Review of Philosophy and Psychology
https://doi.org/10.1007/s13164-021-00584-z

An Externalist Theory of Social Understanding: Interaction, Psychological Models, and the Frame Problem

Axel Seemann [1]

Accepted: 30 August 2021/Published online: 06 October 2021

Abstract

I put forward an externalist theory of social understanding. On this view, psychological sense making takes place in environments that contain both agent and interpreter. The spatial structure of such environments is social, in the sense that its occupants locate its objects by an exercise in triangulation relative to each of their standpoints. This triangulation is achieved in intersubjective interaction and gives rise to a triadic model of the social mind. This model can then be used to make sense of others' observed actions. Its possession plays a vital role in the development of the capacity for false belief reasoning. The view offers an integrated account of the development of social cognition from primary intersubjectivity to level-2 perspective taking. It incorporates insights from interactionism and mindreading theories of social cognition and thus offers a way out of the stalemate between defenders of the two views. Because psychological sense making is perspectival, the frame problem does not arise for social reasoners: the perspective they bring to bear on the action that is to be interpreted constrains the information they can select to make sense of what others do.

1 Introduction

Creatures like us operate in an environment that contains both perceptual objects and perceivers and agents. We make use of our perceptual faculties to acquire knowledge about ordinary objects and to glean insight into fellow subjects' mental lives. One

✉ Axel Seemann
aseemann@bentley.edu

[1] Department of Philosophy, Bentley University, 175 Forest Street, Waltham, MA 02452, USA

family of theories of social cognition treats bodies as origins of behaviour, while mentality is only inferentially accessible.[1] The rival view is that mental life is directly revealed in the perception of action and that inferences are only exceptionally required to make sense of what an agent is doing (e.g., De Jaegher, Di Paolo, & Gallagher, 2010; Gallagher, 2008; Hutto, 2011). It is widely thought that the first family of views gives rise to the frame problem (Zawidzki, 2013, pp. 74–82): if mentality is not directly perceivable in behaviour and if further information is thus required to infer the mental state of the agent, how can the social cognizer be in a position to select amongst potentially unlimited amounts of information she could bring to bear on the interpretation of observed behaviour? Defenders of the second view may, at first glance, appear to be better positioned here: if mind can be directly perceived and experienced in action and interaction, there should be no special frame problem for social cognition. However, it is not self-evident that mentality is directly perceivable in interaction (e.g., Michael, 2011; Schönherr, 2017); and it is not obvious how the second family of views copes with cases in which, on anyone's account, inferential work is required to acquire insight into an agent's mental life.

The question of how humans come to know about others' mental lives through perception has an important ontogenetic aspect: all participants in the debate agree that we can learn much about social cognition by thinking about its development in humans. But the two camps tend to pay attention to different stages of human development: for defenders of Theory of Mind, the all-important finding is that children can solve explicit false belief tasks only around their fourth birthday (e.g., Wimmer & Perner, 1983).[2] For defenders of the view that others' mental lives can be directly experienced in interaction, the focus is on primary and secondary intersubjectivity – the interactions between infants and their caregivers that begin with birth and lead to the shared focus on third objects in joint attention at the end of the first year of life (e.g., Hobson, 2002/2004; Reddy, 2008). By and large, supporters of Theory of Mind argue that mindreading is at the heart of the human capacity for social understanding, while interactionists claim that mindreading occurs only in exceptional circumstances and that embodied interaction is our primary way of understanding others.[3] Neither approach delivers an integrated account of how humans progress from social interaction to observational social cognition and thus from beings who rely on the felt aspect of human experience to creatures who draw on their rational capacities to acquire insight into others' mental lives. But it would be surprising if these two dimensions of psychological sense making were not integrated and mutually supporting. Ordinary human

[1] Within the group of supporters of mindreading approaches to social cognition, defenders of the Theory Theory hold that social cognizers make use of a theory of how mental states inform behaviour. Mental state concepts may be innate (e.g., Leslie, Friedman, & German, 2004) or developmentally acquired (e.g., Gopnik & Meltzoff, 1997). Defenders of the Simulation Theory maintain that we ascribe mental states to others by imaginatively putting ourselves in the other's shoes (e.g., Goldman, 2006). Hybrid views are possible. This is well-trodden terrain and it is not necessary to rehash the debate for present purposes.

[2] For the purposes of this paper, I bracket a group of findings that ascribes the ability to pass some implicit false belief tests to children as young as fifteen months of age (e.g., Onishi & Baillargeon, 2005; Southgate, 2010). My reason for doing so is that some of the relevant findings have recently failed to replicate (Kulke, Reiss, Krist, & Rakoczy, 2018; Kulke, Von Duhn, Schneider, & Rakoczy, 2018) and that, consequently, it is currently controversial whether infant mentalising is a real phenomenon (Kulke, Johannsen, & Rakoczy, 2019).

[3] See Spaulding (2010) for an overview of the respective positions and an argument against the interactionist contention that mindreading occurs only in exceptional circumstances.

perceivers and agents draw on both the felt aspect of human interaction and psychological reasoning to glean insight into the mental lives of others,[4] and sufferers from autism, whose ability to relate to the emotional life of others through interaction is impaired, also struggle with mental state ascription (see e.g. the chapters in Baron-Cohen, Tager-Flusberg, and Cohen (1999)). And, as I hope to show in what follows, the developmental milestones of intersubjective interaction, perspective-taking, and mentalizing are importantly related: their temporal succession in human ontogeny is not an accident but a result of their developmental dependency.

The aim of this paper is to put forward an account of social understanding that integrates key insights from Theory of Mind and interactionism. Such an account, I hope to show, can explain why creatures like us do not face the frame problem. The core view is that psychological sense making takes place in an environment that contains ordinary objects and other perceivers and agents from the beginning of human life. This fact is reflected in its socio-spatial structure: we begin from birth to interact with other perceivers and agents who occupy locations affording distinct perspectives on objects that are thus in public view. Spatial awareness plays a vital and, on the present approach, insufficiently acknowledged role in human socio-cognitive development: the kinds of interactions with caregivers that children undertake in their first year of life lead to the ability to triangulate the location of objects relative to caregivers' standpoints, which facilitates joint attention. And mindreading tasks largely investigate children's knowledge of an observed agent's spatial knowledge: the question children face, in various forms, is typically whether they know that an actor's belief about the location of an object is distinct from their own.[5] The account I shall develop builds on this observation to argue that the socio-spatial structure of humans' perceptual environment plays two cognitive roles: it provides infants who can jointly attend to objects with their caregivers with a practical kind of knowledge of object location in social space, and it subsequently equips young children with a mental model of the triadic relation that produces this knowledge. The social cognizer is a perspective-taker who is operating with a triadic model in which the standpoint of the observed agent forms one constituent. It thus is a relational model of the mind in its social environment, and its relational character explains why social reasoners do not face the frame problem: the mental life of the observed agent, in the context of what she is doing, is always being made sense of relative to the interpreter's own perspective, and this perspective determines the information that can be brought to bear on the interpretation of what another is doing. In developing this view, which I call "social externalism", I integrate the various milestones considered by defenders of the rival approaches of social cognition – primary and secondary intersubjectivity, false belief reasoning, and level-1 and level-2 perspective taking.

The overarching strategy is to explain as many of these milestones as possible by appeal to the structure of the environment in which developing social cognizers operate. It is aimed at parsimony: the fewer, and less demanding, the kinds of mental states whose ascription a theory has to postulate in order to explain how we come to understand what others do the less intractable the frame problem will be. The view I arrive at is that children begin to ascribe the most primitive mental state concept, that of perceptual knowledge of

[4] Empathy is often thought to play a key role in the felt understanding of others. See Stueber (2006) for a discussion of both its felt aspect and its role in folk psychological reasoning.

[5] Though it should be noted that not all tests of young children's psychological reasoning capacities are spatial: see for instance the colour filter tests of Moll and Meltzoff (2011).

object location, to inhabitants of social space when they can consider questions about others' knowledge of object location in their fourth year, prior to passing classical false belief tests. But the ontogenetic primacy of social interaction is not taken to entail that mindreading has no significant role to play in mature reasoners' interpretation of what agents do. I take no view on the proportion of interaction and mindreading in fully developed social sense making. Also, there is no claim that infants and children prior to their fourth year do not sometimes deploy mental state concepts in their attempts to make sense of what others do. It is just that, on the externalist approach I shall develop, a range of developmental milestones up to the age of four can be accounted for without appeal to mindreading capacities; and the kind of spatial reasoning that is required to pass classical false belief tests and solve level-2 perspective taking tasks can be explained in terms of a model of the socio-spatial relation that underwrites joint attention.

The paper proceeds in four steps: I provide an externalist account of the development of social space (I). I explain how the externalist account, combined with psychological model theory, can help explain children's development from joint attention to passing false belief tests, as well as their ability to solve level-1 and level-2 perspective taking tasks (II). I explain how the externalist account avoids the frame problem (III). I end with some brief concluding remarks (IV).

(I) The Development of Social Space

In this section I offer an account of the development of humans' acquisition of a social spatial framework in primary and secondary intersubjectivity. This account makes two key moves. First, it introduces the notion of a "doing" as a description of purposive bodily activity that gets by without positing intentions and then argues that we begin to understand others as doers by interacting with them, first face-to-face and slightly later jointly on third objects. Secondly, it puts forward an externalist theory of the social mind, according to which social understanding begins in joint attention with the acquisition of a spatial frame of reference of an environment in which a variety of locations are occupied by doers whose practical knowledge of the location of the objects of their doings is individuated relative to that environment.

a) *"Doings" As Purposive Activities*

Supporters of mindreading and interactionist approaches to social cognition have quite different views on how to characterise the purposiveness of human activity. Theory of Mind takes it that understanding what others do begins with the ascription of belief-desire pairs (e.g., Leslie, 2000). Interactivists appeal to "motor acts" to account for the facial imitation with which social interaction begins in infancy (e.g., Gallagher, 2005, p. 77). I introduce the technical concept of a "doing" in order to describe purposive object-directed movement while avoiding having to take a view, at the outset of the inquiry, on the nature and role of intentions in human activity:

(DOING) A doing is a proprioceived[6] bodily movement that is directed at a perceptually present object and that the moving creature prolongs.

[6] In what follows, I use the verb "to proprioceive" as a shorthand for "to apprehend by means of proprioception". Thanks go to one of my reviewers for highlighting the need for clarification.

That creatures "prolong" what they are doing is meant to make the notion compatible with a variety of motivations they might have for their purposeful movements. Creatures can prolong what they are doing because they are enjoying the activity or because they are pursuing some external goal, but they can also keep doing what they are doing if they have no apparent reason for doing so at all (think of the doodles you draw while on the phone with someone). Since everything we do eventually comes to an end, the notion of a doing is temporally indexed: it is only within a certain temporal interval that creatures prolong their doings. You thus can be doing something, in my sense of the term, even though what you are doing will end once you have achieved an external goal, or once you don't find it pleasurable anymore, or once it is terminated by external factors (you have to do something else; you fall asleep). You still prolong the doing as long as it is going on. I call this the "intrinsic motivation" that is inherent to a doing. Intrinsic motivations are unlike distal intentions in that they could not be entertained outside of the doing. They are also different from what Searle (1983) calls "intentions-in-action" in that the prolongation of the doing does not require internally represented conditions of satisfaction: it is not that creatures intend to prolong their doings and can succeed or fail in doing so. If they are doing something, they are necessarily prolonging what they are doing within the doing's temporal boundaries. However, all this is compatible with the possibility that the doer might entertain intentions, distal or not, and that these intentions have a causal or explanatory role in the doing. But the notion of a doing is compatible also with the view that intentions play no role in (some of) the things we do and that there nevertheless is a distinction to be drawn between doings and reflex-like bodily movement, such as the twitch of your knee upon the doctor's probing touch.[7]

Doings constitutively involve perceptual objects. The directedness of doings at perceptual objects is designed to accommodate cases in which a perceived object is touched, moved, or otherwise manipulated, as well as cases in which a perceived object is being responded to without being touched. Doings thus comprise both object-directed actions and imitative activity. Infants' repeated manipulations of physical objects, such as the shaking of a rattle or the beating of a toy drum, are doings; so are the movements by means of which they engage with their caregivers in the kinds of social interactions that have come to be called "primary intersubjectivity" (e.g., Trevarthen, 2011).

Creatures who can execute doings[8] enjoy a mental life that is shaped by both their perceptual environments and their bodies. The notion of a doing stresses the bodily dimension of human activity: in order to purposefully move, we need to know how to direct our bodies in our perceptual environment. No particular view is taken, at the outset, on how bodily and environmental information is integrated in such a creature's experience or motor system. Further, the account does not stipulate any kind of self-awareness or (in social doings) the ability to distinguish between "self" and "other", and correspondingly it remains neutral on the question of whether infants can distinguish, conceptually or practically, between social interaction and object-directed

[7] Dreyfus's (1993/2014) Heideggerian notion of "coping" is one version of such a view. Hutto's (2011) enactivism is another example of a way of thinking about human activity without appeal to mental state concepts.

[8] I call such creatures "doers" or sometimes "agents", without thereby meaning to imply that their bodily movements can necessarily be explained by appeal to mental state concepts.

activity. Ascribing doings to children in their first year of life, then, commits you only to the view that they are capable of prolonged object-directed activities in which they draw on proprioception and perception, and that some of these are of a social kind. This view should be compatible with a range of views on the nature of social cognition.[9] In what follows, I build on the notion of a doing to explain how humans, through social interaction, come to acquire a social spatial framework.

b) *Doings in Egocentric and in Social Space*

In this section I argue that depending on whether we manipulate ordinary objects or interact with others, humans use proprioception in different ways. This difference is vital for the present argument. Begin by considering the respective functions of proprioception and perception. Perception informs you about the external world (which may include your own body), proprioception or synesthesia informs you about your bodily movements and the position of your limbs relative to each other. For a creature to be capable of doings, it has to draw on both, so that it knows how it has to move its body in order to act on a perceived object. Suppose a creature has two limbs, akin to human arms, with which it can act on objects in its environment, and that these limbs are attached to opposite sides of its trunk. If the object is presented to the creature as being on its left, the creature has to know that it has to move its left limb. A creature who is capable of executing doings is thus operating in a spatial order in which proprioception and perception are integrated. This spatial order is egocentric: the objects at which its doings are directed are presented to the creature relative to its own location. They could be described by the creature as being to its left or right, above or below it (though the creature need not itself be capable of giving such a description).

Not all egocentrically ordered spaces are action spaces: the sun might be described as positioned above you, but you cannot act on it. A full account of action space needs to take into account the distance between the agent's body and the object of perception and action. The notion of "peripersonal space" captures this consideration. It describes the area around the body within which action on its objects is possible.[10] For purposes of the present discussion, I do not sharply distinguish between egocentric and peripersonal space. The point I shall be developing is that in social interaction, the multimodal integration of proprioception and perception is not restricted to an egocentre that is constituted by the location of the acting creature's body. Social interactions, so the view, play out in a specific spatial order that I call "social space". In this spatial order proprioception is integrated with perception at locations other than the one occupied by the perceiver's body, so that the agent acquires a bodily kind of access to another's movement in social interaction. As is the case with other action-

[9] One way to understand the notion of a doing is by drawing on a conclusion of Borg's (2007) in her discussion of the role of mirror neurons in intention attribution. She suggests that the discovery of mirror neurons can tell us something about which creatures an observer is prepared to treat as minded (and, you might add, which kinds of movements as purposive activities) even though it falls short of explaining how we come to know about the large-scale intentions with which they move. Doings can be thought of as purposive activities whose perception presents the executing creature as minded, irrespective of whether large-scale intention attribution is taking place.

[10] In a seminal article, Rizzolatti, Fadiga, Fogassi, and Gallese (1997) argued that there is a specific kind of spatial format that is represented in the brain as a spatial map that is vital for the control of motor movement. For a recent discussion of the relation between egocentric space and action space, see De Vignemont (2018).

relevant spatial orders,[11] social space is not freestanding: there is no creature that would at any one time only be operating in social space, and making sense of it requires thinking of it as a spatial arrangement relied upon by doers who simultaneously also move in egocentric and peripersonal space. I begin by motivating the need for the notion of social space and then introduce the concept itself.

One perennial puzzle for theories of bodily forms of social cognition arises from studies of infants' interactions with their caregivers. They imitate facial expressions more or less from birth, fixate on caregivers' eyes at two months of age, call for them with shrill vocalizations around two to four months and actively seek repetitions of tickling around four to five months (Reddy, 2011, p. 147). They take delight and seek social contact in cooperative interactions in which mutual gaze and the rhythmic synchronization of bodily expression and activity play a vital role. The increasing range of person-to-person interactions that infants begin to participate in during the first year of life has come to be called "primary intersubjectivity" (e.g., Trevarthen, 2011, p. 86). The question is how the infant deploys her bodily resources in these developmentally early interactions. One possibility is that the infant perceives others with whom she interacts in the same way as she does ordinary objects. Then there is no special use of proprioception in intersubjective interaction: the use of proprioception is restricted to the execution of her own bodily movements and thus to the area occupied by her own body. Proprioception in social interaction then stops at the body's boundaries. This view faces a significant problem with regard to the explanation of facial imitation. The problem as it arises for neonate imitation is this: how can a perceived event (the caregiver producing a facial expression) guide the infant, who can have little or no perceptual awareness of her own body, towards producing a facial expression of the same kind? One possibility is to explain the infant's activity by parsing the imitative event into segments according to their mode of presentation: the infant sees the caregiver's facial expression and matches it by means of proprioceptively presented facial movements of her own. The difficulty is that this description gives rise to a version of Molyneux's problem[12]: given that the imitating infant can have little or no perceptual awareness of her own facial features, you now have to explain how there can be an intermodal "translation" from perception to proprioception.[13] This problem is not trivial: the infant's imitative activity persists over time and throughout obstacles (the child resumes her activity after a dummy has been placed in and subsequently removed from her mouth, for instance (Meltzoff & Moore, 1977)), so you cannot easily appeal to purely reflexive motor processes in your explanation of the child's matching of another's facial expression. But if she is intentionally trying to match a perceived facial expression in proprioception, she must be able to distinguish between better and worse matches between perceived and proprioceived expressions. For that to be possible, she has to have some means of comparison. One way to account for it is by appeal to a

[11] In particular, conceptualising peripersonal space is dependent on the notion of an egocentre occupied by the agent's body.

[12] In its original form, this is the problem of whether a person born blind could immediately identify an object they have known only by touch if they came to see it. For an account of how it pertains to facial imitation, see Meltzoff (1993).

[13] Children begin to pass the "mirror test" (or the closely related "rouge test"), in which their ability to recognise themselves in a mirror is tested, around 18 months (Archer, 1992). It should be noted that the mirror test is subject to considerable criticism.

"supramodal framework" of abstract geometrical and temporal patterns into which expressions in different sense modalities are translated (Meltzoff & Moore, 1977). But this proposal is not unproblematic, since the translation is itself subject to a matching problem: mistakes are always possible and a means of comparison is now needed between the facial expression and the abstract pattern.[14] A vicious regress threatens.

The matching problem has a solution if you suppose that human visual and motor systems are already intermodal. As Gallagher (2005, p. 80) puts it, "The concept of an intermodal code means that the visual and motor systems speak the same 'language' right from birth." Then there is no need for a system that can translate from perception to proprioception and back; the matching problem does not arise.[15] One way to give substance to this view is in terms of the notion of social space. For creatures like us, proprioceptive information about the position and movement of our limbs is always integrated with sensory information from the area in which our bodies are moving. You can distinguish between two kinds of sensory information here. There is, first, what you may call "proximal" sensory information that can only provide you with knowledge about how things are in your immediate environment, such as the temperature of the room you are in or the texture of the clothes you are wearing. Then there is sensory information that you might call "distal". This kind of information can provide you with knowledge about how things are close by but also, crucially, further away. For instance, you can hear a loud noise from across the street or see a plane in the sky. The sensory information that is integrated with proprioception to support awareness of your limbs' relative position and movement is primarily of a proximal kind. For instance, it is in parts because you sense the pressure of your upper on your lower leg that you know they are crossed. This kind of integration can only occur at the location of your own body: there could not be a binding of proprioceptive information from your body with sensory information about the pressure of someone else's upper on their lower leg. After all, you simply cannot have the required sensory- and the other person cannot have the required proprioceptive information.

But there is also integration of proprioception with distal sensory and in particular with distal visual information.[16] We can see our own bodies in the environment and this perception is integrated with our internal sense of bodily movement also. When you see your fingers typing on the keyboard, they are presented to you as fingers whose movement you can also proprioceive. This integration is less tight than its proximal counterpart: for example, you can be tricked into ascribing wiggling fingers on a video screen to yourself if you make a wiggling movement, even if you see the fingers at a rotation of 180 degrees to your fingers' actual position (Van den Bos & Jeannerod, 2002). Consider also the virtual reality experiments by Lenggenhager et al. (2007), in

[14] To see this, compare the situation of the imitating infant to that of the Davidsonian radical interpreter (Davidson, 1973): the interpreter can always point out the features of the environment that are visible to both the speaker and the interpreter in order to ascertain the accuracy of the interpretation at issue. Even though there is no knowing whether speaker and interpreter conceptualise the demonstrated scene in the same way, there is still some kind of means of comparison that enables the interpreter to assess the accuracy of his translation, which is what the imitating infant lacks.

[15] Gallagher (2005, pp. 65-85), Meltzoff (1993); Meltzoff and Moore (1995) offer evidence in support of this idea.

[16] See De Vignemont (2018) for a discussion of the role of action space in the visual experience of nearby and distal objects.

which subjects saw their own backs in a virtual room. If they experienced their backs being stroked while seeing their backs being stroked synchronously in the virtual room, they were likely to identify the figure in the virtual room as themselves, thus mapping their somatosensorily experienced body onto a location it did not occupy. There is thus some evidence that, under certain conditions, our somatosensory systems can integrate proprioception and perception at places not occupied by our bodies. That the spatial alignment of touch and sight can come apart and that it is primarily proprioception that informs visual placement of tactile objects is not a novel idea either: it was already considered by Stratton (1899) in the context of his experiments with inverting lenses.[17]

The suggestion now is that social creatures can integrate proprioceptive with visual information from another's body in a way that is not erroneous: in social space, we integrate proprioceptive information from our own bodies with visual information about the movements of others so that these movements are presented to us in a multimodal format that, though not the same as in our own case (no integration of proximal sensory information from another's body with proprioception), still provides us with a non-observational, direct kind of bodily awareness of others' doings. In social interaction, proprioception is integrated with distal sensory information at places not occupied by our bodies. For this to be possible, such integration must *also* occur at the location occupied by ourselves: it could not be that you *only* proprioceive another's movements that you also see. Rather, your somatosensory system binds proprioception with perceptual information both at the location occupied by you and at the location occupied by the partner in interaction, and so the other's movement is presented to you as a doing. In social space, you have direct multisensory access to the doings of your co-actor.[18]

The hypothesis of social space has the resources to address the matching problem. Because bodily movement in social space is multimodally presented both in one's own case and that of the partner in interaction, no translation is necessary from vision into proprioception or vice versa. Consider again the case of seeing fingers typing on a keyboard (perhaps they are covered with gloves so they are not immediately recognizable to you as your own). Suppose you deploy proprioception in understanding the fingers' movements. Then, when you execute a similar movement, you can draw on your proprioceptive repertoire to achieve a movement similar to that of the observed fingers without running into the matching problem. The hypothesis of social space proposes that in intersubjective interaction, the partner's bodily and in particular their facial expression is apprehended in the same way as the fingers' movements. Hence, by analogy, the facial imitator does not have to solve the matching problem either.

You may wonder, as a reviewer of this paper did, what distinguishes the hypothesis of social space from Goldman's (2006, p. 113f.) notion of "low-level simulation".

[17] For discussion see Briscoe and Grush (2020).

[18] Some recent work offers empirical evidence in support of the existence of a social spatial framework. Costantini and Sinigaglia (2011) find that the perception of objects' affordances for action is modulated not just by the spatial relation between perceiver and object but also by the relation between a potential co-actor and the object. The notion of social space is further supported by a study by Maister, Cardini, Zamariola, Serino, and Tsakiris (2015), who find that shared experience of the enfacement illusion results in a remapping of the representation of the other's peripersonal space as one's own; and, less directly, by a study by Soliman, Ferguson, Dexheimer, and Glenberg (2015), who suggest that the shared manipulation of an object with others builds on a joint body schema. See Seemann (2019) for a detailed discussion.

Simulation theory proposes, very roughly, that creatures come to understand others' mental lives by imaginatively putting themselves into these others' shoes. Low-level simulation is a non-conscious form of bodily mirroring in which the perceiver takes up the other's bodily movement and thereby comes to know about her intentions. The precise details of the proposal that would be needed for a detailed discussion of its relation to social space are not easy to pin to pin down,[19] but the question brings out nicely the difference between the current proposal and simulation-theoretic approaches to mindreading. Simulation Theory, including its low-level variant, is designed to address what is often called the epistemological aspect of the problem of Other Minds (Avramides, 2001): it proposes a mechanism that enables social creatures to come to know what mental states and in particular what intentions drive another observed creature's bodily movements. The neural mechanisms that underwrite low-level simulation can then be specified in terms of the activity of mirror systems that are responsive both to a creature's executed movements and another's perceived movements (Gallese & Goldman, 1998). The thesis of social space, by contrast, is not designed as a solution to the epistemological problem: because it relies on the notion of a doing and because the description of certain bodily movements as doings gets by without the stipulation of intentions, the view that movements can be presented as doings to creatures at locations not occupied by those creatures contributes, by itself, nothing to the question of how the creature can know what intentions drive these movements. It only helps explain how a particular bodily movement can be presented to a creature as a doing and its executor as minded. Along the lines of the thesis of social space, a creature is presented as having a mental life to a perceiver when its movements are apprehended by the perceiver in a multimodal format that includes proprioception. In social interaction, bodily movements that are executed at the location occupied by the apprehending subject and movements executed at the location of its partner in the interaction are presented to the subject in a multimodal format that includes proprioception and hence as a doing. For this to be possible, it is not necessary that there be intentions motivating the creatures' doings; or that, if there are such intentions, the apprehending creature recognize or ascribe them to herself or others. However, there is nothing in the account that denies the possibility of such recognition or ascription: social externalism gets by without taking a view on the question. This theory is not obviously incompatible with the possibility that simulative processes play a role in understanding others' mental lives, but neither does it necessarily support the view that low-level simulation allows a perceiver to come to know about the intentions that drive another's perceived movements.

Social space begins with neonate facial imitation as the most primitive form of a social doing and thus precedes even the earliest instances of apparent examples of implicit mindreading, so the view resists sceptics (Schönherr, 2017) about the primacy of interaction over mindreading in social cognition. However, no argument has been put forward so far in support of the view that mindreading is based on social interaction. I only have argued that such interactions play out in an environment whose

[19] See De Vignemont (2009) for a discussion of the difficulties involved in making the notion of low-level simulation precise.

occupants are intermodally presented to the infant who begins to imitate caregivers' facial expressions in primary intersubjectivity.[20]

iii) *Joint Attention and the Social Spatial Framework*

The upshot from the previous section is that interactive doings are distinct from their non-interactive counterparts with regard to the agent's use of proprioception in apprehending their objects. Creatures participating in primary intersubjective interactions operate in a spatial format in which proprioception is used to apprehend movements both at the location they occupy themselves and at the location of their partner in the interaction. I now build on this view to explain the vital role of joint attention in the development of perspective-taking.[21]

Towards the end of the first year of life, children begin to follow caregivers' gaze and pointing gestures towards third objects. Many species are capable of gaze-following, but only very few can make use of it in order to achieve joint attention.[22] One crucial difference between gaze following and joint attention is that the latter is interactive. It unfolds over time and requires perceivers' continuous and mutually guided shift of attention between a third object and the co-perceiver. The one-year-old child who is beginning to jointly attend to a third object with a caregiver is not merely following the caregiver's gaze so that her focus subsequently comes to rest on the thing. She remains involved in a prolonged process. There is a temptation to think of the child's role in early joint attention as a spectator who lets her gaze be directed by the caregiver. But one-year old infants begin to hold up objects for show and point at distal objects (Reddy, 2011, p. 147). Joint attention, along these lines, is a form of object-involving social activity. Within an episode of joint attention, the child thus carries out the two kinds of doings introduced earlier. She interacts with others and encounters objects with which she deictically engages. The child is exposed to a contrast between two distinct ways of relating to her environment and two kinds of occupants –co-operators and passive objects - of that environment.

Campbell (2005, 2011) characterizes joint attention as a perceptual relation with three constituents. This description does justice to the idea that its subjects encounter third objects together and thus acquire perceptual knowledge about them that is "out in the open". But there is something mysterious about the notion of a triadic perceptual relation.[23] All perception is from somewhere and all perceptual knowledge is

[20] A recent much-discussed study finds, contrary to orthodoxy about social imitation in infancy, that infants do not imitate others' facial expressions: they react with tongue protrusions also when presented with adults' happy faces or finger pointing (Oostenbroek et al., 2016). Note that these findings, if vindicated, do not pose a problem for the current proposal. The thesis of an intermodal social space does not require that neonates imitate facial gestures; it requires only that they deploy perception and proprioception in the intersubjective process.

[21] The crucial role of joint attention in the development of social cognition is prominently stressed by the Shared Intentionality Hypothesis, as defended by Tomasello and his colleagues (e.g., Tomasello & Carpenter, 2007; Tomasello & Moll, 2010). There are clear parallels between this body of work and the present proposal in various respects, amongst them the stress on joint attention as a key event in human sociocognitive development. One important difference between the two approaches is that social externalism attempts to explain the most basic forms of social understanding without having to appeal to intention recognition. Thanks go to a reviewer for highlighting the connection.

[22] It is an open question whether non-human primates are capable of joint attention (e.g., Leavens, 2011).

[23] For a recent critical analysis, see Battich and Geurts (2020).

constrained by the standpoint of the perceiver. Since the participants in an episode of joint attention do not take over each other's standpoints, it is not clear how there can be a perceptual relation with three constituents. The consideration that children in joint attention are exposed to a contrast between two kinds of doings, and two ways of relating to their environment, can help here. It does so by arguing that one-year old children who begin to jointly attend to their surroundings with others acquire a new spatial frame of reference. They begin to operate with a social spatial order in which they can draw a practical distinction between doers who are apprehended by means of proprioception and perception, and the perceptually presented objects at which their doings are directed.

A spatial frame of reference is, for present purposes, an ordering of a creature's perceptual surroundings that allows the creature to carry out doings on the objects in these surroundings. For this to be possible, the proprioceptive and perceptual information it uses to perform a bodily movement has to be integrated at the location it occupies. I call this location a "standpoint".[24] Creatures in primitive social space occupy standpoints. Social interaction involves two doers who occupy different standpoints. But in primitive social space (which humans operate in when they are first capable of primary intersubjectivity), there is no differentiation between locations and standpoints. By contrast, joint attention requires that its participants be capable of this differentiation. It requires, minimally, that each participant be able to work out the location of the object of shared attention by means of an exercise in triangulation that treats the other's location as a standpoint whence doings can be executed. So much is required to draw another's attention to an object and to let yourself be guided towards the object the other tries to make salient to you. When children begin to engage in joint attention, they differentiate between two locations in social space whose occupants are proprioceptively and perceptually apprehended, and a third location that is occupied by the object of their doing. They acquire a new spatial framework in which an object is presented relative to two standpoints so that shared attention to and manipulation of the thing becomes possible. Social externalism substantiates the paradoxically-seeming idea of a triadic perceptual relation by conceiving of the perceivers as doers and by arguing that joint attention is facilitated by a spatial order in which objects' location is determined relative to two standpoints whose occupants are apprehended intermodally.

Creatures who operate in social space thus enjoy a practical kind of knowledge of the location of objects, relative to the standpoints of the two participants in the joint undertaking: they know where the object of their shared attention is in social space. If two perceivers operate in social space, they necessarily know the perceptual object's location in that space; if they don't know its location, they are not operating in social space. The location of the object of joint attention is demonstratively rather than descriptively identified: a description can determine the location of an object relative to objects at other locations, but it cannot capture the fact that this location exists in social space and that the other locations are occupied by doers. The minimal knowledge enjoyed by joint perceivers can be expressed as follows:

[24] I do not here discuss the difficult question of how exactly to think of such a standpoint (or egocentre). One locus classicus is Evans's (1982) work on spatial representation. For proposals of how to think about the notion of an egocentre within an action-based view of perception, see e.g. Grush (2001, 2007) and Schellenberg (2007).

(KL) (*This) is where the object of the interaction is.

(*This) denotes the use of a demonstration by means of which an object is identified in social space. For such an identification to take place, two perceivers must each socially triangulate its location relative to the standpoint of their co-perceiver. This mutual triangulation is interactive, and so their knowledge of object location in social space is necessarily of a bodily and enacted kind. Because the account does not stipulate that the child in early joint attention can discriminate between herself and her co-perceiver, there is no requirement that the child be able to ascribe this knowledge to herself or others. The spatial knowledge enjoyed by children in early joint attention is not what Williamson (2000) calls "luminous": the child who knows (KL) does not therefore also know that she knows it. Knowing (KL) does thus not require the deployment of psychological concepts. One-year old children who are beginning to master joint attention are not, on the present view, psychological reasoners. The child who has singled out an object by participating in an exercise of mutual social triangulation is enjoying a practical kind of spatial knowledge: she knows where the object of an interactive doing is located in social space in a way that allows her to point it out to and manipulate it with her caregiver.[25]

2 From Joint Attention to Perspective-Taking

In this section I offer an externalist account of the trajectory from joint attention to perspective-taking. I address the relevant developmental milestones in chronological order.

iv) *Level-1 Perspective-Taking*

At about 2.5 years of age children can solve level-1 perspective taking tasks in which they have to judge what another can see (Flavell, 1992). But it is only in their fifth year of life that they can judge what an object looks like from a different perspective. There is something paradoxical to the observation that children in their third year can judge what others can see but not how they see it. Perspectives are always "on *something*", as Moll and Tomasello (2006, p. 604) put it, and perceptual objects are not presented to children in their third year as shorn of their standpoint-dependent properties. It can hence seem natural to think that a perceiver who can judge what another can see must also be able to judge how she sees it. Social externalism explains the puzzling ontogenesis of perspective-taking by introducing the notion of a "triadic core model" of the spatial-perceptual relation that obtains in joint attention and that eventually allows the child to judge what is visible from various standpoints.

Theoretical models are ubiquitous in science. Godfrey-Smith(2005), Maibom (2003, 2007, 2009) and Spaulding (2018) have appealed to the notion of a theoretical model to

[25] The epistemological backdrop of the approach I am developing here is the "knowledge first" programme most prominently defended by Williamson (2000). The approach I am recommending is sympathetic to Nagel's (2017) view that the contrast that matters for mental state attribution is not between true and false beliefs but between factive and non-factive mental states, and that observation of perceptual access is a promising entry point for mental state attribution; thanks go to one of my reviewers for highlighting the connection. See section III (b) for more on the epistemology of social space.

explain humans' capacity for folk psychological reasoning – the capacity to attribute mental states to others, typically with the aim of explaining or predicting their behaviour. Godfrey-Smith(2005, p. 7) characterizes such models as follows:

> A theoretical model is a hypothetical system (or family of such systems), specified using some representational medium, that is constructed for the purposes of comparison to a target.

What matters here is that models are unlike theories in that they do not amount to sets of laws, or lawlike generalisations, that make possible the explanation or prediction of a target event. They do not, by themselves, have accuracy conditions: they do not aim at correspondence with states-of-affairs. They are useful only because they can be *applied* in particular contexts so as to make possible "comparisons" with the target. There thus is a distinction between the model, which in itself does not have accuracy or success conditions, and its applications, in which modellers fit the model to particular situations.

These features of models – constructs that do not themselves have accuracy conditions, do not amount to theories, and make possible comparisons with particular targets – make them useful with regard to the question of how a perceiver can solve level-1 perspective taking tasks. As children in their second year continue to pursue and prolong social interactions with their caregivers, they are acquiring a model of the triadic spatial and perceptual relation that underwrites joint attention. The child who can answer the question of what another can see works out what is visible from the two locations in the triadic relation that are occupied by doers by comparing the model to the perceptual context in which she finds herself.

In an experiment of Moll, Carpenter, and Tomasello (2011), two-year old children are familiarized with two objects by a caregiver who then leaves the room and in whose absence a third object is introduced. Upon her return, the caregiver asks for "the one she hasn't yet seen". The child reliably selects the third object. However, if the caregiver remains visually present while her line of sight is blocked, the child selects objects at chance. Moll and her colleagues explain this finding by suggesting that younger children tend to treat the co-presence of an adult with whom they interact as sufficient for shared perceptual experience. This interpretation is in line with the present proposal: the two-year old child is in the process of acquiring a model of the triadic relation that underwrites joint attention but applies it imperfectly. The child cannot yet accurately compare the relation specified by the model to the situation at hand, so that the scenario in which the caregiver is co-present is taken to meet the conditions imposed by the model. It is only when they can accurately compare their perceptual situation with their model of the triadic relation that they are able work out that the interacting adult's sight of the object is blocked by a barrier.

On this view, level-1 perspective taking tasks are solved not by an exercise in perspective taking in the sense of an imaginative simulation of another's viewpoint on a particular object, but by the increasingly sophisticated application of the triadic model. The child who can judge what another can see is not ascribing perceptual knowledge to that perceiver; she is judging what is visible from their standpoint. She is not deploying psychological concepts but has learned to compare the triadic model with her own perceptual situation. This dissolves the apparent paradox of the development of

perspective-taking: the application of the triadic model precedes the ability to ascribe visual perspectives to particular others.

e) *False Belief Tests*

Children pass classical false belief tests around four years of age (e.g., Perner & Lang, 1999; Wimmer & Perner, 1983). When children begin to consider and respond to the question of what another agent knows about an object's location, they demonstrate their ability to ascribe spatial knowledge to a variety of subjects. They have thus made the transition from being "externalists" who apply the triadic model to work out what is visible from various standpoints to psychological reasoners who can attribute to themselves or others knowledge of where the objects of their doings are. But below four years of age, when they cannot yet correctly answer questions about an actor's belief about object location if the object has moved since he last saw it, they ascribe to actors their own knowledge of where the thing is. They have acquired a rudimentary concept of perspective but have no grip yet on its temporal aspect: their concept of perspective is purely perceptual. They take themselves to be operating in an environment whose objects are perceived by all occupants and in which the actor thus shares their perceptual knowledge of the location of the object. It is only once they can discriminate between their own spatiotemporal perspective and that of the actor that they can solve classical false belief tasks. False-belief reasoning, on the externalist view, is a kind of perspective-taking that takes into account the divergence between the mental lives of the occupants of spatially and temporally defined standpoints and that makes use of the concept of the non-factive mental state of belief to this end.

f) *Level-2 Perspective Taking*

Soon after passing false belief tests, children begin to be able to solve level-2 perspective taking tasks. A classic example is the Turtle Task (Masangkay et al., 1974), in which a child and an adult sit at opposite ends of a table with a picture of a turtle between them. Children aged 4.5 years and older are able to acknowledge that the adult sees the turtle in a different orientation, whereas younger children tend to describe the turtle from their own perspective. Solving the task requires that the child make an explicit judgement about another perceiver's spatial knowledge. The child now is engaged in a form of psychological reasoning that builds on and transcends the kind of mental state ascription required to solve false belief tests.

Level-2 perspective-taking requires both that the perceiver can judge another's knowledge of object location and knowledge of the object's spatial properties. The second builds on the first: you cannot know what another perceiver knows about an object's spatial properties if you don't know the location of the object relative to the perceiver. But you can know what another perceiver knows about an object's location even though you don't know what the object looks like from her perspective. The first kind of knowledge is demonstrated by the capacity to solve false belief tasks, the second kind by level-2 perspective taking. So it is no accident that perspective-taking can only be accomplished once the child can attribute perceptual knowledge of object location to another perceiver. At the same time, the two capacities are closely related. You cannot perceive bare objects, shorn of their properties, at particular locations; and

so the four-year old's ascription of perceptual knowledge of, or false belief about, object location to another is incomplete if not coupled with the ability to take the other's perspective on the object. Spatial awareness does not stop at objects' door. Hence it is not surprising that the capacity for level-2 perspective taking and mastery of explicit false belief tasks develop in close developmental proximity.

On Kessler and Rutherford's (2010) view, judging objects' perspective-dependent spatial properties from standpoints other than one's own requires an imaginative bodily rotation. The child who has to judge the turtle's orientation from the perspective of the actor, positioned on the opposite side of the card that displays the turtle, is imagining herself occupying the actor's standpoint relative to the card. Spatial level-2 perspective taking thus requires that the subject have a reflective understanding of the intermodal relation between the body's internal spatial arrangement and the object of perception and action.[26] Social externalism is well equipped to accommodate this consideration: since it argues that our understanding of perspective begins in joint attention, and since it conceives of joint attention in terms of a social doing in which participants use a spatial frame of reference that integrates proprioception and perception at a variety of locations, the reflective appropriation of this framework produces the insight that perceptual objects can be apprehended from various such locations. This opens the door to an account on which level-2 perspective taking requires an imaginative rotation not only of the body but of the body in relation to the object whose perspective-dependent looks the child is asked to judge.

3 The Frame Problem

In this section I explain how social externalism dissolves the frame problem. I have presented an account on which the development of perspective-taking begins with a primitive kind of intermodal social space that gives rise to a social frame of reference. A variety of milestones of children's socio-cognitive development can then be explained in terms of children's increasingly reflective acquisition of that frame, which culminates in their understanding of others as occupants of spatiotemporal perspectives on the objects of perception.

One key consideration implicit in this account is that humans' perceptual environment is presented to them as public. In a public environment, perceptual knowledge about the objects it contains can be demonstratively and descriptively shared with other human perceivers. Infants learn about the public character of their environment early on, when they begin to direct caregivers' attention to the objects of their own doings. In joint attention, everything is out in the open: triadic perceptual relations produce common knowledge about its objects in the participants (Campbell, 2005; Peacocke, 2005) and the doings through which perceivers acquire this knowledge are not in need of psychological interpretation. In psychological sense making, the interpreter undertakes a kind of perspective-taking in which he applies the triadic model in contexts that cannot be understood by perception alone. The introduction of "hidden" mental states to make sense of another's doings in non-joint contexts such as false-belief scenarios can then be understood as the attempt to restore the openness of the environment in which both interpreter and agent operate.

[26] In a related vein, Schellenberg (2007) argues that knowing an object's perspective-transcendent spatial properties requires knowing how to act on it.

g) *The Frame Problem in Theory of Mind*

The frame problem arises for TT-based accounts of mindreading, including model theory. The problem was originally discussed in computational AI and is concerned with the selection of information necessary to carry out some task. A potentially infinite amount of environmental information is available to the AI, but finite computational resources block complete representation of the context in which the task occurs. The context has to be "framed" so that only information relevant to the task at hand is selected for representation. Yet what is relevant can itself not be determined context-independently but only relative to the context in which the task is specified. The problem is that the context itself cannot be independently specified, and so a regress arises (Dreyfus, 1992; Wheeler, 2008). Inferring (hidden) beliefs from observable behavior is an abductive inference to the best explanation (Apperly, 2011, p. 118 f.); as such, it is a "best guess", all things considered, that always leaves open the possibility of rival explanations. Depending on how an action context is described, and how much contextual information is admitted into its description, the theorist's explanatory mentalizing may yield quite distinct results. As Apperly (2011, p. 9) puts it: "Applied to mindreading, the problem is that an agent may have any number of beliefs (and other mental states), any of which might be relevant when trying to judge what the agent will think or do in a given situation."

h) *The Epistemological Backdrop of Social Externalism*[27]

Consider again the most primitive form of knowledge that is enjoyed by creatures who are triangulating the location of an object relative to the standpoints of two perceivers in joint attention:

(KL) (*This) is where the object of the interactive doing is.

(KL) is a kind of knowledge that could not be enjoyed by any perceiver on their own. Because (*This) expresses the use of the demonstrative in episodes of joint attention, the referent is individuated by each participant relative to the standpoints of both. On the developmental trajectory I have presented, children in early joint attention are not yet able to explicitly differentiate between the occupants of distinct standpoints and are thus not in a position to ascribe knowledge of object location to themselves or others. Their knowledge of (KL) is of a practical kind that enables them to learn from and act on the jointly perceived object with a caregiver. It is only when the child begins to operate with the concept of perspective that she can ascribe perceptual knowledge to herself and others. The mature participant in an episode of joint attention, who operates with a reflective conception of the triadic model, knows (KL) in common with co-perceivers, and each participant reflectively knows (KL) in virtue of it being entailed by their common knowledge of (KL). Common knowledge is what Williamson (2000) calls "luminous"[28]: its subjects necessarily know that they have it.[29] The openness of

[27] This section summarises aspects of a much larger discussion in Seemann (2019).

[28] For an argument that there is no such thing as common knowledge, see Lederman (2018). For the view that common knowledge is impossible because subjects can never know whether they have it, see Sperber and Wilson (1995, p. 23). For a comprehensive argument in favour of the possibility of perceptual common knowledge, see Seemann (2019).

[29] This is compatible, of course, with one or more subjects mistakenly believing that they have it.

joint perceptual contexts is thus always known to participants capable of mental state ascription. Mature joint perceivers always know in common where the object of their attention is in social space. They can enlarge the stock of what they know in common about the object by highlighting some of its perceptual characteristics to each other or by introducing non-perceptual facts about it that were not hitherto commonly known. The perceptual context of perceivers is joint to exactly the extent that the facts about it are commonly known. Not everything a joint perceiver is perceiving is being jointly perceived, and not all of a perceiver's knowledge about a jointly perceived object is part of the context that is characterized by epistemic openness.

Ordinary perceptual contexts can be transformed into joint contexts and perceptual knowledge can be transformed into common knowledge by means of communications (verbal or nonverbal) in which pointing gestures play a vital role. I call perceptual environments that can be so transformed "public" or "epistemically open".[30] Social externalism takes no view on the conditions under which such transformation is possible. It does not appeal to "normal" (Schiffer, 1972) or "ideal" conditions under which the facts that obtain in a given domain can be known in common. Such conditions, apart from being notoriously difficult to define, would in any case not be useful in the present context. Suppose there were an epistemic condition on which the individual reasoner knew that the argument he construed in order to interpret a doing was out in the open and could become common knowledge between the inhabitants of the relevant space. The reasoner could still be mistaken about whether the condition obtained. Compare the situation of this reasoner with the situation of a perceiver who believes she is operating in a joint context and thus takes herself to know in common particular facts about the perceptual object with someone else. Even though common knowledge is luminous, so that the joint perceivers' common knowledge that p entails each perceiver's reflective knowledge that she knows p in common with her joint perceiver, it remains possible that the perceiver falsely believes she knows p in common with another perceiver. This is the case, for instance, when I misconstrue your direction of gaze so that I mistakenly take you to be attending to the object I am pointing out to you while you are in fact looking at a different object. Correspondingly, the situation of the psychological reasoner who operates in an epistemically open environment is as follows: her necessary knowledge that she is operating in such an environment when she is does not preclude the possibility of her falsely believing to be operating in such an environment when she is not. Highlighting the public character of ordinary human perceivers' environment therefore does not amount to the claim that the individual perceiver can always know whether a particular perceptual fact she knows can be or is known in common with other perceivers; mistakes are always possible.

The public character of the environment in which human perceivers operate forms the background of the externalist account of mental state ascription. In such an environment, a variety of locations can be occupied by subjects of doings and these doings are presented in relation to the objects they involve. The observer who can see another perceiver as the subject of a doing and who can attribute knowledge about its

[30] Williamson (2000) introduces the notion of a "cognitive home", in which all facts are open to view, and argues against its possibility. Public perceptual environments constitute a social version of such a cognitive home.

object's location to that subject has gleaned the most primitive form of insight into another's mental life.[31] This kind of mental state attribution does not require information that is not perceptually available to the perceiver. But the observer may have interpretive needs that cannot be met by ascribing to the agent the kind of spatial knowledge that can be attributed on the basis of perception alone. Then this context is no longer public: the facts needed to make sense of the doing, relative to the observer's interpretive needs, are not known in common by observer and agent. In this situation, the observer has to introduce new information to satisfy her interpretive need. The need is satisfied just when the epistemic openness of the context is restored.

False belief reasoning is the first instance in human development of this kind of restoration of a perceptual context's epistemic openness. Because an observed doing cannot be made sense of in the context of what is plainly observable, the child introduces mental state concepts as premises in an argument whose conclusion reinstates the public character of the context in which the doing takes place. She thus construes an argument of this kind:

P1: The actor has seen the object being placed in the green box
P2: The object has been moved to the red box
P3: The actor believes the object to be in the green box
C: Hence the actor reaches into the green box

(P1) states what is public knowledge between the child and the actor. (P2) states what is known to the child but not the actor. (P3) introduces a new premise whose stipulation warrants the conclusion (C) that explains the actor's doing. Introducing (P3) is thus aimed at restoring the public character of the environment in which child and actor are operating. Once (P3) is out in the open, the environment is public again: it now satisfies the child's interpretive need.

Consider how the child has restored the public character of the environment that she shares with the actor. The child has in essence carried out an exercise in perspective taking. She has carried out a psychological kind of triangulation: she has addressed the question of why the actor reached into the green box by asking what knowledge was available to the actor and explained his doings in those terms. So the child is making use of the triadic model to restore the public character of the environment in which she observes the actor's doing: she is holding constant her own perspective (whence the question of why the actor performs the doing arises) and the doing while solving for the actor's standpoint. In simple false belief tests she can restore the public character of the environment without having to import any information not provided by the perceptual context: it is sufficient to stipulate a lack of perceptual knowledge on the actor's part. No information that is not available within the perceptual context needs to be imported for the child to restore the openness of the environment in which she and the actor operate.

i) *The Frame Problem Dissolved*

[31] See Nagel (2017) for a related view.

Psychological sense making often demands more than the simple ascription of a false belief about object location to an agent. It then requires the observer to introduce information that is not (or was previously) perceptually available into the context in which an interpretive need arises. Consider the case of Alice, whom you observe as she moves towards the fridge. You want to know why she is moving towards the fridge. The mere observation of the doing, without introducing additional information or mental state premises in an explanatory argument, does not enable you to answer the question. Now the frame problem appears to arise: you could explain what Alice is doing in a potentially unlimited number of ways. Perhaps Alice is going to the fridge because she is hungry and knows that the fridge contains food; perhaps she saw earlier that the lightbulb inside was broken and wants to replace it; perhaps she is only accidentally approaching the fridge as she is practicing a dance move. How do you select amongst these (and many possible other) possible interpretations?

Social externalism addresses this question by arguing that psychological sense making is a form of perspective taking in public space. The interpreter deploys a triadic model of that space of which she herself is a constituent. She herself occupies a particular perspective, and it is from this perspective that her interpretive need arises and that additional information is introduced to interpret another's doing. Perspective taking is always from somewhere, relative to something: to take someone else's perspective on some third event, you have to enjoy a perspective on the event yourself. For the grown-up psychological reasoner, this perspective includes not only physical location but also her cultural, social, cognitive and psychological background[32] (things like norms, character traits, implicit and explicit knowledge, and past experience) and the various constraints under which she is labouring (such as her biological, cognitive and conceptual limitations and the time and effort she can afford to spend on the interpretive task). The adult psychological reasoner thus deploys a psychologically, cognitively, culturally, and situationally enriched version of the triadic model in order to satisfy an interpretive need that arises from her perspective. The need is satisfied when the public character of the environment in which the doing that is being interpreted is restored. Consider this scenario:

> You and Alice have agreed to go to a gallery together. You have put on your shoes and coat, have found your phone and keys and are ready to leave. "Let's go," you say through the open kitchen door. Alice nods and subsequently begins to move towards the fridge. You wonder why she is letting you wait.

A description of your perspective in this scenario might include the following: you want to go to the gallery with Alice; you know it is only open for another hour and you know that she knows this; you have been grocery shopping earlier and know that the fridge contains food. In your interpretation of what Alice is doing, you might include as relevant information that she has skipped lunch today and take this information as evidence for the premise that she is hungry. Because you want to make it to the gallery in time, you need to know why she is letting you wait. Then you can construe the following argument:

[32] The notion of a representation-enabling "background" of nonrepresentational capacities goes back to Searle (1983). I am not, for present purposes, using the notion in strictly nonrepresentational terms.

- P1: Alice is moving towards the fridge (a statement expressing what is publicly perceptually observable in the relevant context)
- P2: Alice is hungry (the introduction of a statement expressing Alice's state of
- mind)
- P3: Alice has skipped lunch today (a statement introducing evidence in support of (P2)
- C: Alice wants to get something to eat

This argument may turn out to be unsound, but *given your standpoint* it best explains why Alice is letting you wait. The qualifier matters: if you had more time, or a different motivation for your question, or knew more about Alice, you might introduce other evidence and construe an argument with a different conclusion. For instance, as you keep waiting you might remember that Alice always checks whether the door of the fridge is closed before going out. Now your perspective has shifted slightly: you can now construe a rival explanation of Alice's doings. There is no determinate saying, from your standpoint, which explanation (if either) is correct. But the frame problem is not that the psychological sense maker is fallible in her interpretive efforts. The question is why humans do not have to select amongst infinite amounts of potentially relevant information when making sense of what another is doing. Explaining why human reasoners don't face the frame problem does not require an account of how they come to choose between rival explanations of what another is doing. All that is required is an account of why we are not in the position of the AI that, because it is not operating in a public context, has no perspective on what it observes. On the social externalist view, we are not in that position because we occupy a particular standpoint in a social kind of space that gets enriched in various ways as we understand more about the conditions under which we operate. This standpoint gives rise to our interpretive needs and determines what information is available in our attempts to make sense of what others do. Because they are occupants of particular perspectives in a public environment, ordinary psychological sense makers are never confronted with an infinite sea of information on which they could draw in order to take another's perspective on what they are doing. Thinking of psychological interpretation in terms of the application of a biologically and culturally enriched version of the triadic model explains why.[33]

The picture, then, is that the psychological reasoner is attempting to make sense of an observed doing in order to satisfy an interpretive need. To this end, he construes an argument that contains as premises statements expressing the ascription of mental states to the subject of the doing and statements introducing information that serves as supporting evidence. The selection of interpretively useful mental states and relevant information is carried out from his perspective, which imposes on him a range of background capacities and constraints. These capacities and constraints may not be cognitively present to the interpreter, but they determine the information available to him in support of the introduction of interpretively useful mental state premises. The soundness of the argument is determined by whether it restores the public character of the environment in which the doing takes place.

[33] There is an interesting question whether an interpreter's perspective determines her interpretation of an observed doing or whether it merely constrains the information she can select to support it. Discussion is not possible here. It is also not required for the present argument.

I said that an environment is public or epistemically open if all facts about it can be known in common by its occupants. Since openness does not entail the inhabitant's ability to always distinguish between environments that are open and those that are not, the interpreter is not always in a position to know whether the openness-restoring argument is sound and whether its premises are true. This is all as it should be: mentalising is both delicate and fallible and there is no saying whether a psychological interpretation of what someone is doing is correct. Even if your interpretation of Alice's movement towards the fridge as motivated by her desire to get a bite to eat before venturing out is apparently vindicated by her subsequent retrieval and consumption of a sandwich, it may still turn out that she is not hungry at all and is eating the sandwich and thus delaying your departure for quite different reasons. Social externalism leaves room for the principled fallibility of psychological interpretation while still insisting that there is a norm – the public character of the shared environment in which doings take place - against which better psychological interpretations can be distinguished from worse ones.

4 Conclusion

The frame problem, though arising in our theorising about mental state ascription, is a difficulty not faced by actual psychological reasoners. One conclusion you can draw is that it is the theory that creates the problem. This is the stance embraced by interactionists about the social mind (e.g., Gallagher, 2012). But the conclusion sits oddly with the observation that we routinely ascribe mental states to others and ourselves in psychological sense making. The frame problem brings to the fore two starkly distinct ways of thinking about the mind and its place in nature. Defenders of interactionism see sociality as constitutive of the mind: it is in our interactions with others that we come to make sense of ourselves and the environment in which we operate. Social interaction immediately enables social perception: mind is revealed in intersubjectivity and mental states can be directly perceived. Defenders of Theory of Mind approaches, by contrast, are individualists: the need to make sense of what others do arises for the individual reasoner. On this view, the interactionist claim that the very perception of others' movements directly reveals their mental lives is simply false: there is no perceptual knowing of others' action-driving beliefs and desires; rival explanations are always possible. The protracted and inconclusive debate between the defenders of both views reveals the two camps' quite starkly distinct foundational assumptions about the nature of the social mind.

Social externalism takes it that both views have something valuable to contribute to this discussion and offers a reconciliation. It is made possible by the externalist insight that mental states can only be individuated relative to the environment in which they are entertained and by taking it that sociality is engrained in the spatial framework with which social creatures operate. On this view, social interaction directly reveals others as bearers of mentality because their object-directed movements, at the places they occupy, are intermodally apprehended. But because agents are conceived as doers, the apprehension of others as creatures possessed of a mental life does not require the ascription of particular mental states to them. It is only with the acquisition of the concept of perspective on objects in a public environment that mental state ascription

becomes necessary and possible. The environment of psychological reasoners is thus inherently perspectival: the objects in them can be perceived from a variety of viewpoints and psychological reasoning begins with the knowledge that bearers of mental states occupy such perspectives. This externalist approach makes it possible to integrate a variety of developmental milestones into an account of the ontogeny of perspective taking, and it can explain why the frame problem does not arise without denying the relevance of mental state attribution in social understanding. It thus offers a way out of the stalemate between defenders of interactionism and theory of mind.

References

Apperly, I. (2011). Mindreaders: The cognitive basis of "theory of mind". Hove: Psychology Press.

Archer, J. (1992). *Ethology and human development*: Rowman & Littlefield.

Avramides, A. (2001). Other minds. London: Routledge.

Baron-Cohen, S., Tager-Flusberg, H., & Cohen, D. (1999). Understanding other minds II. Oxford: Oxford University Press.

Battich, L., and B. Geurts. 2020. Joint attention and perceptual experience. *Synthese, forthcoming.*

Borg, E. 2007. If Mirror neurons are the answer, what was the question? *Journal of Consciousness Studies* 14 (8): 5–19.

Briscoe, R., & Grush, R. (2020). Action-based theories of perception. In E. N. Zalta (Ed.), *the Stanford encyclopedia of philosophy*. Doi:https://plato.stanford.edu/archives/sum2020/entries/action-perception.

Campbell, J. 2005. Joint attention and common knowledge. In *Joint attention: Communication and other minds*, ed. N. Eilan, C. Hoerl, T. McCormack, and J. Roessler, 287–297. Oxford: Oxford University Press.

Campbell, J. (2011). An object-dependent perspective on joint attention. In a. Seemann (Ed.), *joint attention: New developments in psychology, philosophy of mind, and social neuroscience* (pp. 415 - 430ß). Cambridge, MA: MIT press.

Costantini, M., & Sinigaglia, C. (2011). Grasping affordance: A window onto social cognition. In A. Seemann (Ed.), Joint attention: New developments in psychology, philosophy of mind, and social neuroscience. Cambridge, MA: MIT Press.

Davidson, D. 1973. Radical interpretation. *Dialectica* 27: 314–328.

De Jaegher, H., E. Di Paolo, and S. Gallagher. 2010. Can interaction constitute social cognition? *Trends in Cognitive Science* 14 (10): 441–447.

De Vignemont, F. 2009. Drawing the boundary between low-level and high-level mindreading. *Philosophical Studies* 144 (3): 1–10.

De Vignemont, F. 2018. Peripersonal perception in action. *Synthese.* 198: 4027–4044. https://doi.org/10.1007/s11229-018-01962-4.

Dreyfus, H. (1992). What computers still can't do: A critique of artificial reason. Cambridge, MA: MIT Press.

Dreyfus, H. (1993/2014). Heidegger's critique of the Husserl/Searle account of intentionality. In Skillful coping: Essays on the phenomenology of everyday perception and action (pp. 76–91). Oxford: Oxford University Press.

Evans, G. (1982). The varieties of reference. Oxford: Oxford University Press.

Flavell, J. H. (1992). Perspectives on perspective-taking. In H. Beilin & P. B. Pufall (Eds.), *the Jean Piaget symposium series. Piaget's theory: Prospects and possibilities* (Vol. 14, pp. 107-139). Hillsdale, NJ: Erlbaum.

Gallagher, S. (2005). How the body shapes the mind. Oxford: Oxford University Press.

Gallagher, S. 2008. Direct perception in the intersubjective context. *Consciousness and Cognition* 17: 535–543.

Gallagher, S. (2012). Social cognition, the Chinese room and the robot relies. In Z. Radman (Ed.), Knowing without thinking: Mind, action, cognition and the phenomenon of the background (pp. 83–97). London: Palgrave Macmillan.

Gallese, V., and A. Goldman. 1998. Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Sciences* 2 (12): 493–501.

Godfrey-Smith, P. 2005. Folk psychology as a model. *Philosophers' Imprint* 5 (6): 1–16.

Goldman, A. (2006). Simulating minds: The philosophy, psychology, and neuroscience of mindreading. Oxford: Oxford University Press.

Gopnik, A., & Meltzoff, A. (1997). Words, thoughts, and theories. Cambridge, MA: MIT Press.

Grush, R. 2001. Self, world and space: The meaning and mechanisms of ego- and allocentric spatial representation. *Brain and Mind* 1: 59–92.

Grush, R. 2007. Skill theory v2.0: Dispositions, emulation, and spatial perception. *Synthese* 159: 389–416.

Hobson, P. (2002/2004). The cradle of thought. London: Macmillan.

Hutto, D. (2011). Elementary mind minding, enactivist-style. In A. Seemann (Ed.), Joint attention: New developments in psychology, philosophy of mind, and social neuroscience (pp. 307–341). Cambridge, MA: MIT Press.

Kessler, K., and H. Rutherford. 2010. The two forms of visuo-spatial perspective taking are differently embodied and subserve different spatial prepositions. *Frontiers in psychology*(1), 1-12. 1. https://doi.org/10.3389/fpsyg.2010.00213.

Kulke, L., J. Johannsen, and H. Rakoczy. 2019. Why can some implicit theory of mind tasks be replicated and others cannot? *A test of mentalizing versus submentalizing accounts. PLOS One.* 14: e0213772. https://doi.org/10.1371/journal.pone.0213772.

Kulke, L., M. Reiss, H. Krist, and H. Rakoczy. 2018a. How robust are anticipatory looking measures of theory of mind? Replication attempts across the life span. *Cognitive Development* 46: 97–111.

Kulke, L., Von Duhn, B., Schneider, D., & Rakoczy, H. (2018b). Is implicit theory of mind a real and robust phenomenon? Results from a systematic replication study. Psychological Science, *0*(0956797617747090), 29, 888, 900.

Leavens, D. (2011). Joint attention: Twelve myths. In A. Seemann (Ed.), Joint attention: New developments in psychology, philosophy of mind, and social neuroscience (pp. 43–72). Cambridge, MA: MIT Press.

Lederman, H. 2018. Uncommon knowledge. *Mind* 127 (508): 1069–1105.

Leslie, A. 2000. 'Theory of mind' as a mechanism of selective attention. In *The new cognitive neurosciences*, ed. M. Gazzaniga, 1235–1247. Cambridge, MA: MIT Press.

Leslie, A., O. Friedman, and T.P. German. 2004. Core mechanisms in "theory of mind". *Trends in Cognitive Sciences* 8: 528–533.

Maibom, H. 2003. The mindreader and the scientist. *Mind & Language* 18 (3): 296–315.

Maibom, H. 2007. Social systems. *Philosophical Psychology* 20 (5): 557–578.

Maibom, H. 2009. In defence of (model) theory theory. *Journal of Consciousness Studies* 16 (6–8): 360–378.

Maister, L., F. Cardini, G. Zamariola, A. Serino, and M. Tsakiris. 2015. Your place or mine: Shared sensory experiences elicit a remapping of peripersonal space. *Neuropsychologia* 70: 455–461.

Masangkay, Z.S., K.A. McCluskey, C.W. McIntyre, J. Sims-Knight, B.E. Vaughn, and J.H. Flavell. 1974. The early development of inferences about the visual percepts of others. *Child Development* 45: 357–366.

Meltzoff, A. (1993). Molyneux's babies: Cross-modal perception, imitation, and the mind of the pre-verbal infant. In N. Eilan, R. McCarthy, & B. Brewer (Eds.), Spatial representation: Problems in philosophy and psychology (pp. 219–235). Oxford: Blackwell.

Meltzoff, A., and M.K. Moore. 1977. Imitation of facial and manual gestures by human neonates. *Science* 198: 75–78.

Meltzoff, A., & Moore, M. K. (1995). Infants' understanding of people and things: From body imitation to folk psychology. In J. L. Bermudez, A. Marcel, & N. Eilan (Eds.), The body and the self (pp. 43–69). Cambridge, MA: MIT Press.

Michael, J. 2011. Interactionism and mindreading. *Review of Philosophy and Psychology* 2: 559–578.

Moll, H., M. Carpenter, and M. Tomasello. 2011. Social engagement leads 2-year olds to overestimate others' knowledge. *Infancy* 16: 248–265.

Moll, H., and A. Meltzoff. 2011. How does it look? Level 2 perspective taking at 36 months of age. *Child Development* 82: 661–673.

Moll, H., and M. Tomasello. 2006. Level 1 perspective-taking at 24 months of age. *British Journal of Developmental Psychology* 24: 603–613.

Nagel, J. 2017. Factive and nonfactive mental state attribution. *Mind & Language* 32: 525–544.

Onishi, K.H., and R. Baillargeon. 2005. Do 15-month-old infants understand false beliefs? *Science* 308: 255–258.

Oostenbroek, J., T. Suddendorf, M. Nielsen, J. Redshaw, S. Kennedy-Costantini, J. Davis, S. Clark, and V. Slaughter. 2016. Comprehensive longitudinal study challenges the existence of neonatal imitation in humans. *Current Biology* 26: 1334–1338. https://doi.org/10.1016/j.cub.2016.03.047.

Peacocke, C. (2005). Joint attention: Its nature, reflexivity, and relation to common knowledge. In N. Eilan, C. Hoerl, T. McCormack, & J. Roessler (Eds.), Joint attention: Communication and other minds (pp. 298–324). Oxford: Oxford University Press.

Perner, J., and B. Lang. 1999. Development of theory of mind and executive control. *Trends in Cognitive Sciences* 3: 337–344.

Reddy, V. (2008). How infants know minds. Cambridge, MA: Harvard University Press.

Reddy, V. (2011). A gaze at grips with me. In A. Seemann (Ed.), Joint attention: New developments in psychology, philosophy of mind, and social neuroscience (pp. 137–157). Cambridge, MA: MIT Press.

Rizzolatti, G., L. Fadiga, L. Fogassi, and V. Gallese. 1997. The space around us. *Science* 277: 190–191.

Schellenberg, S. 2007. Action and self-location in perception. *Mind* 116 (463): 603–632.

Schiffer, S. 1972. *Meaning*. Oxford: Oxford University Press.

Schönherr, J. 2017. What's so special about interaction in social cognition? *Review of Philosophy and Psychology* 8: 181–198. https://doi.org/10.1007/s13164-016-0299-y.

Searle, J. (1983). Intentionality: An essay in the philosophy of mind. New York: Cambridge University Press.

Seemann, A. (2019). The shared world: Perceptual common knowledge, demonstrative communication, and social space. Cambridge MA: MIT Press.

Soliman, T.M., R. Ferguson, M.S. Dexheimer, and A.M. Glenberg. 2015. Consequences of joint action: Entanglement with your partner. *Journal of Experimental Psychology* 144 (4): 873–888.

Spaulding, S. 2010. Embodied cognition and mindreading. *Mind and Language* 25 (1): 119–140.

Spaulding, S. (2018). How we understand others: Philosophy and social cognition. London and New York: Routledge.

Sperber, D., & Wilson, D. (1995). Relevance: Communication & cognition. Oxford: Blackwell.

Stratton, G.M. 1899. The spatial harmony of touch and sight. *Mind* 8: 492–505.

Stueber, K. (2006). Rediscovering empathy: Agency, folk psychology, and the human sciences. Cambridge MA: MIT Press.

Tomasello, M., and M. Carpenter. 2007. Shared intentionality. *Dev Sci.* 10 (1): 121–125.

Tomasello, M., & Moll, H. (2010). The gap is social: Human shared intentionality and culture. In P. M. Kappeler & J. B. Silk (Eds.), Mind the gap (pp. 331–349). Berlin: Springer.

Trevarthen, C. (2011). The generation of human meaning: How shared experience grows in infancy. In A. Seemann (Ed.), Joint attention: New developments in psychology, philosophy of mind, and social neuroscience (pp. 73–113). Cambridge, MA: MIT Press.

Van den Bos, E., and M. Jeannerod. 2002. Sense of body and sense of action both contribute to self-recognition. *Cognition* 85 (2): 177–187.

Wheeler, M. 2008. Cogition in context: Phenomenology, situated robotics and the frame problem. *International Journal of Philosophical Studies* 16 (3): 323–349.

Williamson, T. (2000). Knowledge and its limits. Oxford: Oxford University Press.

Wimmer, H., and J. Perner. 1983. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition* 13: 103–128.

Zawidzki, T. (2013). Mindshaping: A new framework for understanding human social cognition. Cambridge, MA: MIT Press.