reporting. *Journal of Abnormal Child Psychology*, 42, 1381–1393.
- Horgan, T. E. & Tienson, J. L. (2002). *The intentionality of phenomenology and the phenomenology of intentionality*. In Chalmers (Ed.) (520–533).
- Hull, L., Petrides, K. V., Allison, C., Smith, P., Baron-Cohen, S., Lai, M.-C., et al. (2017). Putting on my best normal: Social camouflaging in adults with autism spectrum conditions. *Journal of Autism and Developmental Disorders*. <https://doi.org/10.1007/s10803-017-3166-5>.
- Ikuta, T. (2004). Tougou-shitchou-shou ni okeru mousou no kouzou [On the structure of delusion in schizophrenia]. *Japanese Journal of Psychopathology*, 25, 111–118.
- Jacob, P. (1997). *What minds can do*. Cambridge: Cambridge University Press.
- Khalidi, M. A. (2010). Interactive kinds. *The British Journal for the Philosophy of Science*, 61, 335–360.
- Khalidi, M. A. (2015). *Natural categories and human kinds*. Cambridge: Cambridge University Press.
- Kistler, M. (2016). Espèces naturelles, profil causal et constitution multiple. *Revue de la Société pour la Philosophie des Sciences*, 3, 17–30.
- Kripke, S. (1980). *Naming and Necessity*. Cambridge, MA: Harvard University Press.
- Kumazaki, T. (2015). Persecutory delusions and first person authority. *Theory and Psychology*, 25, 80–95.
- Kuorikoski, J., & Pöyhönen, S. (2012). Looping kinds and social mechanisms. *Sociological Theory*, 30, 187–205.
- Lai, M. C., Lombardo, M. V., Auyeung, B., et al. (2015). Sex/gender differences and autism: Setting the scene for future research. *Journal of the American Academy of Child and Adolescent Psychiatry*, 54, 11–24.
- Lai, M. C., Lombardo, M. V., Pasco, G., et al. (2011). A behavioral comparison of male and female adults with high functioning autism spectrum conditions. *PLoS ONE*, 6(6), e20835.
- Lai, M. C., Lombardo, M. V., Ruigrok, A. N., et al. (2012). Cognition in males and females with autism: similarities and differences. *PLoS One*, 7(10), e47198.
- Lai, M. C., Lombardo, M. V., Ruigrok, A. N., et al. (2017). Quantifying and exploring camouflaging in men and women with autism. *Autism*, 21, 690–702.
- Lasègue, E.-C. (1873). De l'anorexie hystérique. *Archives Générales de Médecin*. English translation: On hysterical anorexia. *Medical Times and Gazette*, 265.
- Lebrun, Y. (1987). Anosognosia in aphasics. *Cortex*, 23, 251–263.
- Lehnhardt, F.-G., Falter, C., Gawronski, A., Pfeiffer, K., Tepest, R., & Franklin, J. (2016). Sex-related cognitive profile in autism spectrum disorders diagnosed late in life: Implications for the female autistic phenotype. *Journal of Autism and Developmental Disorders*, 46, 139–154.
- LePore, E., & McGlaughlin, B. (Eds.). (1985). *Actions and events: Perspectives on the philosophy of Donald Davidson*. Oxford: Blackwell.
- Mackie, P. (2006). *How things might have been*. Oxford: Oxford University Press.
- McGinn, C. (1995). Consciousness and space. *Journal of Consciousness Studies*, 2, 220–230.
- McGinn, C. (2004). *Consciousness and its objects*. Oxford: Oxford University Press.
- McPartland, J., Reichow, B., & Volkmar, F. (2012). Sensitivity and specificity of proposed DSM-5 diagnostic criteria for autism spectrum disorder. *Journal of the American Academy of Child and Adolescent Psychiatry*, 51, 368–383.
- Mellor, D. H. (1977). Natural kinds. *British Journal for the Philosophy of Science*, 28, 299–312.
- Millikan, R. (1984). *Language, thought and other biological categories*. Cambridge, MA: MIT Press.
- Millikan, R. (1993). *White queen psychology and other essays for Alice*. Cambridge, MA: MIT Press.
- Mundt, C. (1996). Zur Psychotherapie des Wahns [Psychotherapy for delusion]. *Nervenarzt*, 67, 515–523.
- Murphy, D. (2006). *Psychiatry in the scientific image*. Cambridge, MA: MIT Press.
- National Autistic Society. (2010). *SocialEyes: Exploring the social world with people on the autism spectrum*. London: National Autistic Society.
- Papineau, D. (1993). *Philosophical naturalism*. Oxford: Blackwell.
- Parish-Morris, J., Liberman, M. Y., et al. (2017). Linguistic camouflage in girls with autism spectrum disorder. *Molecular Autism*, 8, 48.

- Prigatano, G. P., & Schacter, D. L. (1991). *Awareness of deficit after brain injury: Clinical and theoretical issues*. New York: Oxford University Press.
- Pryor, T., Johnson, T., Wiederman, M. W., & Boswell, D. L. (1995). The clinical significance of symptom denial among women with anorexia nervosa: Another disposable myth? *Eating Disorders: The Journal of Treatment and Prevention*, 3, 293–303.
- Putnam, H. (1970). Is semantics possible? *Metaphilosophy*, 1, 187–201. (Reprinted in Putnam (1975b): 139–152).
- Putnam, H. (1975a). The meaning of ‘Meaning’ (pp. 215–271).
- Putnam, H. (1975b). *Mind, language and reality*. Cambridge: Cambridge University Press.
- Ramachandran, V. S., & Rogers-Ramachandran, D. (1996). Denial of disabilities in anosognosia. *Nature*, 382, 501.
- Rogers, R., & Bender, S. D. (Eds.). (2018). *Clinical assessment of malingering and deception* (4th ed.). New York: Guilford Publications.
- Salmon, N. (1981). *Reference and essence*. Princeton: Princeton University Press.
- Sass, L. A. (1994). *The paradoxes of delusion: Wittgenstein, Schreber, and the schizophrenic mind*. Ithaca, NY: Cornell University Press.
- Slater, M. (2015). Natural kindness. *British Journal for the Philosophy of Science*, 66, 375–411.
- Sperber, D., Premack, D., & Premack, A. (Eds.). (1995). *Causal cognition*. Oxford: Clarendon Press.
- Stone, S. P., Halligan, P. W., & Greenwood, R. J. (1993). The incidence of neglect phenomena and related disorders in patients with an acute right or left hemisphere stroke. *Age and Ageing*, 22, 46–52.
- Tsou, J. Y. (2007). Hacking on the looping effects of psychiatric classifications: What is an interactive and indifferent kind? *International Studies in the Philosophy of Science*, 21, 329–344.
- van Regenmortel, M. H. V. (1992). Concept of virus species. *Biodiversity and Conservation*, 1, 263–266.
- Vandereycken, W. (2006). Denial of illness in anorexia nervosa. *European Eating Disorder Review*, 14, 341.
- Vuilleumier, P. (2004). Anosognosia: The neurology of beliefs and uncertainties. *Cortex*, 40, 9–17.
- Walvoort, S. J., van der Heijden, P. T., Kessels, R. P. T., & Egger, J. I. M. (2016). Measuring illness insight in patients with alcohol-related cognitive dysfunction using the Q8 questionnaire: A validation study. *Neuropsychiatric Disease and Treatment*, 12, 1609–1615.
- Wilson, R. A. (Ed.). (1999). *Species: New interdisciplinary essays*. Cambridge, MA: MIT Press.
- Wilson, R. A., Barker, M. J., & Brigandt, I. (2007). When traditional essentialism fails: Biological natural kinds. *Philosophical Perspectives*, 35, 189–215.
- Wing, L., & Gould, J. (1979). Severe impairments of social interaction and associated abnormalities in children: Epidemiology and classification. *Journal of Autism and Developmental Disorders*, 9, 11–29.
- Yergeau, M. (2010). Circle wars: Reshaping the typical autism essay. *Disability Studies Quarterly*, 30. <http://dsq-sds.org/article/view/1063>.
- Yergeau, M. (2013). Clinically significant disturbance: On theorists who theorize theory of mind. *Disability Studies Quarterly*, 33. <http://dsq-sds.org/article/view/3876>.