The many theories of mind: eliminativism and pluralism in context

Joe Gough

Received: 10 August 2021 / Accepted: 4 July 2022
© The Author(s) 2022

Abstract
In recent philosophy of science there has been much discussion of both pluralism, which embraces scientific terms with multiple meanings, and eliminativism, which rejects such terms. Some recent work focuses on the conditions that legitimize pluralism over eliminativism – the conditions under which such terms are acceptable. Often, this is understood as a matter of encouraging effective communication – the danger of these terms is thought to be equivocation, while the advantage is thought to be the fulfilment of ‘bridging roles’ that facilitate communication between different scientists and specialisms. These theories are geared towards regulating communication between scientists qua scientists. However, this overlooks an important class of harmful equivocation that involves miscommunication between scientists and nonscientists, such as the public or policymakers. To make my case, I use the example of theory of mind, also known as ‘mindreading’ and ‘mentalizing’, and broadly defined as the capacity to attribute mental states to oneself and others. I begin by showing that ‘theory of mind’ has multiple meanings, before showing that this has resulted in harmful equivocations of a sort and in a way not accounted for by previous theories of pluralism and eliminativism.

Keywords Mindreading · Theory of mind · Autism · Polysemy · Pluralism · Eliminativism

1 Introduction
Many scientific terms, including technical terms and fairly recent neologisms, have multiple different meanings. Taylor & Vickers (2017) offer a helpful list of such
terms, each highly disputed. They offer a theory about how such terms come to have multiple meanings, a process which they call ‘fragmentation’: as different areas of science take up a term/concept for different theoretical purposes and using different measures (see also Wilson 2006), precise context-specific definitions and extensions of that term proliferate. The end result is that a single term ends up with multiple distinct but equally legitimate more precise meanings – although the fact that the term has multiple meanings may not be widely noticed.

The phenomenon of scientific terms with multiple meanings was originally noticed and highlighted by scientists and philosophers working in and on the life sciences, especially biology. It was, therefore, seen as insignificant by some, because it might reflect the immaturity of these sciences, or the peculiarities of their subject-matters (see Ruphy 2017, ch.3 for discussion). The physical sciences, physics and chemistry, conversely, were seen as primarily using univocal terms, to refer to well-defined, natural kinds, often thought to be definable in terms of microstructural essences (cf. Bursten, 2018). The idea was that while terms like ‘species’ and ‘human’ might have multiple meanings, terms like ‘electron’ and ‘Argon’ have only a single meaning within the relevant sciences. More recently, however, there has been some pushback on the idea that these sciences are generally free of this phenomenon (e.g., Bursten 2018; Chang, 2012; Ruphy, 2010, 2017; Ruthenberg & Mets, 2020; Slater, 2009; Wilson, 2006) – indeed, so long as these sciences contain different theoretical and explanatory contexts, much of the groundwork has been laid for terms to start accruing meanings (Taylor & Vickers, 2017).

I will use the term ‘theory of mind’ as a case study to explore this phenomenon, arguing that the origin and tenability of its many meanings can only be satisfactorily approached by considering how it is significant in nonscientific ways. In doing so, I will consider some previous theories of ‘pluralism’, which defends the use of terms with multiple meanings, and ‘eliminativism’, which advocates rejecting such terms. Often, this ‘eliminativism’ is driven by worries about miscommunication, equivocation, and empty arguments about the ‘correct’ extension or definition (Taylor & Vickers, 2017). Other forms of eliminativism are more directly ‘ontological’ (e.g., Churchland 1981, 1993), that is, more directly concerned with existence claims than with the use of terms/concepts – although it is not clear from a pragmatist point of view that this distinction stands up (e.g., Peirce 1878). Conversely, ‘pluralists’ defend these terms as beneficial – as better reflecting the way the world is (as discussed in Ereshefsky 1992), or as playing certain key ‘bridging roles’, facilitating communication and integration between different disciplines, subdisciplines, or researchers with different, legitimate conceptions of the term (Neto, 2020).

Taylor & Vickers (2017); Haueis (2021b) offer heuristics or criteria for deciding between pluralism and eliminativism, or for ensuring pluralism does not lead to problems. Many papers discussing eliminativism and pluralism focus on scientific

1 There are other factors that support the idea that this phenomenon is extremely widespread. Haueis (2021b) highlights that polysemy is an extremely widespread feature of natural language. Additionally, according to at least Pritchard (2019), studies of analogue cognition suggest that word meanings are not for categorization, but instead primarily for suggesting analogies in some way, and in particular in a way that encourages rather than hinders contextual variation in reference and extension (see also, e.g., Lakoff 1987; Recanati, 2017).
discourse over ordinary discourse. Many of these papers focus on terms with limited or largely inconsequential contemporary use outside scientific discourse. Perhaps because of these facts, the heuristics or criteria offered have focused on heuristics and criteria heavily geared towards avoiding miscommunication between scientists. Taylor and Vickers offer two heuristics, both with *ceteris paribus* clauses – first, that the more ‘theoretical roles’ a term takes on, the stronger the case for elimination; secondly, the more pivotal the term is in arguments, the stronger the case for elimination. Haueis claims that pluralism is defensible wherever the term is associated with reliable measurement techniques, a homogeneous domain, and pick out properties significant in description, explanation, or classification of behaviours of entities in that domain.

If the question is when it is legitimate to make use of scientific terms with multiple meanings, then it is my aim to highlight a largely unacknowledged aspect of answering this question. In particular, the meanings and significance of a term outside scientific contexts can both lead to miscommunications and equivocations between scientists and nonscientists, and increase the likelihood of equivocations within science. Indeed, one of the most potent roles that a scientific concept can have is in ‘[affecting] the conceptual self-understanding of a society as a whole’ (Haueis and Slaby 2022, p.3). In some cases, making use of a scientific term with nonscientific significance unnecessarily invites misuse and misunderstanding of one’s conclusions. In such a scenario, one presumably ought to avoid using the term. All one needs for such a conclusion is a plausible general principle that if one *can* avoid using terms that are likely to lead to the misinterpretation of one’s conclusions by the public, politicians, and policymakers, one *ought* to avoid using such terms.²

In the following, I argue that nonscientific significance sits at the heart of a dilemma for accounts that aim to adjudicate between accepting and rejecting a term with multiple meanings. Either they ignore nonscientific factors, and are unable safely to rule even on core examples in the debate, or they countenance such factors, and are unable to offer prescriptive guidance that is both detailed and generalizable. I further suggest that the second horn of this dilemma is preferable.

My argument proceeds by examining a case study, ‘theory of mind’, and the (by definition) synonyms, ‘mindreading’ and ‘mentalizing’. My aim here is not to argue that we should eliminate the term ‘theory of mind’. Instead, I argue that several features relevant to determining whether we should eliminate the term are ignored by any theory that focuses solely on the roles terms play within science and the ways terms with multiple meanings may hamper science. Likewise, my aim is not to take a stand on the debate between pluralism and eliminativism in general. Instead, my aim is to bring out the importance of these ‘nonscientific’ factors, which have generally not received much attention within this debate, and to argue that they are relevant to this debate because they are relevant to many of the particular cases of scientific terms where eliminativism and pluralism are live options.

The term ‘theory of mind’ has many different meanings (§ 2) – even though it is generally defined as the *capacity to attribute mental states to oneself and others*. Its

---
² For broader discussion on the role of (non-epistemic) values in science, see Douglas (2000); Longino (1990); Putnam (1982, 1984).
scientific use results in harmful equivocations outside scientific contexts (§ 3). Its significance outside scientific contexts also appears to be part of what causes equivocations in scientific contexts (§ 4). If the aim is to avoid harmful equivocations based on scientific terms with multiple meanings, this cannot only be a matter of examining the term’s use within science (§ 5). Unfortunately, this means that any theory that tries to avoid scientific equivocations by looking solely at scientific factors is at risk of ignoring some of the most important factors, the first horn of the dilemma. The second horn of the dilemma stems from a principled pessimism about the possibility of a detailed theory for adjudicating between pluralism and eliminativism that accounts for scientific equivocations based on nonscientific factors, let alone one that accounts for nonscientific equivocations (§ 6).

2 The many meanings of theory of ‘mind’

The term ‘theory of mind’, used interchangeably with ‘mindreading’ and ‘mentalizing’ which were introduced as synonyms to ‘theory of mind’, was originally defined as the ability to attribute mental states to oneself and others, in the paper Does the chimpanzee have a theory of mind? (Premack & Woodruff, 1978). The first ‘cognitive’ version of the social theory of autism (which largely supplanted previous sensory theories of autism; Robertson & Baron-Cohen 2017) was put forward in Baron-Cohen et al., (1985) paper, Does the autistic child have a theory of mind? which picked up the term, proposing that a theory of mind deficit was core to autism, and recalled the title of Premack and Woodruff’s paper (see also Happé & Frith 2020).

More recently, many researchers in animal psychology have come to use ‘theory of mind’ as a ‘generic label … covering a wide range of processes of social cognition’ (Tomasello et al., 2003b, p.239; see also Tomasello, Call, & Hare, Tomasello et al., 2003a). On this construal, different animals have different theory of mind abilities – for example, Tomasello, Call, and Hare claim that chimpanzees can understand seeing, but not believing. Other animal psychology researchers, for example, Povinelli & Vonk (2003) define theory of mind as an ‘all-or-nothing’ capacity, either present or absent in any given species, that involves ‘an attributed experience’ represented in a ‘non-behavioural code’ that cannot be reduced to representations of abstract classes of observables, such as ‘abstracted spatio-temporal invariances’ (p.157) or abstract classes of behaviour (Penn & Povinelli, 2007).

In autism research, the idea that a deficit in theory of mind is the ‘core deficit’ of autism has been abandoned by many researchers. Sometimes it is abandoned in favour of theories that emphasise that autism is an umbrella term describing a kind of syndrome not a specific aetiology or specific set of symptoms. Sometimes it is abandoned in favour of domain-general theories. These domain-general theories often have a more sensory flavour. They often work on the idea that autism involves an inability to sort signal from noise, thereby impairing gestalt perception, perhaps including the perception of mental states. For example, there is a Bayesian theory according to which autism is some kind of overweighting of sensory input over top-down expectations (e.g., Karvelis et al., 2018).
However, the idea that impaired theory of mind is the core characteristic symptom of autism has not been widely abandoned. The claim that autism is a deficit in theory of mind is now offered by many as a characterization of rather than a hypothesis about autism. For example, Gernsbacher & Yergeau, found that.

over 75% … of the top 500 articles indexed by Google Scholar (for “theory of mind” and “autism”) simply assert that autistic people lack a theory of mind rather than provide original data to buttress the claim. (2019, p.103)

An uncharitable interpretation of this practice of asserting that autistic people lack a theory of mind without evidence is that people are asserting a highly contentious hypothesis about autism entirely without justification. A more charitable interpretation is that they are offering a characterization of autism, which they take to be merely descriptive and largely uncontroversial. This gives us perhaps the most significant meaning of ‘theory of mind’ in autism research – the general sort of social ability thought to be impaired or atypical in autism. It is worth noting that the cluster of social abilities atypical in autism includes various abilities to do with the attribution of features not generally considered ‘mental’, for example, social rank recognition (Ogawa et al., 2019).

Some articles, however, talk not of an impaired or atypical theory of mind, but of a lacking theory of mind – an idea also implied by terms like ‘mindblindness’ (for review see Gernsbacher & Yergeau 2019; Gernsbacher, 2018). When talking of autistic people as lacking a ‘theory of mind’, the term can mean one of at least two things. A common (but unjustified, and in my view clearly false – see § 3) claim is that autistic people are entirely unable to understand the minds of others (and perhaps themselves) – unable to ‘imagine what others are thinking, or even that they are thinking’ (Soper & Murray, 2012, p.125); such claims are generally confined to popular books, textbooks, and encyclopaedias in psychology, and rarely made in research papers. In this first case, the ‘theory of mind’ supposed to be lacking is simply the ability to understand minds.

Another common way to flesh out the claim that autistic people lack a theory of mind, which is less obviously false although in my view still unjustified (again see Gernsbacher & Yergeau 2019), is that autistic people lack certain basic mechanisms that embody some basic understanding of mental states, develop along clearly defined tracks, and whose presence in neurotypical humans enables them to effortlessly develop a full suite of abilities for understanding minds (e.g., Bloom & German 2000). In this second case, ‘theory of mind’ refers to this set of basic mechanisms.

Many uses of ‘theory of mind’ in autism research refer to a construct developed primarily in some other specialism. The term ‘theory of mind’ has also been adopted by psychometrics, developmental psychology, and neuroscience. In developmental psychology, the term can be used to name a body of conceptual knowledge (e.g., Wellman & Liu 2004; see Apperly 2012). This notion of ‘theory of mind’ cannot be identified with any of the previous notions, but it has important relations to several – for example, it is thought to underlie the ability to make explicit attributions of mental states, and thought to develop as a result of the operation of the basic mechanisms that enable social development. In developmental psychology, ‘theory of mind’ is
also used to refer to a multi-component system defined in terms of multiple developmental milestones and responsible for a wide domain including the detection of intentionality and goal-directed behaviour (Bermúdez, 2020).

Use of the term is perhaps messiest in psychometrics, where even excluding single-task measures, there is no single kind of mental state attribution or ability assessed in even close to a majority of ‘theory of mind’ measures (Beaudoin et al., 2020), and where some even question whether ‘theory of mind tasks’ can legitimately be described as measuring mental state attributions (Quesque & Rossetti, 2020). Worse still, many of these measures and tasks fail to converge with one another, and to predict behaviours and abilities in the domain of pretheoretic interest such as prosocial behaviour, empathy, and everyday social skills (Gernsbacher, 2018).

One could take one of at least three stances on this state of affairs. The first option, introducing no new meaning of ‘theory of mind’, is that ‘theory of mind’ measures are an attempt to measure one of the other preexisting notions of theory of mind, and that they do so with widely varying quality. The second option, likewise introducing no new meaning of ‘theory of mind’, is that ‘theory of mind’ measures as a whole fail to measure anything – that they represent a failed attempt to find or define a psychometrically valid construct. The last option, introducing vastly more meanings of ‘theory of mind’, is that ‘theory of mind’ in each case refers to whatever each ‘theory of mind’ measure actually measures, and that this varies measure to measure. There are other options, but most represent a combination of these three strategies (e.g., holding that some ‘theory of mind’ measures fail to measure anything, while others measure something specific to that measure, and some cluster of measures serve as good measure of some preexisting notion of theory of mind). My aim here is not to take a stand on the status of potential psychometric conceptions of ‘theory of mind’, since there are enough other meanings of the term to show that it is a candidate for pluralism and eliminativism.

In neuroscience, the ‘theory of mind mechanism’ is a group of neurones in the right temporoparietal junction (rTPJ) thought to be responsible for the explicit attribution of certain complex intentional states like beliefs and thoughts (Saxe, 2009, 2010; Saxe & Kanwisher, 2003; Saxe & Powell, 2006; Scholz et al., 2009) – but not the attribution of pain or emotion (Kosakowski & Saxe, 2018; Saxe & Powell, 2006). It is worth noting that some think that the rTPJ, including this group of neurones, is responsible for gestalt perception and attentional gating, tying in nicely to domain-general theories of autism (Bloechle et al., 2018; Huberle & Karnath, 2012; Rennig et al., 2013; Schuwerk et al., 2017). The ‘theory of mind network’ is a loose collection of brain regions responsible for a variety of ‘theory of mind’ tasks, including most of the so-called ‘social brain’ thought to underlie social cognition (Baron-Cohen, 2009; Schaafisma et al., 2015). In each case, ‘theory of mind’ can be taken to refer either to the underlying neural system (for the ‘theory of mind mechanism’, the group of neurones in the rTPJ), or to refer to the function of that neural system (in this case, explicit attribution of certain complex intentional states). For the ‘theory of mind network’, specifying its function is difficult – the collection of brain regions is individuated by the fact it is involved in several theory of mind tasks, but as discussed above, the nature of these tasks and what they measure is extremely difficult to specify.
There are, therefore, many meanings of ‘theory of mind’, making it an appropriate case study for discussing theories of pluralism and eliminativism. Setting aside the problem of ‘theory of mind’ measures, the term ‘theory of mind’ can be used to mean, at least, two neural systems, their functions, a multicomponent system responsible for *(inter alia)* the detection of intentionality and goal-directed behaviour, a body of conceptual knowledge that enables the explicit attribution of mental states, the possession of any understanding of minds whatsoever, the basic mechanisms involved in the development of neurotypical adult mental state attributions, the cluster of social abilities thought to be impaired or atypical in autism, the ability to explicitly attribute of mental states where this is irreducible to the ability to recognize abstract classes of observables, and a generic label for whatever abilities to understand whatever mental states a given species has.\(^3\)

### 3 Unscientific harms

These are far from the only scientific meanings that the term has taken on. Of course, that is not a problem in itself – especially since the term is largely recognized as an umbrella term, or shorthand, in at least animal psychology and autism research. Bermúdez (2020, p.335) describes the term ‘mindreading’ as ‘a very general label for the skills and abilities that allow us to make sense of other people and coordinate our behaviour with theirs’. Although ‘theory of mind’, ‘mindreading’, and ‘mentalizing’ have taken on many subtly distinct meanings, and although there is no reliable measure associated with most uses of the terms, they are in this regard not obviously problematic. There are very few examples of serious scientific errors based on equivocation between these meanings – which is not to say there are *no* plausible cases of such errors (see § 4).

However, in terms of the reception and marketing of theories of ‘theory of mind’, there are some very serious errors based on equivocation between these meanings. The central error is the lumping together of autistic people, especially autistic children, and nonhuman animals as lacking a ‘theory of mind’, where theory of mind is seen as ‘one of the quintessential abilities that makes us human’ Baron-Cohen 2000b.

---

\(^3\) One might object to the preceding as follows: ‘theory of mind’ is essentially defined as *the ability to attribute mental states to oneself and others,* and all other working definitions are somehow derived from this ‘generic’ definition; it is therefore univocal, and not in the scope of the debate. I suspect that this is a red herring. In my view, the implicit consensus is that a term in the scope of this debate – a ‘pluralistic’ term – is one whose use is governed by different working definitions in different contexts, where these working definitions are importantly distinct and yet not in legitimate competition (i.e., do not reflect genuine disagreement about the nature of a phenomenon; see also Feyerabend 2001; Kidd, 2012). This is an important category not because of the underlying linguistic mechanisms but instead because of the scientific, metaphysical, and especially methodological issues raised by such terms (see also Bursten 2018). If I am right, many pluralistic terms may even come out as *univocal* by some definitions because it may be possible to give them a ‘generic’ definition so long as it is appropriately parametrized and imprecise (Akagi, 2018, 2021). More pressingly, in the sense relevant to linguistics, the question of univocity may be unanswerable – most tests of univocity/ambiguity rely on speaker-intuitions about whether utterances are felicitous, and are therefore hard to bring to bear on technical terms used only in isolated specialist areas. See also footnotes 8; 9.
p.169), rendering autistic people as ‘biologically set apart from the rest of humanity in lacking the basic machinery’ (Baron-Cohen, 2009, p.73).

In a clear sense, such claims are only rarely part of the scientific study of autistic people or animals – they generally play no serious role in the conclusions or argument. Instead, they form part of a ‘marketing strategy’, a ‘hook’ for drawing in readers and presumably funding. It is also based on a clear equivocation. While animals’ purported lack of theory of mind turns on their purported over-reliance on low-level cues, and inability to explicitly attribute such states (e.g., Heyes 2014; Penn & Povinelli, 2007), the precise opposite is true for autistic people. Autistic people are claimed to be too reliant on explicit, perhaps linguistically-mediated attribution of mental states and impaired in their ability to tune into and easily make use of low-level cues (e.g., Gernsbacher & Yergeau 2019). Even supposing that both theories are correct, the ‘theory of mind’ lacked by animals and the ‘theory of mind’ lacked by autistic people are simply not the same thing.

Part of the trouble is that the terms ‘theory of mind’ and ‘mindreading’ have taken on a great deal of ethical significance, given the role they have taken on in contemporary theories of personhood. Dennett (1976), whose 1978 commentary on Premack and Woodruff’s paper had a major influence on the direction of research, helping cement the false-belief paradigm, directly links the ability to attribute mental states to reason, reflectiveness, decision-making, and personhood. More recently, Lurz (2011, p.4–5) claims ‘that of the attributes that define personhood, mindreading is the most central.’

Baron-Cohen (2000a, p.266) treats autistic people as some sort of morality tale about the significance of theory of mind. He claims that ‘autism is a clear illustration of what human life would be like if one lacked a theory of mind’. He envisages this as a world in which humans see each other as grotesque, terrifying ‘bags of skin … stuffed into pieces of cloth’, ‘noisy skin-bags’ with ‘no way of explaining … or predicting’ each other (Alison Gopnik, who is not autistic, as quoted in Baron-Cohen 1995, p.4–5).

This dehumanization is stigmatizing in itself, and is in my view the primary harm that results from this way of construing autism. However, it is not the only way in which conveying autism as a ‘theory of mind’ deficit has proven stigmatizing (e.g., Gernsbacher & Yergeau 2019; Yergeau & Huebner, 2017). For example, conveying autism this way is frequently linked to dismissals of the opinions of autistic people on autism – if an autistic individual demonstrates a significant level of introspection, it is claimed that they must not really be autistic (Yergeau & Huebner, 2017), and so their personal account of autism can be dismissed.

For example, Oliver Sacks, upon first reading Temple Grandin’s autobiographical account of her life and experience with autism, Emergence: Labeled Autistic (1996), could not help being suspicious of it: the autistic mind, it was supposed at the time, was incapable of self-understanding and understanding others and therefore of authentic introspection and retrospection (Sacks, 2012, p.241).

This common belief at the time, that autistic people lack the ability to introspect, was rendered plausible by the messiness of the term ‘theory of mind’. There are at least
two senses of ‘theory of mind’ at play here, worth reintroducing. The first is ‘theory of mind’ as the ability to attribute mental states to oneself or others. The second is ‘theory of mind’ as the basic mechanisms that normally support the development of this ability in neurotypical humans. Arguably, the research literature when Sacks was writing offered some support for the claim that autistic people were impaired in the first, and some support that they lacked the second. Even so, because autistic children apparently struggled on false-belief tasks in a way not accounted for by general cognitive impairment, many came to believe that autistic people were entirely unable to introspect. It is hard, perhaps impossible, to see how people en masse could have acquired this harmful, false belief without the construct theory of mind to facilitate the inference from extremely specific developmental evidence, to an ostensibly absurd conclusion about autistic people.4

In my view, the above situation represents a nonscientific harm, resulting from a nonscientific equivocation, in turn resulting from nonscientific factors, including prior normative views, the social context of research, and the nonscientific significance of ‘theory of mind’. The term ‘theory of mind’ builds bridges between the study of the differences between autistic people and neurotypical humans, the differences between humans and nonhuman animals, and theorizing about the conditions of personhood. This interfaces very poorly with the fact that autistic people are a highly stigmatized group, and that dehumanization is an important mechanism for stigma. The equivocation seemingly results in part from a prior willingness to take a dehumanizing view of autistic people. It is worth stressing here that this equivocation does more than merely reflect preexisting harms. Certainly, the equivocation may be caused in part by prior social attitudes towards autistic people. Nevertheless, the equivocation may further cement these attitudes by lending them spurious scientific legitimacy. This equivocation is nonscientific in that, rather than primarily being a matter of miscommunication between scientists in their roles as scientists, the issue is that the term misleads important groups of nonscientists, such as policymakers and the public.

4 Scientific harms, unscientific causes

It is not only nonscientific equivocations and harms that result from nonscientific factors – such factors can also increase the risk of scientific equivocations. One of the few plausible significant cases of a scientific equivocation based on ‘theory of mind’ results from the nonscientific significance of ‘theory of mind’. In the debate over theory of mind in nonhuman animals, two senses of ‘theory of mind’ appear to be conflated regularly. This appears to be a result of a particular sort of theorizing about human specialness, which goes beyond a scientific concern.

4 To his credit, Sacks’ response to this situation was to meet with Temple Grandin, to write about her in a careful and empathetic manner, and to encourage others to read her autobiography. When his view of autism was challenged by a first-person account of autism, he changed his view of autism. However, as Yergeau & Huebner (2017) show, this has been far from a universal response to such accounts of autism. Many researchers have doubled down, dismissing such accounts of autism as showing that their authors are not really autistic.
Those who hold that there are nonhuman animals that have a ‘theory of mind’ tend to take a relatively ‘deflationary’ view of theory of mind. According to this more deflationary view, the attribution of mental states is manifest in sensitivity to equivalence classes best defined in relation to mental state terms, and not definable in relation to any particular observable – this meaning of ‘theory of mind’ is not mentioned in § 2, but is associated with the view of theory of mind as a generic label for certain social abilities of whatever given species. Conversely, those who hold that nonhuman animals lack, or have not been shown to have, a ‘theory of mind’ tend to use a more ‘inflationary’ definition. According to this definition, the attribution of mental states requires the explicit internal representation of mental states qua mental states (Burge, 2018; Heyes, 2014; Lurz, 2011).

The more deflationary definition of ‘theory of mind’ is arguably the default – some of its opponents seem to acknowledge it as such (e.g., Heyes 2015). Its opponents’ definition, however, is not obviously illegitimate (although it is linked to some possibly intractable methodological problems; see Halina 2015; cf. Burge, 2018). What is illegitimate, however, is using one legitimate sense of ‘theory of mind’ to object to claims made using another legitimate sense of ‘theory of mind’.

It is plausible that this goes on in many discussions of the ‘logical problem’ (e.g., Penn & Povinelli 2007; Povinelli, 2004; Povinelli & Vonk, 2003, 2004). The logical problem is used as an objection to all previous evidence of theory of mind in nonhuman animals. It is claimed that no previous experimental paradigms are exacting enough to distinguish theory of mind from ‘behaviour reading’. For example, if it is found that chimpanzees track and respond appropriately to other chimpanzees’ gaze direction, it is claimed that this cannot prove that they attribute vision. According to this objection, they could just be inferring directly from gaze to behaviour, without explicitly representing vision as mediating between their conspecific’s gaze and behaviour.

This objection relies on the more inflationary sense of ‘theory of mind’ (see also Halina 2015). The reason for this is that the deflationary definition of ‘theory of mind’ entails that theory of mind grades into behaviour-reading, and that to be capable of sufficiently smart behaviour reading just is to have a theory of mind (Whiten, 1996). To demonstrate the presence of ‘theory of mind’ in this more deflationary sense, it is neither possible nor necessary to rule out behaviour reading.

However, the logical problem is used as an objection to conclusions pitched in terms of the more deflationary sense of ‘theory of mind’. If both are legitimate, as I think that they are, then these kinds of objections cannot work. Moreover, with notable exceptions (e.g., Heyes 2015), those who make use of the inflationary sense, and object to claims made using deflationary sense, do not pitch their arguments as to do with the meaning of ‘theory of mind’. They appear to end up talking across previous researchers, on the basis of an equivocation between different senses of ‘theory of mind’.

Why does this occur? It appears to be partly a matter of the nonscientific significance of ‘theory of mind’. A key factor is the ethical significance of ‘theory of mind’

---

5 This definition was developed by Whiten (1996; 2001), foreshadowed by Gómez (1991), and endorsed by, e.g., Hare et al., (2000) and Halina (2015).
in theories of personhood, discussed above. Another factor is the widespread use of psychological theorizing to support theorizing about what makes human beings special, ethically, culturally, and psychologically (Sorabji, 1993). The role of ‘theory of mind’ in such ethically-loaded theorizing lends it normative connotations that different researchers may want ‘on their side’ (Gallie, 1955). Part of the motivation for putting forward the logical problem and objecting to previous research is that it casts nonhuman animals as inappropriately similar to humans, in the eyes of those who offer the objection (e.g., Caporeal & Heyes 1997; Heyes, 2014, 2015; Penn et al., 2008; Povinelli, 2004; Povinelli & Vonk, 2003). Compare the example of ‘vision’. ‘Vision’ has been applied to systems as different to humans as individual bacteria (e.g., Nilsson & Colley 2016; Schuergers et al., 2016). Such claims generate comparatively little controversy: vision is not widely seen as a particularly ethically significant capacity; it has not been widely held to be unique to humans.

The ethical significance of ‘theory of mind’, and the increased motivation to defend theories of human specialness, are not entirely a matter of a term’s use within science. However, it appears that the use of the logical problem as an objection is, in many cases, an instance of obstructive equivocation, and researchers talking across one another. As such, the term’s nonscientific roles appear also to be relevant to the risk of miscommunication within science.

5 The pervasiveness of nonscientific significance

‘Theory of mind’ therefore suffers from exactly the sort of problem that motivates many eliminativists – it has many meanings, and this results in harmful equivocations, both nonscientific (§ 3) and scientific (§ 4). However, several key factors behind these equivocations are nonscientific – social context, normative assumptions, discriminatory biases, and the ethical significance of the term. My primary aim across the rest of this paper is to argue that nonscientific factors present an interesting challenge to theorists of pluralism and eliminativism.

In particular, I aim to press a dilemma. Any theory of eliminativism and pluralism, even one focussed solely on scientific equivocations, faces this dilemma. The first horn is to ignore nonscientific factors and thereby lose a significant amount of scope. The other is to countenance such factors and, as I will argue in § 6, thereby become intractable. In this section, in order to push the first horn of the dilemma, I will argue that nonscientific significance is pervasive among the key terms in the scope of the debate, and that nonscientific factors can play several key roles relevant to the debate.

Previous theories of eliminativism and pluralism generally do not account for nonscientific factors. This is not necessarily a problem on their own terms, and certainly need not reflect a view that such factors should be set aside. There may also be terms on which it is possible to come to a ruling without any great consideration of nonscientific significance. Many of the key terms, and especially key candidates for pluralism over eliminativism, lack any great nonscientific significance (e.g., Haueis’ 2021a ‘cortical column’), or have a form of nonscientific significance that is innocuous and perhaps even helpful for communication and identification of a coarsely-demarcated domain or phenomenon of pretheoretic interest (e.g., Wilson’s 2006 ‘hardness’).
Even so theories that do not account for nonscientific factors seem ill-placed to account well for many of the central candidates for eliminativism or pluralism. Included on Taylor and Vicker’s (2017) list are, among others, ‘consciousness’, ‘race’, ‘intelligence’, ‘health’, ‘life’, ‘scientific explanation’, ‘scientific method’, and ‘scientific theory’. All of these terms, in my view, have a great deal of nonscientific significance – the latter three least obviously so, but given the public importance of the demarcation problem (distinguishing science from nonscience, especially pseudoscience), I think it is hard to deny that they do. There are many other possible examples, not included on this list, for example, ‘wellbeing’ (Alexandrova, 2017) and ‘democracy’ (Krauss, 2016; Nagel, 1961) – both of which have many meanings, and both of whose scientific and nonscientific use is heavily shaped by nonscientific factors.

Even if a theory remains squarely focussed on (avoiding) scientific equivocations, it is hard to see how it could reliably come to the right ruling in such cases if one focusses solely on scientific factors. There are two main reasons for this. The first is that the normative content of a term may increase the likelihood of equivocations – independently of the scientific factors discussed by Haueis (2021) and Taylor & Vickers (2017). For ‘theory of mind’ to result in the kind of equivocation that I think is going on in the use of the logical problem as an objection to previous animal psychology research on the topic, it need only have two scientific meanings (cf. Taylor & Vickers, 2017), both of which could be associated with a reliable measure and significant in explaining the relevant animals’ behaviour (cf. Haueis 2021) – in particular, it need only have the meaning assigned to it by Povinelli & Vonk (2003), and the meaning discussed by Halina (2015).

Even if these were the only scientific meanings, and even if both were associated with a reliable measure, it might still lead to harmful equivocations within science on the basis of its nonscientific, ethical significance. The point here is not that Haueis’ theory or Taylor and Vickers’ theory would come to the wrong ruling, which they may or may not; the point here is that their theories appear not to track the relevant factors in raising the likelihood of scientific equivocations based on ‘theory of mind’. Where there are normative connotations to compete over, these have long been recognized as increasing the risk of equivocation (Gallie, 1955). As such, a theory that tracks the risk of harmful scientific equivocations based on scientific factors may well fail to track the actual risk of harmful scientific equivocations, making it less informative even on cases central in anchoring the contemporary debate.

There is an even more troublesome role that nonscientific significance may play. In certain cases, the nonscientific significance of a scientific term may make it indispensable, ruling out eliminativism even though there is an extremely high risk of equivocation. Such a defence does not, in my personal opinion, work in the case of ‘theory of mind’. It seems fairly plausible to me that ‘theory of mind’ links animal psychology and autism research to a particularly bad or pointless kind of ethical discourse. For example, Sorabji (1993) identifies a pattern of reasoning common across the history of Western philosophy, looking for a single feature that distinguishes humans from all other animals. However, there need be no one feature that so distinguishes us; additionally, it need not be the same feature that distinguishes us from
each other kind of animal (Carpenter, 2018; Sorabji, 1993) also highlights that the search for a single distinguishing feature is a peculiarly Western concern.

Compare, however, discussions of wellbeing. If the normative significance of ‘wellbeing’ – the normative component shared across different definitions and meanings of the term – is something like an inherently desirable state which it would be good for (such-and-such) individuals to be in, then it is hard to see how one could eliminate the term ‘wellbeing’ from scientific discourse, even if it is inevitable that people will adopt different definitions of the term, and will talk across one another after having done so. The reason for this is that one might think that the benefits of scientifically-informed theories of wellbeing for humans (and subclasses thereof; Alexandrova 2017), as well as the benefits of scientific theories of how to attain such a state, outweigh the inevitable equivocations that come from making use of such a normatively-loaded term in science.

In other words, while ‘theory of mind’ may link together certain areas of science inappropriately linked to one another and a particularly unhelpful kind of normative theorizing, ‘wellbeing’ may link together certain areas of science appropriately linked to one another and an extremely helpful kind of normative theorizing. While the former casts the resultant equivocations as needless and egregious harms, the latter casts the resultant equivocations as necessary but fairly minor evils in pursuit of a much greater good. Overall, what this means is that the nonscientific significance of scientific terms may also make equivocations more or less tolerable, and may therefore mean that eliminativism is inappropriate even in the face of widespread scientific equivocations. More pressingly for a theory concerned solely with scientific equivocations, it may render such equivocations inevitable – eliminating a term like ‘wellbeing’ may be an entirely pointless endeavour, if it means that another term will simply come to replace it (see also Wilson 2006).

There are also other ways that nonscientific significance and nontechnical meanings can support pluralism in a particular case. I mentioned earlier that in Wilson’s (2006, ch.6) example of ‘hardness’, it is extremely plausible that the ordinary meaning of the term helps the term to play its roles within science – that ‘hardness’ is a useful scientific term in part because it has a nontechnical meaning that is both widely understood, and helpful for generating legitimate and rarely-confused technical meanings (corresponding to different ways of probing materials). More generally, one might think that it is often a good idea to use terms with nonscientific significance in science, since such terms tend to be easily comprehensible and tied to coarsely-demarcated domains of public and pretheoretic interest.

Any theory of eliminativism and pluralism, even one focussed solely on scientific equivocations, therefore loses a significant amount of scope if it does not countenance non-scientific factors. If such a theory deals solely with scientific factors, then it cannot come to an overarching ruling on terms with nonscientific significance, and this includes many key terms in the debate. This is because nonscientific factors may play many important roles relevant to choosing between eliminativism and pluralism. They may increase the likelihood of equivocations because they may lend terms normative connotations that researchers want associated exclusively with their preferred

---

6 This point was stressed to me by an anonymous reviewer.
technical definitions. They may also mean that eliminativism even with respect to a term that leads to many equivocations is undesirable from a broader perspective, and pointless even if one’s sole interest is avoiding scientific equivocations because another term with the same normative significance will come to replace it.

6 The prospects of a general theory

The other horn of this dilemma is that one accepts that theories of pluralism and eliminativism (even those focussed solely on scientific equivocations) need to account for the nonscientific significance of scientific terms, and the range of factors to which they must be sensitive becomes much larger, perhaps indefinitely large. This makes generalizable normative guidance extremely difficult to specify, a point that I will press in this section.

I do not believe that it is likely that we will be able to build a normative theory of pluralism and eliminativism that manages to be both sensitive to enough factors and sufficiently specific in its prescriptions and criteria, that it can offer clear and correct guidance in different cases. It is plausible, in my view, that there are no general rules that govern when scientific terms with multiple meanings lead to equivocations that are harmful, avoidable, and not outweighed by the benefits that such terms can bring. If this were so, any assessment or decision would have to be open-ended and case-by-case – which in my view is neither particularly surprising, nor particularly bad news!7

Of course, at a very general level, many different cases have certain abstract features in common.8 As highlighted by an anonymous reviewer, one important feature which comes out clearly from discussion of ‘theory of mind’ is that scientific terms gain traction in nonscientific contexts because they (appear to) have some ‘ordinary’ meaning which one can grasp without grasping their more technical meanings (again, see Wilson 2006). Many scientific terms with many meanings, including neologisms and technical terms, are introduced precisely because they have some ordinary meaning approximately appropriate for the relevant technical notion. It is the precise nature of the gap between ordinary and technical meanings, in addition to

7 Indeed, this open-endedness may be an inevitable feature of the nature of concepts and conceptual practices if for example, conceptual practice is normative by nature (e.g., Haueis & Slaby, 2022), and if normative evaluation is necessarily open-ended (e.g., Rouse 2007), grounded in a relationship of mutual accountability between (candidate-)performers of a practice (including those criticising a performance).

8 One reason that it is difficult to get beyond this very general level is that the phenomenon of concern is not, in my view, a linguistically-unified phenomenon (or in other words, pluralistic terms do not constitute a linguistic kind). One issue is that multiple kinds of polysemy, vagueness, and context-sensitivity all plausibly play a role in generating new ‘meanings’. A more significant issue is that various mechanisms special to scientific contexts also do so, including operational definitions (e.g., Haueis 2021b; Wilson, 2006), scale-dependency (e.g., Bursten 2018; Haueis, 2021b), resolution-dependency (e.g., Ruphy 2017), and the use of concepts that ‘bridge’ observable/phenomenological and microstructural/explanatory levels of description (see again Bursten 2018). One particularly important phenomenon highlighted by an anonymous reviewer is that of ‘normal cotravellers’, where one term ends up referring (or is discovered to refer) to several different properties because they normally (appear to us to) co-occur (see also Field 1973).
the differences between technical meanings, that leads to many of the problems with ‘theory of mind’.

Such terms can help to build bridges between a domain of pretheoretic interest, areas of scientific research, and areas of theoretical or normative discussion. In an age of increasing specialization, this can be a hugely important service. It can also help to facilitate discussion and integration between different disciplines – for example, Neto (2020) discusses the services done to by biology in this way by ‘molecular gene’, ‘evolutionary novelty’, and ‘biological lineage’, all of which he counts as examples of imprecise concepts. The way these terms build bridges is plausibly by encouraging the perception of similarity, and discouraging the perception of difference, by a trick as simple as giving things the same name (or at least, letting them have the same name). This in turn provides a sort of flexible glue.9 Flexibility is afforded by differences in meaning. Glue is provided by encouraging people to see different areas as discussing the ‘same thing’10 (also, in the modern day, by the way scientific work is indexed and searchable).

In any given case, we can ask whether it is a good idea to build such bridges in the first place, and whether those bridges have been built well. We ought not to encourage the perception of relevance and similarity between some kinds of areas, especially not in a way that encourages the underplaying of differences. It can facilitate stigmatizing equivocations and enable illicit cultural obsessions to influence and obstruct scientific research. Thoroughly assessing the choice between pluralism and eliminativism means considering social attitudes – because these attitudes can raise the prior likelihood of the harmful equivocations potentially facilitated by a term with multiple meanings. It also means considering the legitimacy and necessity of the cultural pursuits that have laid claim to a term.

My aim in this section is to explain why I do not hold much hope for a general theory significantly more detailed than this brief sketch. The primary reason for this is that nonscientific factors of the sort considered in previous sections provide a seemingly open-ended way for terms to acquire new meanings, and for their various meanings and forms of significance to interact with each other and to behave in atypical ways.

Nonscientific significance interfaces with scientific use of terms reciprocally, and not in a particularly well-behaved manner. For example, ‘theory of mind’ took on some of its nonscientific significance partly because it was applied to autism, an already stigmatized condition that people were presumably all too ready to believe reflected a reduced level of humanity and an inability to understand others or introspect. Autism thereby came to serve as a sort of ‘proof’ of the significance of ‘theory

9 Ostensibly, some psychological evidence also supports this point (for discussion, see Hauens 2021b). However, such evidence primarily bears on how a term might have multiple related meanings in an individual’s idiolect, and how an individual might store and represent these different meanings. But no one individual need store or represent all the different meanings associated with a pluralistic technical term – indeed, often the problem is precisely that different speakers are unaware that there are different meanings in use.

10 This is not necessarily wrong – ‘discussing the same thing’ is often taken to be as coarse-grained as concepts, and therefore not as fine-grained as conceptions of, descriptions of, or beliefs about that ‘thing’ (e.g., Sawyer 2018).
of mind’ to humans in general, as discussed in § 3; this also seemingly increased the
profile of animal psychology research on the topic. It is, in turn, the success of these
areas of study that led to the proliferation of meanings of ‘theory of mind’, as objec-
tions to and defences of this broad scheme proliferated. Scientific uses can seemingly
therefore affect nonscientific significance, which can in turn affect scientific uses.

The specific nature of these effects, once more, appears to depend on the content
and context of the discussion. For example, there appear to be cases in the history
of science where the nonscientific significance of terms has led them to fall out of
favour, because scientists at the time were suspicious of the particular normative con-
notations involved – this is arguably true of the behaviourists’ scepticism of ‘mind’
and ‘consciousness’, and true on a larger timescale of ‘self’ and ‘soul’ (Martin & Bar-
resi, 2006). This opens the door to various kinds of disorderly case – for example, the
possibility that a scientific use of a term might increase its nonscientific significance,
in turn leading to scepticism of the term among scientists, and thereby resulting in
certain kind of harm, the needless abandonment of a useful scientific term.

One key issue here is that no good account of pluralism and eliminativism can be
‘formal’ if it is to account for nonscientific significance. The specific content and the
specific context matter – it matters who or what is being discussed, in what context,
for what reason, and so on. In particular, one’s view of the legitimacy of the cultural
pursuits that have laid claim to a term, and the prior normative views of research-
ers regarding what they are researching, affect the decision between eliminativism
and pluralism. More generally, where ‘theory of mind’ results in harmful equivoca-
tions, scientific and nonscientific, it is because of what (and who) is discussed in the
contexts that it links together, the stigma associated with these groups, and broader
cultural concerns. Much of the problem with ‘theory of mind’ is what kinds of areas
it links together. It is a bad idea to build a bridge between the study of autistic people,
who are a highly stigmatized group, and an area of animal psychology where many
researchers are looking for a psychological explanation of human specialness. Autis-
tic people are a stigmatized and marginalized group, and dehumanization is a core
mechanism of stigmatization and marginalization.

As such, it cannot solely be a matter of counting the range of purposes and gaug-
ing the significance of the term in arguments (cf. Taylor & Vickers, 2017), nor can
it solely be a matter of ensuring that the term stands in appropriate relations to sci-
entifically significant properties and reliable measures (cf. Haueis 2021) – which is
not to deny that any of these features is nevertheless highly significant. In fact, my
suspicion is that these more formal features are highly significant, thereby adding to
the already large range of potentially highly significant factors.

There is a still greater range of factors if one wishes also to avoid harmful nonsci-
entific equivocations – as I believe one ought. It is worth briefly defending my view
on this point. The core issue, as I see it, is that if a theorist of pluralism and elimina-
tivism deliberately sets aside nonscientific equivocations and harms, then they risk
finding themselves in a deeply morally dubious position.11 Suppose that there is a

11 It is very important here that I be clear that I am not accusing previous theorists of pluralism and elimi-
nativism of any moral failing. This would be a doubly absurd accusation. First, because many authors who
advocate for the retention of scientific terms with multiple meanings focus on morally innocuous terms

Springer
term that is scientifically in credit overall, but which leads to extremely harmful non-scientific equivocations. The theorist might offer their theory as claiming that the term is scientifically in credit, but not as advocating for its retention on this basis, but practically this seems a difficult and unstable position to hold in a debate. More likely, they will end up (or at least end up being understood as) advocating for pluralism and the retention of the term — not because they believe that the scientific benefits outweigh the nonscientific harms, but because they have set aside the nonscientific harms as being outside their purview.

Putting pluralism and eliminativism in context may be indispensable for making a well-informed decision, but it may also mean that no complete, overarching normative theory of eliminativism and pluralism is possible. It is important, however, not to overestimate how grim this would be. It opens the door to the possibility that different cases will look very different to one another. It may also make it hard to be sure that all relevant factors have been considered. Neither of these things, however, means that the decision would be impossible, or even necessarily difficult, in any particular case (see also Haueis & Slaby, 2022).

7 Conclusion

The term ‘theory of mind’ has many meanings and results in harmful equivocations. Previous theories of eliminativism and pluralism often identify this as the key danger of using scientific terms with multiple meanings. However, some of the most significant harmful equivocations that it leads to are not covered by current theories of pluralism and eliminativism. Additionally, the features of the term that lead to these equivocations are not those identified in current theories of pluralism and eliminativism.

The term ‘theory of mind’ and the research areas in which it is used are highly socially significant. This is not apparent when looking solely at scientific discussions of ‘theory of mind’. ‘Theory of mind’ does not lead to problems because it has an especially wide range of uses, fails to be associated with reliable measures, or fails to pick out scientifically significant properties. Instead, it leads to problems because it builds bridges between theorizing about a stigmatized group, research on nonhuman animals, and a particular kind of theorizing about human specialness.

It may be that in an ideal world, such problems would not arise, because the public and policymakers would not ‘misappropriate’ the findings of science even when scientists use ostensibly ordinary, nontechnical terms to describe their findings. Perhaps science, conceived as a process of collective inquiry, would be better-off in such a world, able to proceed without fear of misleading the public and policymakers. We (e.g., Haueis 2021a, b; Wilson, 2006), and although many authors who advocate for the elimination of terms that are not morally innocuous do largely ignore their moral dangers, they are working in an entirely morally defensible way — once they have gathered what they see as sufficient, maximally persuasive (which here likely means ‘scientific’) evidence in favour of elimination, it seems deeply unreasonable to insist that they gather more, more morally-salient evidence. Secondly, because explicit discussions of eliminativism and pluralism remain a fairly young, burgeoning area of discussion, and it is deeply unreasonable to accuse other authors of being immoral because they did not make the points I made by building on their insights.
do not live in that world. When it is sufficiently feasible to use alternative scientific terms that are less likely to encourage such misuses of science, scientists ought to do so. It surely would not undermine scientific progress to use different terms for the capacity to explicitly represent mental states, as nonhuman animals have often been claimed to lack, and the shortcuts neurotypical humans use to work out others’ mental states, as autistic people have often been claimed to lack.

Scientific publishing does not rigorously stick to a distinction between work that is intended for public consumption, and work that is intended for other scientists. Even if it did, it seems unlikely that the way science is reported in newspapers, and the way scientific conclusions are used in public discourse, politics, and policy would respect that distinction. So long as scientific writings are discussed in nonscientific contexts, there should be some norms that discourage their misuse. It might be that whenever these clash with the norms that best support collective inquiry, collective inquiry must win out. However, this does not preclude norms that do not undermine collective inquiry – for example, carefully considering one’s use of scientific terms with multiple meanings and significant nonscientific uses.

Discussions of pluralism and eliminativism therefore apparently need to account for the nonscientific uses of scientific terms with multiple meanings. This cannot be a matter solely of purely formal properties of such terms. Some such terms exploit the flexibility that they afford in order to facilitate beneficial interaction and integration. Others lead to harmful equivocations and miscommunication. In order even to spot, let alone to mitigate, these problems, one has to consider the broader social systems in which scientific research is embedded.

Acknowledgements With thanks to Andy Clark, Riana Betzler, Alex Raubo, Nathan Davies, Helen Gough, and two anonymous reviewers for their indispensable feedback.

Funding N/A

Availability of data and material N/A

Code Availability N/A

Declarations

Conflicts of interest/Competing interests N/A

Ethical approval N/A

Consent to participate N/A

Consent for publication N/A

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use
References

Akagi, M. (2018). Rethinking the problem of cognition. *Synthese, 195*(8), 3547–3570. doi:https://doi.org/10.1007/s11229-017-1383-2

Akagi, M. (2021). Cognition as the sensitive management of an agent’s behavior. *Philosophical Psychology, 1–24*. doi: https://doi.org/10.1080/09515089.2021.2014802

Alexandrova, A. (2017). *A Philosophy for the Science of Well-Being*. Oxford: Oxford University Press

Apperly, I. A. (2012). What is “theory of mind”? Concepts, cognitive processes, and individual differences. *Quarterly Journal of Experimental Psychology, 65*(5), 825–839

Baron-Cohen, S. (2000a). The Evolution of a Theory of Mind. In M. Corballis, & S. E. G. Lea (Eds.), *The Descent of Mind: Psychological Perspectives on Hominid Evolution* (pp. 261–277). Oxford: Oxford University Press

Baron-Cohen, S. (2000b). Theory of mind and autism: a review. *International Review of Research in Mental Retardation, 23*, 169–184

Baron-Cohen, S. (2009). Autism: the empathizing-systemizing (E-S) theory. *Annals Of The New York Academy Of Sciences, 1156*(1), 68–80. doi:https://doi.org/10.1111/j.1749-6632.2009.04467.x

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the Autistic Child have a “Theory of Mind”? *Cognition, 21*(1), 37–46

Beaudoin, C., Leblanc, É., Gagner, C., & Beauchamp, M. H. (2020). Systematic Review and Inventory of Theory of Mind Measures for Young Children. *Frontiers in Psychology, 10*, 2905–2905. doi: https://doi.org/10.3389/fpsyg.2019.02905

Bermúdez, J. L. (2020). *Cognitive Science: An Introduction to the Science of the Mind* (3 ed.). Cambridge: Cambridge University Press

Bloechle, J., Huber, S., Klein, E., Bahnmueller, J., Moeller, K., & Rennig, J. (2018). Neuro-cognitive mechanisms of global Gestalt perception in visual quantification. *Neuroimage, 181*, 359–369. doi:https://doi.org/10.1016/j.neuroimage.2018.07.026

Bloom, P., & German, T. P. (2000). Two reasons to abandon the false belief task as a test of theory of mind. *Cognition, 77*(1), B25–B31. doi:https://doi.org/10.1016/S0010-0277(00)00096-2

Burge, T. (2018). Do infants and nonhuman animals attribute mental states? *Psychological Review, 125*(3), 409–434

Bursten, J. R. (2018). Smaller than a Breadbox: Scale and Natural Kinds. *The British Journal for the Philosophy of Science, 69*(1), 1–23. doi:https://doi.org/10.1093/bjps/axw022

Caporeal, L. R., & Heyes, C. (1997). Why Anthropomorphize? Folk Psychology and Other Stories. In R. W. Mitchell, N. S. Thompson, & H. L. Miles (Eds.), *Anthropomorphism, Anecdotes, and Animals* (pp. 59–73). New York: SUNY Press

Carpenter, A. (2018). Illuminating Community: Animals in Classical Indian Thought. In P. Adamson, & G. F. Edwards (Eds.), *Animals: A History* (pp. 63–90). Oxford: Oxford University Press

Chang, H. (2012). *Is Water H2O? Evidence, Realism and Pluralism*. Boston Studies in the Philosophy and History of Science

Churchland, P. M. (1981). Eliminativism and the Propositional Attitudes. *Journal of Philosophy, 78*(Feb), 67–90

Churchland, P. M. (1993). Evaluating Our Self Conception. *Mind and Language, 8*(2), 211–222

Dennett, D. C. (1976). Conditions of Personhood. In A. O. Rorty (Ed.), *The Identities of Persons*. Berkeley: University of California Press

Dennett, D. C. (1978). Beliefs about beliefs. *Behavioral and Brain Sciences, 1*(4), 568–570

Douglas, H. (2000). Inductive Risk and Values in Science. *Philosophy of Science, 67*(4), 559–579. Retrieved from http://www.jstor.org/stable/188707

Ereshefsky, M. (1992). Eliminative Pluralism. *Philosophy of Science, 59*(4), 671–690. Retrieved from http://www.jstor.org/stable/188136

Feyerabend, P. (2001). *Conquest of abundance: a tale of abstraction versus the richness of being*. University of Chicago Press
Field, H. (1973). Theory change and the indeterminacy of reference. *Journal of Philosophy*, 70(14), 462–481

Gallie, W. B. (1955). Essentially Contested Concepts. *Proceedings of the Aristotelian Society*, 56, 167–198. Retrieved from http://www.jstor.org/stable/4544562

Gernsbacher, M. A. (2018). Critical review of autism and theory and mind: A technical report. *Open Science Framework*

Gernsbacher, M. A., & Yergeau, M. (2019). Empirical Failures of the Claim That Autistic People Lack a Theory of Mind. *Archives of Scientific Psychology*, 7(1), 102–118

Gómez, J. C. (1991). Visual behaviour as a window for reading the mind of others in primates. In A. Whiten (Ed.), *Natural Theories of Mind: Evolution, Development and Simulation of Everyday Mindreading* (pp. 195–208). Oxford: Blackwell

Grandin, T. (1996). *Emergence: Labeled Autistics*. New York: Grand Central Publishing

Halina, M. (2015). There Is No Special Problem of Mindreading in Nonhuman Animals. *Philosophy of Science*, 82(3), 473–490

Happé, F., & Frith, U. (2020). Annual Research Review: Looking back to look forward–changes in the concept of autism and implications for future research. *Journal of Child Psychology and Psychiatry*, 61(3), 218–232

Hare, B., Call, J., Agetta, B., & Tomasello, M. (2000). Chimpanzees know what conspecifics do and do not see. *Animal Behaviour*, 59(4), 771–785

Haueis, P. (2021a). The death of the cortical column? Patchwork structure and conceptual retirement in neuroscientific practice. *Studies in History and Philosophy of Science Part A*, 85, 101–113

Haueis, P. (2021b). A generalized patchwork approach to scientific concepts. *The British Journal for the Philosophy of Science*. doi:https://doi.org/10.1086/716179

Haueis, P., & Slaby, J. (2022). The humanities as conceptual practices: The formation and development of high-impact concepts in philosophy and beyond. *Metaphilosophy*, n/a(n/a), https://doi.org/10.1111/meta.12551

Heyes, C. (2014). Submentalizing: I Am Not Really Reading Your Mind. *Perspectives on Psychological Science*, 9, 131–143

Heyes, C. (2015). Animal mindreading: what’s the problem? *Psychonomic Bulletin & Review*, 22(2), 313–327. doi: https://doi.org/10.3758/s13420-014-0704-4

Huberle, E., & Karnath, H. O. (2012). The role of temporoparietal junction (TPJ) in global Gestalt perception. *Brain Struct Funct*, 217(3), 735–746. doi:https://doi.org/10.1007/s00429-011-0369-y

Karvelis, P., Seitz, A. R., Lawrie, S. M., & Seriès, P. (2018). Autistic traits, but not schizotypy, predict increased weighting of sensory information in Bayesian visual integration. *eLife*, 7(e34115)

Kidd, I. (2012). Feyerabend, Pseudo-Dionysius, and the Ineffability of Reality. *Philosophia*, 40(2), 365–377

Kosakowski, H. L., & Saxe, R. (2018). “Affective Theory of Mind” and the Function of the Ventral Medial Prefrontal Cortex. *Cognitive and Behavioral Neurology*, 31(1), 36–37

Krauss, A. (2016). The scientific limits of understanding the (potential) relationship between complex social phenomena: the case of democracy and inequality. *Journal of Economic Methodology*, 23(1), 97–109. doi:https://doi.org/10.1080/1350178X.2015.1069372

Lakoff, G. (1987). *Women, Fire and Dangerous Thing: What Categories Reveal About the Mind* (22 vol.). University of Chicago Press

Longino, H. E. (1990). *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry* (25 vol.). Princeton University Press

Lurz, R. (2011). *Mindreading Animals: The Debate Over What Animals Know About Other Minds*. Cambridge, MA: MIT Press

Martin, R., & Barresi, J. (2006). *The rise and fall of soul and self: An intellectual history of personal identity*. Columbia University Press

Nagel, E. (1961). *The Structure of Science: Problems in the Logic of Scientific Explanation*. Harcourt, Brace & World

Neto, C. (2020). When imprecision is a good thing, or how imprecise concepts facilitate integration in biology. *Biography and Philosophy*, 35(6), 1–21

Nilsson, D. E., & Colley, N. J. (2016). Comparative Vision: Can Bacteria Really See? *Current Biology*, 26(9), R369–371. doi:https://doi.org/10.1016/j.cub.2016.03.025

Ogawa, S., Iriguchi, M., Lee, Y. A., Yoshikawa, S., & Goto, Y. (2019). Atypical Social Rank Recognition in Autism Spectrum Disorder. *Scientific Reports*, 9(1), 15657. doi:https://doi.org/10.1038/s41598-019-52211-8
Peirce, C. S. (1878). How To Make Our Ideas Clear. Popular Science Monthly, 12(Jan.), 286–302
Penn, D. C., Holyoak, K. J., & Povinelli, D. J. (2008). Darwin's Mistake: Explaining the Discontinuity between Human and Nonhuman Minds. Behavioral and Brain Sciences, 31(2), 109–130
Penn, D. C., & Povinelli, D. J. (2007). On the Lack of Evidence that Non-Human Animals Possess Anything Remotely Resembling a ‘Theory of Mind’. Phil Trans R Soc B, 362, 731–744
Povinelli, D. J. (2004). Behind the Ape’s Appearance: Escaping Anthropocentrism in the Study of Other Minds. Daedalus, 133(1), 29–41
Povinelli, D. J., & Vonk, J. (2003). Chimpanzee minds: suspiciously human? Trends In Cognitive Sciences, 7(4), 157–160. doi:https://doi.org/10.1016/s1364-6613(03)00053-6
Povinelli, D. J., & Vonk, J. (2004). We Don’t Need a Microscope to Explore the Chimpanzee’s Mind. Mind & Language, 19(1), 1–28
Premack, D., & Woodruff, G. (1978). Does the Chimpanzee have a Theory of Mind? The Behavioural and Brain Sciences, 4, 515–526
Pritchard, T. (2019). Analogical Cognition: an Insight into Word Meaning. Review of Philosophy and Psychology, 10(3), 587–607. doi:https://doi.org/10.1007/s13164-018-0419-y
Putnam, H. (1982). Beyond the Fact–Value Dichotomy. Critica, 41(41), 3–12
Putnam, H. (1984). The Craving for Objectivity. New Literary History, 15(2), 229–239
Quesque, F., & Rossetti, Y. (2020). What do theory-of-mind tasks actually measure? Theory and practice. Perspectives on Psychological Science, 15(2), 384–396
Recanati, F. (2017). Contextualism and Polysemy. Dialectica, 71(3), 379–397. doi:https://doi.org/10.1111/1746-8361.12179
Rennig, J., Bilalic, M., Huberle, E., Karnath, H. O., & Himmelbach, M. (2013). The temporo-parietal junction contributes to global gestalt perception-evidence from studies in chess experts. Frontiers In Human Neuroscience, 7, 513. doi:https://doi.org/10.3389/fnhum.2013.00513
Robertson, C. E., & Baron-Cohen, S. (2017). Sensory perception in autism. Nature Reviews Neuroscience, 18(11), 671–684. doi:https://doi.org/10.1038/nrn.2017.112
Rouse, J. (2007). Social Practices and Normativity. Philosophy of the Social Sciences, 37(1), 46–56. doi:https://doi.org/10.1177/0048393106296542
Ruphy, S. (2010). Are Stellar Kinds Natural Kinds? A Challenging Newcomer in the Monism/Pluralism and Realism/Antirealism Debates. Philosophy of Science, 77(5), 1109–1120. doi:https://doi.org/10.1086/656544
Ruphy, S. (2017). Scientific pluralism reconsidered: A new approach to the (dis) unity of science. University of Pittsburgh Press
Ruttenberg, K., & Mets, A. (2020). Chemistry is pluralistic. Foundations of Chemistry, 22(3), 403–419. doi:https://doi.org/10.1007/s10698-020-09378-0
Sacks, O. W. (2012). An Anthropologist on Mars. London: Picador
Sawyer, S. (2018). The Importance of Concepts. Proceedings of the Aristotelian Society, 118(2), 127–147
Saxe, R. (2009). Theory of Mind (Neural Basis). In W. P. Banks (Ed.), Handbook of Theory of Mind, 401–409). Cambridge, MA: Elsevier: Academic Press
Saxe, R. (2010). The right temporo-parietal junction: a specific brain region for thinking about thoughts. In Handbook of Theory of Mind (pp. 1–35)
Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people: The role of the temporo-parietal junction in “theory of mind”. Neuroimage, 19(4), 1835–1842
Saxe, R., & Powell, L. J. (2006). It’s the Thought That Counts: Specific Brain Regions for One Component of Theory of Mind. Psychological Science, 17(8), 692–699
Schaafsma, S. M., Pfaff, D. W., Spunt, R. P., & Adolphs, R. (2015). Deconstructing and Reconstructing Theory of Mind. Trends in Cognitive Science, 19(2), 65–72
Scholz, J., Triantafyllou, C., Whitfield-Gabrieli, S., Brown, E. N., & Saxe, R. (2009). Distinct regions of right temporo-parietal junction are selective for theory of mind and exogenous attention. PLoS One, 4(3), e4869. doi:https://doi.org/10.1371/journal.pone.0004869
Schuermers, N., Lenn, T., Kampmann, R., Meissner, M. V., Estevés, T., Temerinac-Ott, M., & Wilde, A. (2016). Cyanobacteria use micro-optics to sense light direction. eLife, 5. doi:https://doi.org/10.7554/eLife.12620
Schuurk, T., Schurz, M., Muller, F., Rupprecht, R., & Sommer, M. (2017). The rTPJ’s overarching cognitive function in networks for attention and theory of mind. Soc Cogn Affect Neurosci, 12(1), 157–168. doi:https://doi.org/10.1093/sca/nsx163
Slater, M. H. (2009). Macromolecular Pluralism. Philosophy of Science, 76(5), 851–863. doi:https://doi.org/10.1086/605817
Soper, H. V., & Murray, M. O. (2012). Autism. In C. A. D. Noggle, & R. S. Horton, A M (Eds.), *The Encyclopedia of Neuropsychological Disorders* (pp. 125–128). New York, NY: Springer.

Sorabji, R. (1993). *Animal Minds and Human Morals: The Origins of the Western Debate*. London: Duckworth.

Taylor, H., & Vickers, P. (2017). Conceptual Fragmentation and the Rise of Eliminativism. *European Journal for Philosophy of Science, 7*(1), 17–40.

Tomasello, M., Call, J., & Hare, B. (2003a). Chimpanzees understand psychological states—the question is which ones and to what extent. *Trends in Cognitive Science, 7*, 153–156.

Tomasello, M., Call, J., & Hare, B. (2003b). Chimpanzees versus humans: it’s not that simple. *Trends in Cognitive Science, 7*, 239–240.

Wellman, H. M., & Liu, D. (2004). Scaling of theory-of-mind tasks. *Child Development, 75*(2), 523–541. doi:https://doi.org/10.1111/j.1467-8624.2004.00691.x

Whiten, A. (1996). When does smart behaviour-reading become mind-reading?. In P. Carruthers, & P. K. Smith (Eds.), *Theories of Theories of Mind* (pp. 277–292). Cambridge: Cambridge University Press.

Whiten, A., & Suddendorf, T. (2001). Meta-representation and secondary representation. *Trends in Cognitive Science, 5*, 378.

Wilson, M. (2006). *Wandering Significance: An Essay on Conceptual Behavior* (57 vol.). Oxford: Clarendon Press.

Yergeau, M., & Huebner, B. (2017). Minding Theory of Mind. *Journal of Social Philosophy, 48*(3), 273–296.

**Publisher’s Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.