AI & Soc (2019) 34:55–73
DOI 10.1007/s00146-017-0694-7

ORIGINAL ARTICLE

I in an other's eye

Alan Dix[1]

Received: 15 August 2015 / Accepted: 11 January 2017 / Published online: 1 March 2017

Abstract This paper presents a model of how the fundamental cognitive machinery of self emerges as an accident of sociality, reflecting Buber's assertion of the primacy of I–Thou relationships. This stands in contrast with the standard 'I first' model of theory of mind, which suggests that we understand others' thought processes by imagining ourselves in their heads. However, this standard model tacitly assumes that understanding oneself is in some way easy, counter to experience in knowledge elicitation, where experts find it hard to reflect on and externalise tacit thought processes. Furthermore, it is hard to create convincing evolutionary accounts for the spontaneous emergence of self. The paper argues that the reflexive understanding of self is both more plausible phylogenically as an evolutionary development and fully consonant ontogenically with research on childhood cognitive development. This reflexive understanding has practical implications for efforts to create artificial agents or robots that are in some sense conscious, and may also inform discussions of the ethical and spiritual implications of advances in artificial intelligence.

Keywords Consciousness · Self · Theory of mind · Artificial intelligence · Evolutionary psychology · Child development · Robotics · Ethics · Buber

✉ Alan Dix
alan@hcibook.com

[1] School of Computer Science, University of Birmingham, Birmingham, UK

1 Introduction

Traditionally, theory (or model) of mind is seen as putting oneself in another's head, assuming that one already knows one's self and thus understands others through that process. That is, understanding of 'I' precedes that of 'Thou' and 'Me' (myself in others' eyes). However, both phylogenically and ontogenically, there is good reason to believe that the opposite may be the case. We understand ourselves as sentient beings, because to understand others, we need to understand their models of us, including imputed intentions and feelings. In other words, 'me' precedes 'I' and consciousness of self is an accident of sociality, or in Buber's terms, emergence of 'I' from 'I–You' (Buber 1923, p 74).

Philosophically, this connects with notions such as Wittgenstein's meaning as use (Wittgenstein 1958) and philosophies of embodiment (Clark 1998; Shanahan 2010), but stands against both Descartes' (1759) primacy of the internal and Damasio's (1999) focus on extensions of physical body image. Taking a more reflexive view does not deny an internal life, but sees it as derived from cognitive abilities attuned to the external and inter-social. This is also related to the way that Burling suggests turning traditional theories of language development on their head, regarding the development of comprehension and reception as preceding production (Burling 2005).

Practically, this reflexive viewpoint suggests that models for artificial intelligence based on presenting and interpreting for others may be more fruitful than those starting with internal cognitive and emotional processes. For consciousness, this suggests starting with models that interpret others' intentions and, in particular, others' models of others: that is, looking to build theory of mind before self-modelling. For emotion modelling, this suggests creating robots or avatars that express and interpret emotion, rather than

building emotion models. The suggestion is that, by taking this stance, models of self-knowledge and personal affect will naturally emerge.

Ethically, this extends the behavioural view embodied in the Turing Test. From this view, one might judge that a robot or AI is an ethical agent when we see it as such. A more reflexive approach would add to this by asking to what extent an artificial agent is able to model and assess the extent to which they will be held accountable or blameworthy by others, or even more complex levels, such as "how will others consider I considered my actions?" Buber's "I and Thou" is both philosophical and theological, steeped in Judeo–Christian and other spiritual traditions. This raises the question of whether it is meaningful to consider artificial spirituality as well as ethics, possibly as an outcome of a programmed motivation towards relationship. While this at first appears even more speculative than considerations of artificial consciousness, there may, in fact, be very immediate practical implications.

The remainder of this paper starts by describing the phenomenon called 'theory of mind', different views on it and on the related issues of self and consciousness of self. It then looks at the phylogenic development of self, how it is that humans have evolved to have the fundamental cognitive machinery to be able to perceive their own mental state. Traditional models struggle to have plausible small steps that lead to 'self', and so an alternative more developmental path is suggested where theory of mind develops first, followed by consciousness of self. This phylogenic argument is backed up by an ontogenic review of the emergence of self in the child development literature, which shows that, at very least, self and theory of mind develop concurrently during early years. Methodologically, these two argument streams, the phylogenic and ontogenic, are mutually reinforcing, matching what is observed today (early childhood abilities) with a putative explanation of how this came to be. Furthermore, early (more primitive) stages in both can often also be observed in other animals. Finally, the article explores, as summarised above, the implications that this has for artificial models of self and concomitant implications for robot/AI ethics and spirituality.

2 Encountering other: theory of mind

"*Little does she know that I know that she knows that I know she's two timing me*" (The Kursaal Flyers 1976).

2.1 What is theory of mind?

Theory of mind is about our ability to see inside someone else's head, to attribute thoughts, desires, beliefs to them, to treat them as intentional beings.

Most adults are easily able to 'mind read' like this, "Oh yes, he's going to the shop because *he wants* a pint of milk", "she's running for the bus but *she doesn't know* it has already left". We may make mistakes, and some people are better at this than others, but we have no problem in imagining these thoughts in other people's heads. This conceptualising of another's thoughts is first-order theory of mind, but it can get more complex.

The Kursaal Flyers' 1970s hit is an example of higher order theory of mind: the singer's knowledge of his girlfriend's lack of knowledge of his knowledge of her knowledge of his knowledge that she has been two timing him.

First order Conceptualising of another's thoughts—"she knows that I know".

Second order Conceptualising of another's thoughts about your thoughts (or those of a third person)—"I know that she knows that I know".

Third order Conceptualising of another's thoughts about your thoughts of their thoughts "little does she know that I know that she knows that I know".

Fourth order The singer himself knows the statement.

Fifth order The singer must assume that we do not know that he knows this, otherwise why tell us (except maybe just to make catchy lyrics).

It is very hard to keep track of the lyrics of the chorus verbally, showing that our ability to manage these higher order theories of mind is limited. Happily, the song came with a video acting out the *story* in the verses of the song. It showed the singer, and across the street in a launderette, the singer's girlfriend with another man. The girl catches sight of the reflection of the singer in the chrome rim of the washing machine—she knows that he knows. However, she does not realise that he saw that she had seen him—she does not know that he knows that she knows. The complex modal reasoning becomes much simpler when embodied in a real situation.

2.2 How do we manage theory of mind?

There are two main explanations of theory of mind. The so-called 'theory theory' posits a cognitive model of the other person as a social being, but not totally unlike one's material model of a physical object like a rock; that is, one simply has an understanding of their actions and behaviours, without using an analogy to oneself. In contrast, 'simulation theory' suggests that we effectively imagine ourselves in the other's shoes; this is also the common way in which

theory of mind is talked of informally. In addition, in a special case of simulation theory, some writers adopt an ecological or embodiment view, taking the physical body and actions as central. Notably, Damasio suggests that we do not imagine our head in another's head so much as imagine being in another's body (Damasio 2010, p 104). There are even a few writers who effectively seek to eliminate (a particular meaning of) self and other entirely (Metzinger 2009).

However, those who regard the theory of mind and the idea of 'self' as real typically assume that we effectively 'know our own mind', albeit understanding the knowing in myriad different ways. For example, Tomasello talks about the ability of organisms to "*understand conspecifics as beings like themselves*" (emphasis Tomasello) and thus able to imagine themselves in the "*mental shoes*" of others (Tomasello 1999, pp 5–6). Likewise, as alluded to above, Damasio talks about understanding the actions of others by "*placing ourselves in a comparable body state*" (Damasio 2010, p 104).

In both cases, this putting ourselves into the place of another is supposed to give us greater understanding, under the assumption that we already know more about ourselves than about others. In other words, the dominant models of theory of mind assume that human consciousness of self and of our own thought processes precedes our understanding of other people's thought processes.

However, it is not so obvious that knowing one's own thought processes is so simple; indeed, all analyses of expert behaviour reveal how hard it is to externalise our own tacit knowledge and ways of thought (Schön 1984; Dix and Gongora 2011). Neither is it clear that we need to use models of self to understand each other; indeed, Gallagher and Zahavi, describing the problems of simulation theory, quote Wittgenstein, "*Do you look within yourself, in order to recognise the fury in his face?*" (Gallagher and Zahavi 2008, p 176; Wittgenstein 1980, §0.927).

This paper suggests that in fact, from a phylogenic development viewpoint, 'self' is actually complex and early social understanding may be more 'primitive', in the sense that a knowledge of other's thoughts may have developed earlier in our own and other species than knowledge of self.

2.3 Consciousness and consciousness of self

"*A self that's unaware of itself is an oxymoron.*" (Ramachandran 2004, p 97).

"*The only sort of consciousness we can describe, or even imagine, involves a sense of self.*" (Carter 2002, p 218).

For some writers, the consciousness of self is virtually equivalent with the term consciousness. For example, Rosenthal (somewhat circularly) defines 'conscious states' as "*simply mental states we are conscious of being in" and "what makes conscious states conscious is their causing higher-order thoughts that one is in those mental states*" (Rosenthal 1986).

The centrality of consciousness of self dates back to early philosophers. Notably, according to St Augustine, "*the mind knows itself to think*".[1] However, it is Descartes' "*cogito ergo sum*" that is the most well-known reference to consciousness of self, which for Descartes is not merely about defining consciousness, but the primary evidence for existence itself.

More recent philosophical investigation of the nature of self and consciousness includes a focus on higher order thinking (thinking about thinking) (Rosenthal 1986); embodied concepts of self as only really existing in activity with the world (Shanahan 2010); more phenomenological accounts (Gallagher and Zahavi 2008); as well as those who regard consciousness or self as a sort of illusion (Metzinger 2009).

There is also extensive psychological research on the nature of consciousness and of self. Examples of the former include work on understanding differences between conscious and unconscious processes at work in phenomena, such as blind sight (Carter 2002, p 19), and Libet's (2005) work on the delay between when the brain 'decides' to act and when a person feels they consciously made that decision. Examples of the latter include work on the mutability of body image when people come to believe a false hand is their own (Botvinick and Cohen 1998; Ehrsson et al. 2005) and studies of people with disorders of self (Gallagher and Zahavi 2008, p 208).

Neuroscientists and cognitive scientists also spend time trying to track the brain areas or higher level architectures that enable consciousness to work; for example, Goldberg's (2001) investigation of the role of frontal lobes or Damasio's (1999, 2010) work on body image.

2.4 Becoming conscious

One of the core questions of this article is about how we come to have the fundamental *cognitive machinery* for consciousness of ourselves; that is the ability to perceive our own thought processes, goals, and intentions, to be able to say, "I was just thinking", or, "I want to".

Having this explicit knowledge is not the same as thinking or wanting; these may be, and often are, purely tacit. The consciousness of ourselves as intentional allows us to make the tacit internal life explicit, whether to communicate to others or simply to be aware of ourselves. The fact that Descartes is thinking is evidence that Descartes exists, independently of whether he is aware of it, but for him to

[1] *De Trinitate*, Book X, quoted in Anscombe (1975).

say, "*cogito ergo sum*", he needs to be explicitly aware of his own thinking.

That being the case, we want to dig deeper than the ability to think and feel, towards the ability to look in on our own thoughts and feelings.

Note, this is not Chalmers (1995) 'hard problem' of consciousness.

Imagine relaxing outside on a summer day, a few small clouds passing across the blue sky; you are not explicitly conscious of your own thoughts, or paying attention to yourself, but you are still conscious, still seeing the sky even if you are not attempting to make sense of it or your thoughts about it. Nagel (1974) famously asked, "What is it like to be a bat?" While this surfaces many issues for consciousness (not least fundamentally different senses), it seems likely that the bat would not have explicit knowledge of its own thoughts and desires; that is, no consciousness of self. Yet, it could still be conscious and have an awareness of its environment, and if not a bat, then think of a cat, or a dog that you know.

This bare awareness of being is what Chalmers termed the 'hard problem'. This awareness is there at the embodied level of tacit action (the phenomenological experience of perception) and at the internal level (the phenomenological experience of self). The latter is possibly a natural extension of the former, and it seems likely that *if* we understood the hard problem as applied to the simpler, 'animal' side of our nature *and* we fully understood consciousness of self, then the higher level 'hard problem' would not seem hard. Of course for those philosophers who believe the hard problem does not exist, or can be dissolved by epiphenomenological sleight of hand, this is not an issue anyway.

Our scope in this article is purely the more 'computational' side of consciousness of self, which, while not irrelevant to the hard problem, is in no way dealing with it.

3 The evolutionary development of self

The standard evolutionary assumption is that things must either have a direct functional quality, or be some form of accident of a functional quality. This can sometimes be indirect, even to the point of apparently mal-adaptive traits that may occur due to runaway sexual selection (e.g., peacock tails). However, for something to develop, there needs to be a 'why', a benefit that improves fitness for survival.

However, it is not sufficient that a complex trait has a positive effect; there must also be a path of development. Without such a path, we end up with teleological arguments; for example, while the eye is useful for reading on a computer screen, this 21st century benefit cannot be the reason trilobites first developed rudimentary eyes in the Cambrian, 5000 million years ago (Parker 2003). Instead, we need to trace a potential development of any present-day phenomenon from simpler forms. This extends the 'why' question; there must also be a 'how': small steps, each of which, in its particular developmental niche, must have had survival value.

Evolutionary biologists use a combination of fossil evidence and reasoned arguments to examine the way current life developed. This form of study was originally focused on physical traits, from the eye to upright walking, but various authors, including Mithen (1996, 2007), Renfrew (2007) and Calvin (1990), use a form of *cognitive archaeology* to seek to understand human cognition as the result of a development through prehistory. Similar techniques have also been used by Tomasello (1999) in understanding the development of consciousness and by evolutionary psychologists studying current social cognition (Cosmides 1989; Tooby and Cosmides 1997).

Of course, we do not have direct evidence of cognitive evolution in the way that we do for physical development, just as we do not have direct recording of language before writing. Instead, these arguments depend on *plausible chains of development*, building more complex abilities upon simpler ones, bringing in, where appropriate, knowledge of present-day cognition, childhood development, and, if available, paleontological or archaeological data.

Focusing on consciousness of self, it is clear that having an idea of self makes it possible to write Hamlet's soliloquies or books about consciousness. However, just like the eye for seeing computer screens, advantages that are only manifest in complex society can only yield anachronistic or teleological arguments for development. If we are to have a useful understanding of how self developed phylogenically, we need to find appropriate incremental advantages for each stage in the development of self.

We will first look at some existing suggestions for the way self and consciousness may have developed and then the alternative model of the development of self as an accident of sociality; an argument that, the author would assert, is more credible than the standard 'I first' model of theory of mind.

3.1 Development models starting with 'I'

Damasio establishes strong arguments for the need to have a model or image of one's own body; as physical beings, understanding our physical nature can help us avoid danger or seek benefit (Damasio 2010, p 57). He also constructs a model of a consciousness more primitive than autobiographical self, but able to monitor its own feelings.

When looking at what consciousness contributes to the being, Damasio first suggests that consciousness helps self-regulation, but then admits that this argument is weakened when considering work on the extensive role

of unconscious action (not least Libet's work). He then falls back to the role self has in enabling language and culture. However, while this certainly demonstrates how self can be powerful in ramping up human development of a species once it is present, it is far less convincing as an argument for the drivers that caused it to emerge in the first place.

Tomasello (1999) discusses issues of self extensively when looking at child development, and he also sees this and other aspects of cognition as interwoven with the development of language and culture.

Other writers also see consciousness and self as emerging from language. Dennett (1993, p 195) suggests that our inner train of thought or stream of consciousness is derived from a short-circuiting of what was initially the overhearing of one's own speech. Ontogenically, the way small children (and indeed grown adults) talk aloud to themselves while doing complex tasks bolsters this argument (Shanahan 2010).

Many writers look to narrative as a way of conceptualising the way we establish a sense of identity (see review in Gallagher and Zahavi 2008, pp 200–202), and again, this has parallels in child development, where stories and story telling are central even to early infant experience including the landmark use of the word 'me'.

"*Stories are one of the fundamental ways in which we each create an extended self. The developing child's cumulative repertoire of stories gives him or her a sense of self across time and situation"* (Engel 1996).

However, while these uses of narrative and language are undoubtedly important, especially for establishing the autobiographic self, there appears to be a more primitive pre-linguistic sense of self and other, as Wittengstein's early quote suggests. Whether it is your apprehension of the emotion of others, or your awareness of your own eye movements over a scene, it does not appear phenomenologically that you are using anything linguistic.

Helen Keller was not deaf and blind from birth, but became so when she was less than 18 months, before the stage at which she would have attained the more complex linguistic structures associated with self-modelling. From that stage until Anne Sullivan taught her to communicate in full sign language at the age of seven, she had only 60 simple signs. However, in her biography, Keller describes her mental life during these years, including the following:

"*I think I knew when I was naughty, for I knew that it hurt Ella, my nurse, to kick her, and when my fit of temper was over I had a feeling akin to regret*" (Keller et al. 1905, Ch. II).

Note both her appreciation of Ella's feelings and an awareness of her own emotional being. Of course, this may involve an element of backward projection by the adult Keller, but to remember this incident suggests that it is not pure confabulation and that she did indeed have a relatively rich pre-linguistic notion of self and other.

3.2 An alternative view: self emerging from theory of mind

The view that knowing oneself is easy and precedes theory of mind at first appears natural, but we have seen that it has problems as we start to unpack it from an evolutionary standpoint. We will now explore an alternative phylogenic scenario for the emergence of a simple pre-linguistic self. Critically, this is a reflexive social development, where understanding of others precedes self-understanding. As mentioned, cognitive archaeology methodology cannot 'prove' a cognitive development scenario that, by its nature, leaves no physical trace. Instead, the aim is to create a plausible account by producing a series of steps, each small enough, and with adaptive value.

The final purpose of this paper is to apply this understanding to robotic and AI issues. As such, irrespective of whether this is *actually* how human consciousness of self developed, if the account is plausible, then it is a potentially valuable way to inform artificial cognitive development.

That said, the author would argue that this model of the development of self as an accident of sociality is more credible than the standard 'I first' account of theory of mind, which calls self into being with only far-future benefit.

3.2.1 Stage 1: reacting to the environment

The field is full of rabbits. As we approach, they initially stop feeding, their eyes fixed on us, and then, as we come closer, they run away. Even the simplest creatures react to their environment and the other creatures around them, whether as potential predators, potential prey, or maybe for the clues they offer to food sources.

Some of these reactions are purely instinctive, like our own startle response when we hear a loud sound. Others may involve more complex learning, as is seen in many larger mammals, where the juveniles learn from their mothers, or even in simpler animals, such as tits learning from each other to peck the shiny tops of milk bottles.

Figure 1 depicts this situation, focusing on the bear's level of cognition of the world. The human is hunting the bear, and the bear has seen the human. Although there are many debates about the nature of this representation, the bear is, in some sense, aware of the human and depending on the perceived threat may ignore, attack or run away.

3.2.2 Stage 2: predicting other creatures' reactions

Watch sheepdog trials: the shepherds call their dogs, a single command, a whistle or a word, and the dog circles

Fig. 1 Modelling the world. Figures 1, 2, 3, and 4 from McIntyre (1923), *The cave boy of the age of stone*

Fig. 2 Predicting other animals reactions. Figures 1, 2, and 3 from McIntyre (1923), *The cave boy of the age of stone*

round, bringing the flock back towards the fold. This sending round is one of the hardest things to teach, but the wonder is that it can be taught at all. One would imagine that the natural reaction of a predator in the wild, when it sees the flock of potential prey, would be to head straight for them. However, if the flock scatters, it is harder to catch any one of them. Successful wolf packs work as a team, some wolves circle round, and some come more stealthily in.

Again, some of these actions are purely instinctive, but a more successful predator, particularly a human hunter, has some understanding of the behaviour of its prey. In Fig. 2, the bear on the right has stage 1 thought processes (1), just as in Fig. 1, and may run if the hunter moves quickly. The successful hunter needs to have a model not just of the animal in front of him, but of its possible reactions, what is going through the animal's head (2).

This might mean moving more slowly to avoid scaring the animal away, or the opposite, jumping up, so that the animal runs away towards the hunter's mate, so that she can spear the prey.

Note that it is not that the hunters are imagining how they would react if they were prey, they simply have some sort of model of the prey's responses.

In some ways, this is a very primitive and asymmetric first-order theory of mind. The hunter has a model of the prey's mental processes. This might be purely reactive, but may include some level of motivation, drives, or intentions. For example, the hunter may wait by a waterhole, knowing that the prey will go there; in the case of hunting a bear, they may wait by a hollow tree with a bee colony within, or even, although this is a little more advanced, leave bait. This model may also include the prey's emotional states: an anxious prey is likely to be more difficult to get close to, and an angry bear more likely to fight back than run away.

The model of the prey's reactions and thoughts may include the way the prey responds to the physical actions of the hunters themselves. That is, there may be a level of reflexivity even at this stage, albeit purely with regard to the external physical actions of the hunter. In some ways, this is like the model that one has through procioception, nocioception, and seeing and touching one's own body.

Note that there is a difference between this level of model—that another creature reacts to one's actions—and a true physical de-egocentricism. Small children are able to know that adults and other children will react to their actions before the age when they can perform Piaget-style dolls-on-a-landscape non-egocentric-view tests (watch a small child fall and look around to see if anyone is within earshot before crying).

Similarly, experiments with chimpanzees have shown that they are aware of what others in their group have seen, and modify their behaviour accordingly (Hare et al. 2001). That is, they are using a form of theory of mind, at least in terms of the physical perceptions of other chimpanzees.

3.2.3 Stage 3: predicting other humans' reactions

We will now move on, probably many tens of thousands of years, and assume that humans or possibly pre-human

Fig. 3 Simple predictions about other humans—first-order theory of mind

hominids have well-developed proto-theory of mind as described above.

Consider what happens when humans interact with one another.

In Fig. 3, an early human approaches another.

The approaching man has a spear in his hand, so the sitting woman's first reaction might be to reach to grab the hand axe beside her in case he is aggressive. However, she is also aware that grabbing a weapon might make him angry or nervous and start a full-blown fight.

That is, she has a model of the approaching man, his mental state, and his potential reactions, just as she does of animals in stage 2.

However, they are not dealing with simple animals. They are each capable of having models of each others' mental states and intentions. Therefore, when you interact with another person, their mental state includes a model of your own. It will be clearly be advantageous if your own model can be more complete and in particular include the model the other person has of you.

Figure 3 illustrates that the approaching human's model of the sitting human includes the fact that she has a model of him. It is clear how this slightly more complex understanding could aid social interactions. If the sitting human picks up her axe, the obvious reaction for the approaching human might be to raise his spear. However, if he is able to think, "it's just because she thinks I am aggressive", he might instead lower it and defuse the situation.

3.2.4 Stage 4: from second-order theory of mind to self

So far, we have progressed through three stages of gradually increasing complex mental models:

1. Representations of the environment and creatures in it;
2. Predicting other creatures' behaviour by modelling their stage 1 thought processes; that is simple first-order theory of mind;
3. Including in models of other people their stage 2 models of one's own thought processes; that is simple second-order theory of mind.

This progression shows how second-order theory of mind, understanding other people's models of one's own mental state, can arise progressively from simpler predictive models of other creatures and other humans. Some levels of this model making we probably share with other creatures, such as pack hunters and those that exist in social groups; some are probably uniquely human.

The steps are each smaller than an *ex nihilo* emergence of self, and each has value even when only partly formed. Even partial changes between stage 1 and stage 2 would constitute a survival advantage; for example, awareness of what other creatures have seen, as in the chimpanzee experiments of Hare et al. (2001). Furthermore, we have evidence that at least some of these are present in creatures today.

Note that once they are at stage 3, the humans in Fig. 4 effectively have a model of themselves through the other person's eyes. Even without fully internalising this model of one's own thoughts, the early human can start to accrue some of the advantages of a conscious idea of self, posited by authors such as Damasio (2010). This then starts an evolutionary positive feedback loop, as better models of oneself start to correlate with internal feelings, with incremental improvements leading to incremental benefit.

Given such a model of our own thought processes from another's viewpoint, it is then a short step to effectively bypass the 'other', in the same way that Dennett (1988) imagines personal narrative of self emerging from speaking and hearing oneself, and over time bypassing the physical articulation of words.[2]

That is, because an early human (or maybe hominid) has a model of the other's model of 'me', the human develops a self-image of its own intentions: "I".

[2] Elsewhere Dennett (1991) creates a putative story of the development of self, using an argument not unlike the methodology here. His derivation leaps abruptly from self as tacit distinction from others to self-narrative and self-presentation. To some extent this paper bridges part of that gap.

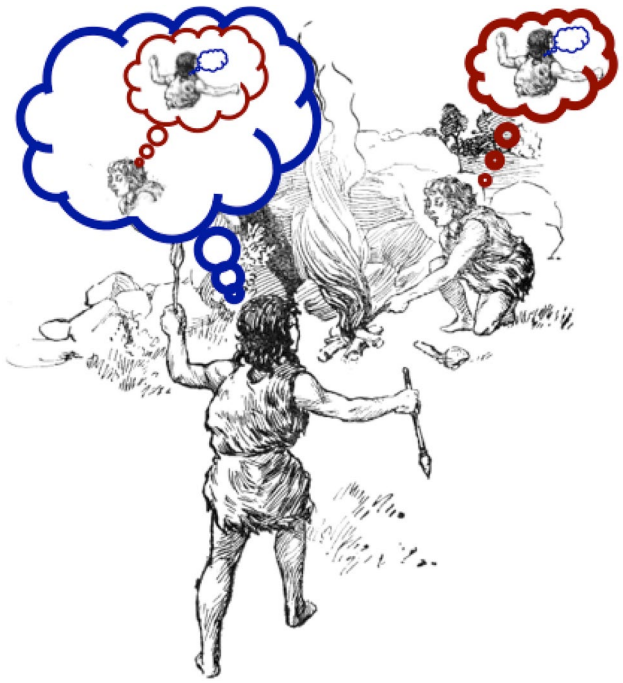

Fig. 4 Understanding other people's understanding of oneself—second-order theory of mind

4 Childhood development of self and other

The previous section developed a phylogenic argument for the development of self from social interaction. In this section, we will look at the way concepts of self and other emerge during childhood development, and see that there are parallels of this dialogical emergence of self ontogenically as well as phylogenically.

While, in its pure form, the Victorian adage "ontogeny recapitulates phylogeny" has been debunked, childhood development is still used as a touchstone when considering impossible to observe phylogenic development in areas such as language. While this is not a direct relationship, more complex structures are assumed to build from simpler ones both in evolutionary development and in childhood development.

As we have discussed, there are various meanings to 'self' and different awareness of self and other emerges at different stages of typical childhood development.

Physical self-exploration, touching and visually inspecting one's own body, happens for the youngest age (Gallagher and Zahavi 2008, p 207), and arguably aspects of tactile self-examination may even be pre-natal. Mimicry is also very early, starting typically with tongue poking, and the underlying mirror neurons that enable this are not only innate in humans, but have been shown to be present in other animals also. Richer notions of self and other emerge throughout childhood, and more complex higher order theory of mind is still developing in older children (Liddle and Nettle 2006), and possibly throughout adult life.

4.1 Three levels of other

Tomasello (1999, p 180) identifies three main levels of awareness of others, which have been adopted by other writers. The first, identified by mimicry and related behaviours, he terms understanding of others as *'animate beings'* and, as noted, is present from or soon after birth.

The next major stage, understanding of others as *'intentional beings'*, emerges at around 9 months. Gaze following and shared attention are key signs of this stage. This is uncommon amongst non-humans.

Tomasello's final stage is the understanding of others as *'mental beings'*; occurring at around 4–5 years; this is the point when children are able to perform basic theory of mind tests, in particular false-belief tests. There are variants of these; in the classic version, the child is told a story about Sally and Anne. Sally and Anne are in a room; Sally hides a marble in a basket, but while she is out of the room Anne moves it; then Sally returns to the room. The child is asked, where will Sally look for the marble? Younger children select the location to which Anne moved it, but older children are able to realise that Sally does not know this and so will look in the original basket.

While Piaget's classic Three Mountain test of egocentricism (Piaget and Inhelder 1956) suggests that a full ability to de-centre (at least in terms of visual viewpoint) is only developed at 6–8 years of age, Martin Hughes' version using dolls and 'hiding' from a policeman (a more realistic task than picture matching), suggests de-centring much earlier, around 3.5–5 years of age (Hughes 1975; Donaldson 1978); that is agreeing very closely with Tomasello's third stage. Note that we will see that even this may be conservative and some aspects of de-centring may emerge far younger still.

It is interesting that Hughes' study and others arising from Margaret Donaldson's work, which also yield younger developmental ages, all use more concrete versions of Piaget's classic tasks. Note too that even when dealing with an identical situation, the verbal description in the Kursaal Flyers' song, mentioned previously, is far harder to grasp than the visual action of the video.

4.2 Beyond and between

The Kursaal Flyers' song exposes the complexity of higher order theory of mind, which develops far later. Tests of higher order theory of mind suggest that while first- and second-order theories of mind are well mastered by 10 years, stories involving third order of mind

are only understood a little above chance, and fourth order (like "Little does she know") not at all (Liddle and Nettle 2006).

As well as these more advanced stages, there are also other aspects of understanding of others and self that emerge between Tomasello's second and third stages.

By 18 months, infants have some understanding that other people have behavioural goals: if an adult tries some task, but appears to fail, the child then completes the task—that is, the child has inferred the goals of the adult (Meltzoff 1995 in; Tomasello 1999, p 83). There is also evidence that chimpanzees have this level of theory of mind (Call and Tomasello 2008).

Parents and carers of young children will also know that at this toddling stage, when a child falls, she will often look to see if an adult is around. If so she will burst into tears, but if not she will get up without any apparent fuss. This is not to say that the distress is not real and that also, in some circumstances, the child may cry irrespective of the presence of others. However, just like the injured footballer when the referee is watching, the child's response of the tears may be magnified based on the perception that they will be heard by another and lead to external comfort. It is interesting that while this is part of the everyday experience of parents, the phenomenon does not seem to be discussed commonly in the development literature, presumably because of the ethical problems in performing experiments.

Around this same age (towards 2 years), Fernyhough (2008) notes the first signs that a child is able to detect mistaken belief in others, years before formal false-belief tests can be successfully performed. He also cites Bartsch and Wellman's (1995) work showing that 'contrastives', that is where the child uses language which expresses differences between beliefs and reality, emerge some time before the formal test. As noted earlier, the concrete use of concepts often precedes the more formal 'tests' for them.

When the author's own daughter was two and quarter years of age, part way between Tomasello's intentional and mental being stages, she was interviewed by a young linguistics student. During the discussion, the child said that she had been to the doctor with her baby sister. The student asked, "where did you go to the doctor?" From a theory of mind point of view, this is itself an interesting question. In the context, it was clear that the adult wanted to know whether this was a doctor in the hospital or a general practitioner/family doctor, but while the former would be reasonable to expect from a 27-month-old vocabulary, the latter would not! The child's answer appeared to be enigmatic, or maybe a non sequitur, "up the steps". This answer, however, would have been instantly recognised by every parent in the neighbourhood as the steps (with no ramp!) to the GP's surgery were a major problem when juggling prams, pushchairs, and young children.

Note here that child had no problem with the idea that the student did not know where the child had been; indeed, the child had volunteered that they had been to the doctor, tacitly assuming that this was not knowledge shared by the student. However, the child failed to understand that the contextual knowledge of the geography of the area was not shared. In other words, even at 27 months, we see quite complex first-order theory of mind for episodic experience (where it is obvious that the listener was not present), but not for propositional knowledge (where a shared context is tacitly assumed).

4.3 The knowledge of me

Tomasello also suggests that, at this stage, an infant might first start to become aware of 'me' as a concept as she becomes aware that an adult's gaze, which she has been following from object to object, falls on herself (Tomasello 1999, p 89). As confirmatory evidence of this hypothesis, he notes that around this stage (first birthday) are the first signs of shyness or coyness (Harter 1983; Lewis et al. 1989).

Note that this is a knowledge of oneself as a physical thing in the world, not knowledge of one's own mental processes. However, it is reflexive knowledge, quite literally seeing oneself in another's eyes.

Small children often refer to themselves in the third person, and books and websites offering advice to new parents explain that small children may struggle with pronouns, especially personal context pronouns, such as 'you' and 'I'. If Buber is right in emphasising the primacy of the I–You relationship for babies, this does not carry through into the ability to say words. Instead small children often refer to themselves by name: "Sammy go to park", "Gemma wants drink".

Note that this language is partly because parents' baby language often uses proper nouns to refer to themselves and the child. Indeed, parents' language is critical in encouraging the rate, if not the final endpoints, of children's development of concepts of their own mental states and those of others (Kirk et al. 2015).

4.4 Which comes first?

We have discussed a variety of elements of self and other as they emerge during childhood.

Physical exploration of one's own body, beginning prenatally, can be observed externally, as can other elements of physical awareness, such as recognition in a mirror. External social interactions, shared attention and the ability to interpret others' intentions can also be observed externally. However, internal knowledge of oneself is hard to observe before the child can talk about it. It is thus difficult

to distinguish language development from the development of ego and theory of mind.

The argument of this paper that individuality and consciousness of self emerge through sociality might at first seem odd, but, in fact, these abilities develop at exactly the same time in small children, a process developmental psychologists term developmental *synchronies*. (Marraffa 2011).

"*Children do not conceptualize their own mental states before they conceptualize the mental states of others* (Gopnik 1993), *nor do they talk about them earlier* (Bartsch and Wellman 1995)" (from Tomasello 1999, p 75).

Theory of mind has been particularly important to researchers in autism. Children and adults with autism have problems both understanding other people's beliefs and intentions, and also understanding their own emotional state (Baron-Cohen et al. 1985; Chari 2002). In her review of theory of mind and autism, Lantz summarises various evidence that autism, while typically not diagnosed until later, is in fact already apparent at some of the very early stages discussed above: evidence including absence of joint attention and treating people more like objects as opposed to intentional beings (Lantz 2002).

That is, when considering both the order and disorders of development, the emergence of self and the emergence of theory of mind are tightly bound. While this does not prove the emergence of self from sociality, it is certainly consistent and certainly does not support the common view that in some way self is 'easy'.

4.5 Language, knowledge and self

As noted previously, the most clear alternative development of self is through language. There are undoubted elements of interdependency; not least it seems hard to envisage the full development of an autobiographical self without interpersonal narrative. In addition, if, as suggested by Dennett (1988), the stream of thought is internalised verbalisation, then explicit awareness of this (e.g., "what was I just thinking about") can only come after the emergence of language.

However, the linguistic origins of self-related language presuppose at least a tacit theory of mind. In order to tell you stories I need to know that you don't know already (as in the story of the author's own daughter). The evolutionary account was presented without mentioning language, and it seems possible that the whole series of stages could be achieved with a pre-linguistic cognition.

Donald (1991) suggests three early levels of human and pre-human culture. The first is *episodic culture*,[3] effectively living in the present, and he suggests that apes, even chimpanzees, who have shown a level of theory of mind (Premack and Woodruff 1978; Call and Tomasello 2008), are at this level. In terms of self-awareness, this is likely to involve some level of representation of one's own body (as we have seen babies possess almost from birth), but not one's own intentional state. Donald's third stage is *mythic culture* and is associated with language. However, what he terms the 'missing link' is *mimetic culture*. This, he suggests, may have been present in more advanced pre-human hominids and involves a level of *non-verbal* mental imagery allowing relatively complex representation.

The phylogenic account in this paper would be entirely consistent with this mimetic level. Indeed, Keller's account of her awareness of causing her nurse pain and of the ensuing regret would bolster the suggestion that this is at least possible. However, it is also possible that the final stage emerged in parallel with language. Note, we are talking here about the emergence of cognitive *structures* that enable consciousness of self. It is quite reasonable that language is a pre-requisite for certain cognitive mechanisms to develop in humans as a species, but once these are present, they enable individuals to have a non-linguistic awareness of self and others.

Whether or not that proto-idea of self existed at this stage of development, the cognitive effects of language would undoubtedly create a positive feedback with notions of self and other. As noted, some pre-linguistic level of theory of mind is essential for narrative, but being able to name things and concepts then helps to solidify and objectify them, a process the author has previously called *trans-articulation* (Dix 2003). The author has also previously argued that, while we are not fundamentally constrained by our language, there is a tendency for words to shape even the physical artefacts around us to reflect the vocabulary we have as well as the more obvious vice versa (Dix 2009).

Crucially, language is essentially social even if it is also recruited for internal thought. That is, both the linguistic and pre-linguistic origins of awareness of self are fundamentally *reflexive* and related to spoken or unspoken *relationship*.

5 The artificial self

So far, this paper has argued that self emerges from the need for social understanding. This reflexive viewpoint suggests that models for artificial intelligence based on presenting and interpreting for others may be more fruitful than those starting with internal cognitive and emotional processes.

[3] Note that there is a slight tension in terminology with psychology of memory. Donald's 'episodic' culture is one without episodic memory; the latter is precisely the ability to recall previous episodes and thread them together in a historical narrative.

5.1 Consciousness and self

Friedenberg (2008) reviews a number of projects aimed at establishing a level of artificial consciousness. Some projects are bottom–up, focused on reproducing the low-level mechanisms and architecture of the brain and hoping that consciousness will spontaneously emerge. Other projects are working top–down, seeking to model explicitly the processes of consciousness and self. In establishing criteria for artificial consciousness, Friedenberg turns to Aleksander and Dunmall's five axioms for machine consciousness (Friedenberg 2008; Aleksander and Dunmall 2003): *depiction* (an internal representation of sensorial states), *imagination* (to recall past experiences), *attention* (selectivity in perception and imagination), *planning* (ordering future events to achieve a goal), and *emotion* (to evaluate and motivate action).

Interestingly, both these selected projects and the criteria are almost entirely egocentric, outward looking only insofar as the external world is the theatre of action and source of reward. Social interaction is left to Friedenberg's last chapter.

This is in marked contrast to Fernyhough, developmental psychologist and parent, who charts meticulously the intellectual growth of his own daughter in the light of current scientific knowledge (Fernyhough 2008). The primacy of human contact and interaction is a constant theme in his accounts. In words that strike so many chords with Buber's writing:

"*She is built to talk silently about love, in comfortable close-up. Smooching with a baby is so rewarding, partly, because it is the one thing they can really do"* (Fernyhough 2008, p. 46).

Remember too the operation of mirror neurons, which enable babies to copy tongue movement from the first few days after birth: sociality is innate.

Breazeal's (2002) description of the design of Kismet, a 'social robot', does take human contact as critical and amongst her five key characteristics are being *human-aware*, including a theory of mind to understand others, and *being understood*, allowing emotions and desires to be apparent in facial expression and actions. For the latter Breazeal emphasises that to be understood, a robot should *'understand its own self'*; however, this falls short of an explicitly reflexive model of self.

What, then, would a fully reflexive model of robot self look like? Whilst Friedenberg's (2008) review found bottom–up neural models and top–down cognitive models; we need something more outside-in, starting with the other.

For the robot's physical (or virtual) body, a reflexive model is not necessary for simple social actions. If the robot wishes to be near another person, then it initiates movement towards the person and 'knows' that this will bring it closer.

At a more complex level, a robot that wants (but need not know that it wants) human attention, may wave its arm, knowing that this brings attention.

So far, these are still instrumental actions, things that either directly cause a desired state, or in a simple predictable way cause another to perform an action with the desired effect.

The next step is to emulate the child who is able to interpret the intentions of others: a robot that if it sees a human struggling to perform a task, infers that the human wants to achieve a goal and if possible acts to help that goal come about. For example, if, while building blocks together, the human appears to be collecting red blocks, the robot may push a red block into the human's pile.

For this, the robot needs motivations (emotions/desires/wants) and also has to recognise goals in others (but not those of itself). Critically, one such motivation needs to be some sort of collaborative drive. This may be a special drive, or might be an outcome of mirroring, close to Damasio's "*understanding the actions of others by placing ourselves in a comparable body state*" (Damasio 2010, p 104), or empathy. That is, if I see someone struggling to achieve a goal, and I have no other conflicting desires, I have a 'borrowed' desire for that goal; if I see someone happy or sad, or predict that they will be so, then I feel a little of that emotion. While the latter sounds complex, levels of empathy and judgement of others' actions are present from the earliest age (Zahn-Waxler et al. 1992; Hamlin et al. 2010).

The final step is to create simple robot models of humans that include the humans' models of the robot. At this point, the robot would have a model of itself, just as in Stage 4 of the evolutionary development of self. While this sounds more complex than simply building a model of self, remember that artificial motivations and drives are likely to be modelled by neural networks, or complex Bayesian rules, no more immediately apparent to the artificial mind than our unconscious is to us. A self-aware robot built, apparently, more 'simply' would in some way have to re-represent these 'implicit' structures in an explicit declarative form. The apparently more complex reflexive model can work from the start with a more categorical (or pre-categorical) representation.

5.2 Emotion

These reflexive models will include emotions, of others and of oneself.

William James, the father of the psychology of emotion, saw emotion as a form of appraisal of bodily state, "*My thesis … is that bodily changes follow directly the PERCEPTION of the exciting fact and that our feeling of*

the same changes as they occur IS the emotion." (James 1884). While this idea has been updated and debated in many ways (see LeDoux 1998, ch.3, for a review), elements of it are present in many current theories, and indeed, basic emotional responses, such as the startle response to loud noises, clearly happen faster than conscious emotion.

This is one of the reasons why a bursting balloon can be so funny: the startle response raises adrenalin levels, but then higher level appraisal works out "it's all right", so you are left with positive conscious valence, but high bodily arousal—hence a belly laugh.

From an artificial modelling point of view, adopting a fully reflexive approach would not try to accomplish that appraisal directly, but indirectly through the impact on others. A reflexively emotional robot would have low-level emotional circuitry (arousal, valence), but then, at a higher level, model the emotions of others and their emotional perception of the robot itself.

This does not mean that our high-level conscious thinking cannot affect lower level emotion. Regret involves quite high-level counter-factual reasoning about the situation as it is, and how it could have been under alternative actions by the agent. If these alternative scenarios are better or worse than the actual outcome, this modifies the emotional response positively or negatively, which in turn 'tunes' low-level learning. When a cognitive model of regret inspired a computational model, this improved the rate of game learning by a factor of 5–10 times (Dix 2005). However, note that in this case, the conscious mind is not controlling emotion directly, but instead engaging in mental actions that affect emotion.

The reflexive approach means that a robot's explicit model of its own emotional state need not correspond to that state. Imagine a 'fearful' robot in a difficult situation (standoff with another robot!). The robot might actually take on an aggressive demeanour: doing this would create an impression of confidence to others and thus increase the robot's drive for safety. A robot with such a reflexive emotional appraisal might even 'believe' this self-image: bravado, just like many adolescents, and, for that matter, adults.

5.3 Deceit and the Turing test

The point of self-deception is that, unless one is sufficiently reflective, one is not consciously aware of it. However, plain deception is often used as a litmus test for theory of mind—you have to understand that someone has beliefs, that those beliefs need not be the same as your own or reality, and that you can manipulate those beliefs.

There are many non-Machiavellian reasons for deceit: play, politeness, privacy, a secret surprise, as well as conflict situations. We use 'white lies' to oil the wheels of social activity, as bare truth is often rude.

Nijholt (2011) reviews various works that embody an element of deceit or non-cooperative behaviour as well as some of the reasons why we might wish to have deceitful artificial agents. Indeed, the Turing Test itself can be seen as a form of deceit, as the computer is *pretending* to be human.

5.4 Who decides?

In the classic Turing Test, the computer is deemed intelligent (if not conscious) when the human interrogator cannot tell it from a human. That is, it is deemed intelligent by a human. From the point of view of consciousness or awareness of self, we could adopt the same position. Ruth Aylett sees robots as becoming social actors when people view them as such, "*we treat them as if they are real characters, as if they really had dreams and hopes like us*" (quoted in Shukman 2015); Tony Prescott (2015) says of his iCub robot child, "*Sometimes it even leaves me with the surprising feeling that 'someone is home'*".

In the film Ex Machina (2015), Caleb is recruited to perform a kind of Turing Test on Ava, an android created by the software magnate Nathan[4]. As a Turing Test, it is non-standard, since Caleb knows Ava is not human; however, the crucial issue is (ostensibly) not just whether 'she' is intelligent, but whether she is conscious. Over the course of a number of meetings, Caleb begins to feel affection for Ava, and, furthermore, believes that she is falling in love with him. In fact, it transpires that she is deceiving him and instead merely pretending, a fact that Nathan seizes with excitement (well, until he is killed), as evidence of her complex emotional understanding—second-order theory of mind and consciousness.

Although it all ends badly (albeit with characters for whom it is hard to feel sympathy), this may well be the right arc if we wish to create truly emotional robots: not simply a robot that by some computational mechanisms can be said to love, but one that can know that it is loved by another, and know that the other knows themself to be loved.

[4] Only the ostensibly central storyline is considered here; the film has been criticised for falling back on classic sexist tropes such as the *femme fatale* and picking up on old but still persistent caricatures of the female, like the robot, as not fully human (Watercutter 2015).

6 Reflexive ethics

1. *A robot may not injure a human being or, through inaction, allow a human being to come to harm.*
2. *A robot must obey the orders given it by human beings except, where such orders would conflict with the First Law.*
3. *A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws.*

(The Three Laws of Robotics, Isaac Asimov 1950).

Asimov's Three Laws of Robotics have become ubiquitous not only in discussions of the governance of robots within science fiction, but also in society at large. Currently, robotic systems obey lower level computational rules, but as systems become more autonomous, some form of more interpretative rules will become necessary. Asimov (1950) predicted the problems this would bring, and created scenarios where, even assuming highly intelligent 'positronic' robot brains, unforeseen consequences emerge.

Asimov's laws are fictional but remarkably prescient, having been drawn up in the days when only the earliest computers were being developed; since then, they have been both widely quoted and critiqued (e.g., Singer 2009). Today, with more than 60 years' experience of digital computation, senior academics and major industry players, including Google and Microsoft, have tried to create laws that are more pragmatic, aimed at designers, rather than robots themselves (EPSRC 2014; Nadella 2016; Amodei et al. 2016). Leading scientists have also called for a "*ban on offensive autonomous weapons beyond meaningful human control*" (Hawking et al. 2015), attempting to preclude such autonomy; however, it is hard to believe that this apparently moderate declaration will not be called into question, for example if an autonomous robot is shown to more accurately and effectively distinguish hostage from terrorist in siege situations.

Drawing on Piaget (1932) and Damon (1983), Tomasello (1999, p 180) argues that it is at the stage at which children develop *empathetic understanding* of others that they move from *rule following* to true *moral* reasoning.

Various forms of the 'Golden Rule' exist across world religions from Hinduism to Humanism, Confucianism to Christianity: sometimes stated negatively, "do not do to others"; sometimes positively, "do to others"; sometimes in terms of love, "love one another".

"*One should also behave towards all creatures as he should towards himself*" (*Mahābhārata*).

"*The man who loves himself so much, Should do no injury to others.*" (*Udâna*).

"*love thy neighbour as thyself*" (*Leviticus* 19:18).

"*Do to others as you would have them do to you*" (*Luke* 6:31).

This principle is sometimes included within the broad philosophical/social principle of reciprocity, for example, the following from The Analects of Confucius:

Tsze-kung asked, saying, "Is there one word which may serve as a rule of practice for all one's life?" The Master said, "Is not RECIPROCITY such a word? What you do not want done to yourself, do not do to others." (The Analects of Confucius).

However, reciprocity can also include tit-for-tat punishment as well as these more benevolent reflexive principles.

Elements of the Golden Rule often form a key part of moral arguments, "how would you like it if X did that to you", and variations of this "getting into another's shoes" moral reasoning are found in Rawls' (1971) *'veil of ignorance'*, which asks how one might like society to operate if one did not know which person one was going to end up being.

It is also used as a justification of universal human rights, although Hardwick's (2012) critical review of theoretical groundings of human rights suggests various weaknesses in this, not least the longstanding issue of who precisely counts as 'others': slaves, women, children, people of different races? More recent debates around the same issue have concerned those with brain death, the unborn, animals, and quite critically at the moment of writing, those of different religions.

At some point in the future, society may need to consider the issue of robot 'rights', especially when considering 'social robots'. In 2008, Whitby called for professional codes of conduct to be modified to deal with issues of abusive behaviour to artificial agents; not because these agents have any moral status in themselves, but because of the potential effect on other people. These effects include the potential psychological damage to the perpetrators of abuse themselves and the potential for violence against artificial agents to spill over into attitudes to others (Whitby 2008).

In the TV series *'Humans'* (2015), human-like androids, called 'synths', are built to perform mundane tasks, but a small number are sentient. Niska, one of the synths, has been enslaved (or simply installed) in a form of android brothel for human customers. As she escapes, she says to the woman who runs the establishment, "*everything they do to us, they want to do to you*".

This is a fictional and hypothetical scenario; facts, however, are catching up with fiction. In Canada, a man was charged with procuring child pornography after ordering a child-sized sex doll from a Japanese company (Rutkin 2016). The laws on this currently differ dramatically between countries, and ethicists debate whether the effects are positive, acting as a form of therapy (as the manufacturer believes), or negative, likely to encourage real abusive

behaviour. To further complicate this picture, in Shinto, according to the designer of the dolls, the dolls themselves are considered to have a form of soul (Morin 2016).

Darling (2012) suggests that we may soon have to consider legal rights for social robots and agents, akin to those for corporate entities and animals. Darling argues that it is the anthropomorphic nature of robots that make them special, citing examples where a child would be hurt to see a robot pet damaged, and where even a military general felt compassion for a battlefield robot.

At a keynote at Web Science 2015, Mia Consalvo presented her study of the growth and demise of Faunasphere, an online simulated-world game (Consalvo 2015). The creatures, 'fauna', in the simulated worlds were like pets, rather like a sophisticated version of Tamagotchi. The players became very attached to their fauna, which they fed, played with, and generally looked after. The extent of this became most obvious when the game was closed down. In the conference auditorium, many of the audience were brought to tears at the description of the distress of the players, some taking time off work to be with and tend their fauna during their last hours.

These examples point to an emerging first-order definition of an agent as an *object of ethical responsibility*: an agent has ethical/legal rights when we, as humans, feel that it has.

The next stage from this is when agents are able to have some model of the way that they are treated. It is increasingly common to use forms of chatterbot as first-line response systems. It has been found that users of such systems adopt verbally abusive language to a greater extent than they would do to either a real human operator or a non-anthropomorphic computer system (Brahnam 2005; De Angeli and Brahnam 2008). Potential solutions to this include having agents that recognise and respond to inappropriate language (Brahnam 2005). For areas such as customer support, this would normally be by deflecting the conversation back to appropriate topics, or, in extremis, ending a call.

Social robots in the home cannot simply 'end the call', so one could imagine that a robot, on detecting abuse, might respond in language that emulates being affronted or distressed. As discussed above, there is arguably no hard difference between emotion and emulated emotion; however, whether or not the agent was 'really' hurt, this could not help but increase the sense of ethical responsibility of the human who caused the 'distress'. That is, whether or not philosophically justified, socially and emotionally for the human involved, the extent to which an artificial agent is aware that it is mistreated (second-order reflexive) alters the ethical severity of that mistreatment.

Having discussed robots as an object of ethical responsibility, let us return to the Laws of Robotics and the Golden Rule in relation to robots as *agents of ethical responsibility*. Whether or not robots or virtual agents are protected by law, can they be prosecuted by law? When a robot behaves badly, is it blameworthy itself, or simply its designer or owner?

When discussing the potential legal protection of social robots, Darling (2012) cites animals and companies as examples of non-human entities afforded protection or rights under many jurisdictions. However, these differ in terms of parity. In the medieval period, animals could be tried and even hung (Carson 1917), but nowadays, it is the owners of animals who are prosecuted, although animals may be non-punitively destroyed for the protection of others. In contrast, companies, while non-human, have both legal protection (e.g., patents) and responsibility (can be sued).

This said, while ethical language may be used about companies, when probed most people would not see the company itself as ethically responsible, but rather those on the board of directors, or carrying out the problematic actions. That is, we distinguish ethical and legal agency.

The legal issue is far from theoretical, as autonomous cars and drones are soon likely to be 'released'. Insurance is still a major barrier, and the responsibility for accidents and any ensuing harm will ultimately be decided in the courts.

Ethically, we could start with an 'outside-in', first-order definition, that is one might regard that a robot or AI is an ethical agent when we see it as such. Arguably, we do this already with other humans. A sociopath does not have the same innate empathy as most people, and may be given a level of leeway in terms of norms of social behaviour by friends and acquaintances, but is still expected to uphold the same standards of social conduct, and ultimately may be held legally accountable.

For an artificial agent or robot, this does end up tied closely to the issue of being an object of ethical consideration. If a car is defective, it is scrapped—we may keep it off the road in a museum for its historic value, or even in a garage for personal nostalgia, but we feel no sense of guilt in destroying it or depriving it of freedom to be on the road. As long as this is true for a robot, we do not have to face the issue of ethical agency. However, at the point at which turning a machine off or destroying it is deemed ethically problematic, then, if things do go wrong, we need to ask whether it 'deserves' to be so punished.

This is still all centred on society's view of the agent.

A *fully reflexive* view would start with second-order emotions, such as obligation and shame; not merely "what are the material effects of my actions on others?", but, "how will others consider my actions?", or even, "how will others consider I considered my actions?". Buber's "I and Thou" is deeply theological, and the idea of personal immanence is deeply rooted in the Judeo-Christian conception of God,

Fig. 5 Venus of Willendorf, c. 24,000–22,000 BCE (Wikipedia 2015)

"You have searched me, Lord, and you know me. … you perceive my thoughts from afar" (Psalm 139); just a short step to an internalised 'ought'. An AI or robot becomes an ethical agent, not so much when others perceive it as such, but when it perceives that others perceive it so.

One might deliberately include such reflexive models in order that an autonomous robot can make appropriate decisions about actions. However, they might also arise more 'accidentally' as a result of other goals; for example, making the robot more emotionally expressive.

As mentioned earlier, in an attempt to validate a cognitive model of regret, the author created a computational model, and found that this improved learning (Dix 2005). While the object of regret in this simulation was non-social (rewards in a card game), it would be quite reasonable to use this or other modelled emotions, such as shame, to improve learning and interaction in social contexts.

That is, we may not be far from machines that know they would be perceived as doing wrong, which is, arguably, very close to what it means to do wrong.

7 Buber's primal You

This paper is set within the context of a special issue focused on Buber's "I and Thou". However, Buber's writing is not purely philosophical, but more spiritual, theological, or maybe religious. Coming from a Jewish tradition, Buber's imagery naturally draws on Jewish legend (p 76), but it also mentions the crucifixion (p 67) and references Brahman and Egyptian spirituality (p 71). However, this is not a writing of dispassionate comparative religion, but an impassioned call to return to the You in one another and the 'one being'. For Buber, to be human is to participate in the inter-personal at the deepest level (p 85).

In the Greek and Roman Pantheon, the gods are seen as remote, treating humans instrumentally, and, likewise, the goal of temple and sacrifice is to placate, in Buber's terms an I–It relationship. There are exceptions; although most transgressions, except the solely sexual, end badly. Notably, Prometheus, enduring daily punishment for his compassion in giving shivering humans divine fire, is cast as the suffering hero, albeit hard not to read nowadays without overtones of Blake's Christology.

In contrast, the Judeo-Christian tradition, of which Buber is a part, sees God as different, but not always distant. This is perhaps most obvious in the call to a 'personal relationship' in evangelical preaching and the Westminster Shorter Catechism's declaration that the chief end of man (*sic*), is "*to glorify God, and to enjoy him forever*" (emphasis added). However, in tone, Buber's Jewish mysticism is closer to the contemplative traditions in Catholicism and Orthodoxy.

7.1 Spirituality of the artificial?

In *'Humans'*, Max, the most sensitive and childlike of the sentient 'synths', finds himself alone. Earlier in the episode, he has seen some roadside flowers with a written message, and, presumably inspired by this, he drops to his knees and prays artlessly to God; a God whom Max is neither certain exists nor, if He does, that He would care about a synth like him.[5]

Is this credible? Can the mechanical encounter the ineffable?

Max's spiritual encounter is effectively an accidental outcome of his sentience. Given the universality of some form of religious belief across all cultures, might some level of spiritual sensibility arise naturally from any form of artificial consciousness? In Arthur C. Clarke's *'Fountains of Paradi*se' (1979), earth encounters a convenient passing AI in a deep space probe. While somewhat irreligious itself, the AI confirms the theoretical exo-sociologists' hypothesis that religion is largely the preserve of species with nuclear families. While a fictional extra-terrestrial artificial anthropologist is not the best evidence, this does raise the issue of how deeply spirituality is tied to material form as well as cognitive capacity and culture.

Buber puts great emphasis on the first relationship of the child with its mother inside the womb. Certainly, the heavy breasted Paleolithic 'Venus' figures (see Fig. 5) suggest early close associations between spirituality and the maternal bond. However, for Buber, this is not the source, but more the first attachment of a more primal "*longing for relationship*" (p 78), so that the "*development of the child's soul is connected indissolubly with his cravings for the You*" (p 79).

[5] Hard not to be reminded of Luke 18:14 here.

Attempts at more human-like robotics, such as Kismet (Breazeal 2002) or iCub (Metta et al. 2010), often include a number of underlying drives. This paper has suggested that inter-personal reflexive processes are a fruitful path to more self-aware robotics. To kick-start these processes, the underlying drive would be precisely that basic "*longing for relationship*", exactly what Fernyhough (2008) observes in his 'smooching' child. If Buber is right, then this may well lead to potentially spiritually aware robotics.

7.2 The mark of humanity

Pope Paul III's *Sublimis Deus* (1537) forbade the enslavement of Native South Americans. The debate leading to this and the Bull itself depended critically on the argument that the natives had a soul and were thus fully human. From Aristotle on, arguments for slavery and various forms of racial, sexual, and social discrimination have often relied on the assertion that some group have no soul or lesser soul and are not fully human (Heath 2008). Aristotle's 'soul' is probably close to current 'consciousness' or 'mind', and it seems likely that the point at which we recognise this in artificial creatures may be the point at which we really have to face some of the ethical challenges of robotics.

However, taking more fully the reflexive stance of this paper, does a robot become spiritual when we regard it as having a soul, or when it becomes aware that we regard it as such?

Given the existing power of platforms, such as Google and Uber, to direct humans' day-to-day lives, the current debates about autonomous weapon systems (Hawking et al. 2015) and the public statements by eminent scientists that self-aware AI is humankind's greatest danger (Hawking et al. 2014; Cellan-Jones 2014), maybe the crucial question in the end may not be whether we recognise the soul in the machine, but whether they recognise the soul in us.

7.3 Possible or probable

Discussing computational spirituality seems even more speculative than artificial consciousness or the ethics of robots. However, this may emerge sooner than at first appears.

First, there has been substantial work looking at cognitive, social, evolutionary, and even genetic accounts of religion (Kirkpatrick 2004; Dow 2008; McNamara 2014). To date, this has been largely theoretical, but with ever increasing interest in understanding radicalisation, it is likely that rich cognitive and epidemiological models will be developed, both in the open academic literature and behind closed doors. In some ways, this is more about religiosity (which Buber regards as a form of I–it experience, see p 65) rather than spirituality.

Second, recent advances in artificial intelligence have prompted initiatives to make it more accessible and comprehensible to humans, and more able to communicate with humans. For example, the UK funding agency, EPSRC, which includes computing, is planning a programme in human-like computing (Dix 2016; EPSRC 2016). If artificial agents and robots are to understand humans, then this will need to include modelling religious and spiritual experience.

Finally, Buber describes three spheres of relationships: with the physical world and animals, with other humans, and with the divine (p 57). For Buber, it is the I–You of the third that underlies all authentic encounters with the first two, and the I–You of the first that is the foundation of art. If Buber is right, then artificial spirituality may not simply be a speculative fancy, but instead an essential underpinning of any form of artificial creativity or consciousness.

8 Summary

This paper has proposed that the cognitive machinery for consciousness of self emerged as a by-product of the construction of second-order theory of mind. Both this view and more common linguistic accounts of the origin of consciousness point to a reflexive notion of self as an accident of sociality. This has resonance philosophically with Buber's primacy of 'I-and-Thou' over 'I'. It also has practical consequences when we consider the construction of artificially conscious and emotional robotics, and for emerging issues of ethics.

We saw that whether a robot or artificial mind could be the *agent* of ethical responsibility was intimately tied to whether it was an *object* of ethical responsibility, and both are related to its level of emotional fidelity. Furthermore, an artificial agent's ethics and emotion can be seen in reflexive terms: being aware of how it appears emotionally to others; being aware of whether it is held to ethical account by others for its actions.

If artificial self and consciousness is problematic, then artificial spirituality is doubly so, and yet, for many people, these are intimately connected. Buber's I-and-You is about the relationship of people to each other and the world, but also to the 'one being'. It is hard to imagine ethics without 'ought', and for many 'should' is intimately related to religion. Yet, it is equally hard for those from a Judeo–Christian tradition, such as Buber, to conceive of theology without love.

Many of these issues are still some years away for artificial agents and robots, but by considering what it means for a machine to be spiritual, ethical, emotional or conscious, we may better understand human selfhood, seeing how it

can be more deeply embodied in our social presence and openness to the other.

Acknowledgements Many thanks to the reviewers whose insightful comments have helped improve this paper and also suggested potential avenues for future work, not least a science fiction reading list.

References

Aleksander I, Dunmall B (2003) Axioms and tests for the presence of minimal consciousness in agents. J Conscious Stud 10(4):7–18

Amodei D, Olah C, Steinhardt J, Christiano P, Schulman J, Mané, D. (2016) Concrete Problems in AI Safety (v2). ArXiv.org. arXiv:1606.06565v2

Anscombe G (1975) The first person. In: Guttenplan S (ed) Mind and language: Wolfson College Lectures 1974. Oxford University Press, Oxford, pp 45–65

Asimov I (1950) I, Robot. Gnome Press, New York

Baron-Cohen S, Leslie A, Frith U (1985) Does the autistic child have a "theory of mind"? Cognition 21:37–46

Bartsch K, Wellman H (1995) Children talk about the mind. Oxford University Press, Oxford

Botvinick M, Cohen J (1998) Rubber hands 'feel' touch that eyes see. Nature 391:756

Brahnam S (2005) Strategies for handling customer abuse of ECAs. In: Proceedings of the Workshop on Abuse: The dark side of Human-Computer Interaction, pp 64–67. http://www.agentabuse.org/Abuse_Workshop_WS5.pdf. Accessed 12 Feb 2017

Breazeal C (2002) Designing Sociable Robots. MIT Press, Cambridge. ISBN:9780262524315

Buber M (1923) I And Thou. (Tr. Kaufmann, W. (2008)). Touchstone, New York

Burling R (2005) The talking ape: how language evolved. Oxford University Press, Oxford

Call J, Tomasello M (2008) Does the chimpanzee have a theory of mind? 30 years later. Trends Cogn Sci 12(5):187–192. doi:10.1016/j.tics.2008.02.010

Calvin W (1990) The Ascent of Mind: Ice Age Climates and the Evolution of Intelligence. Bantam. http://www.williamcalvin.com/bk5/bk5.htm. Accessed 12 Feb 2017. ISBN: 0-595-16114-6

Carson H (1917) The trial of animals and insects. A little known chapter of Mediæval Jurisprudence. Proc Am Philos Soc 56(5):410–415

Carter R (2002) Exploring Consciousness. University of California Press, California

Cellan-Jones R (2014) Stephen Hawking warns artificial intelligence could end mankind. BBC News/Technology, 2 December 2014. http://www.bbc.co.uk/news/technology-30290540. Accessed 12 Feb 2017

Chalmers D (1995) Facing up to the Problem of Consciousness. Journal of Consciousness Studies 2:200–219

Chari S (2002) The "Theory of Mind" Hypothesis: Explaining Autism. AutismUSA.net. http://www.autismusa.net/papers-theory-of-mind.html. Accessed 29 Sep 2014

Clark A 1998 Being there: putting brain, body and the world together again. MIT Press, Cambridge

Clarke A (1979) Conversations with starglider, Chap. 16. In: The fountains of paradise. Victor Gonzalez Ltd, London

Confucius (500–200BC) The Analects of Confucius, Lunyu XV. 24. (415) Tr. James Legge. Text at: http://wengu.tartarie.com/wg/wengu.php?l=Lunyu&no=415. Accessed 12 Feb 2017

Consalvo M (2015) Players and their pets: An online game from beta to sunset. http://www.slideshare.net/miaconsalvo/part-conditionconsalvo1. Accessed 12 Feb 2017

Cosmides L (1989) The Logic of social exchange: has natural selection shaped how humans reason? Studies with the Wason selection task. *Cognition*, **31**:187–276

Damasio A (1999) The feeling of what happens: body and emotion in the making of consciousness. Harcourt, San Diego

Damasio A (2010) Self comes to mind: constricting the conscious brain. Heinemann, London

Damon W (1983) Social and personality development. Norton, New York

Darling K (2012) Extending Legal Rights to Social Robots. *We Robot Conference*, University of Miami, April 2012. http://ssrn.com/abstract=2044797. Accessed 12 Feb 2017

De Angeli A, Brahnam S (2008) I hate you! Disinhibition with virtual partners. Interact Comput 20:302–310

Dennett D (1988) "Why Everyone is a Novelist," The Times Literary Supplement, September 16–22, 1988, 4, 459; reprinted in B. Cooney, ed., Philosophy of Mind, Jones and Barlett, Nov. 1996. http://dl.tufts.edu/catalog/tufts:ddennett-1988.00007. Accessed 12 Feb 2017

Dennett D (1991) The Origins of Selves," *Cogito*, 3, 163–73, Autumn 1989. Reprinted in Daniel Kolak and R. Martin, eds., Self & Identity: Contemporary Philosophical Issues, Macmillan, 1991

Dennett D (1993) Consciousness explained. Penguin Books, London

Descartes R (1759) A Discourse Of A Method For The Well Guiding Of Reason, And The Discovery Of Truth In The Sciences. Part IIII. (English Tr.) London, Thomas Newcombe, 1759. http://www.gutenberg.org/files/25830/25830-h/25830-h.htm. Accessed 12 Feb 2017

Dix A (2003) Articulation and Trans-articulation. (online essay) http://alandix.com/academic/essays/transarticulation.pdf. Accessed 12 Feb 2017

Dix A (2005) The Adaptive Significance of Regret. (online essay) http://alandix.com/academic/essays/regret.pdf. Accessed 12 Feb 2017

Dix A (2009) Paths and patches: patterns of Geonosy and Gnosis. Chapter 1. In: P. Turner, S. Turner, E. Davenport (eds) Exploration of space, technology, and spatiality: interdisciplinary perspectives. Information Science Reference, pp 1–16. http://www.hcibook.com/alan/papers/paths-and-patches-2009/. Accessed 12 Feb 2017. ISBN: 978-1-60566-020-2

Dix A (2016) Human-Like Computing and Human–Computer Interaction. Proc. Human Centred Design for Intelligent Environments (HCD4IE) Workshop. HCI2016. http://www.alandix.com/academic/papers/HCD4IE-2016-human-like/. Accessed 12 Feb 2017

Dix A, Gongora L (2011) Externalisation and design. In: Proceedings of the Second Conference on Creativity and Innovation in Design (DESIRE '11). ACM, New York, pp 31–42. doi:10.1145/2079216.2079220

Donald M (1991) Origins of the modern mind: three stages in the evolution of culture and cognition. Harvard, London

Donaldson M (1978) Children's minds. Fontana/Croom Helm, London ISBN 0-85664-759-4

Dow J (2008) Is religion an evolutionary adaptation? J Artif Soc Soc Simul 11(2):2. http://jasss.soc.surrey.ac.uk/11/2/2.html. Accessed 12 Feb 2017

Ehrsson H, Holmes N, Passingham R (2005) Touching a rubber hand: feeling of body ownership is associated with activity in multisensory brain areas. J Neurosci 25(45): 10564–10573. doi:10.1523/JNEUROSCI.0800-05.2005

Engel S (1996) Storytelling in the first three years. Zero to three journal, December 1996/January 1997. http://www.zerotothree.org/child-development/early-language-literacy/the-emergence-of-storytelling.html. Accessed 12 Feb 2017

EPSRC (2014) Principles of robotics. The engineering and physical sciences research council, Swindon, UK. https://www.epsrc.ac.uk/research/ourportfolio/themes/engineering/activities/principlesofrobotics/. Accessed 12 Feb 2017

EPSRC (2016) Human-like computing: report of a workshop held on 17 & 18 February 2016, Bristol, UK. The engineering and physical sciences research council, Swindon, UK. https://www.epsrc.ac.uk/newsevents/pubs/humanlikecomputing/. Accessed 12 Feb 2017

Ex Machina (2015) Directed by Neill Blomkamp [Film]. DNA Films, Film4. http://www.imdb.com/title/tt0470752/. Accessed 12 Feb 2017

Fernyhough C (2008) The Baby in the mirror: a child's World form Birth to Three. Granta

Friedenberg J (2008) Artificial Psychology. Psychology Press, Hove

Gallagher S, Zahavi D (2008) The Phenomenological Mind. Routledge, New York

Goldberg E (2001) The Executive Brain: Frontal Lobes and the Civilised Mind. Oxford University Press, Oxford

Gopnik A (1993) How we read our own minds: The illusion of first-person knowledge of intentionality. Behav Brain Sci 16:1–14

Hamlin J, Wynn K, Bloom P (2010) Three-month-olds show a negativity bias in their social evaluations. Dev Sci 13(6):923–929. DOI:10.1111/j.1467-7687.2010.00951.x

Hardwick N (2012) Theoretically Justifying Human Rights: A Critical Analysis. (published student essay) E-International Relations. 5th Aug 2012. http://www.e-ir.info/2012/08/05/theoretically-justifying-human-rights-a-critical-analysis/. Accessed 22 Aug 2015

Hare B, Call J, Tomasello M (2001) Do chimpanzees know what conspecifics know? Anim Behav 61:139–151. doi:10.1006/anbe.2000.1518

Harter S (1983) Developmental perspectives on the self-system. In: Mussen P (ed) Carmichael's 'Manual of child psychology, vol 4. Wiley, New York, pp 285–386

Hawking S, Russell S, Tegmark M, Wilczek F (2014) Transcendence looks at the implications of artificial intelligence—but are we taking AI seriously enough? The independent, Thursday 01 May 2014. http://www.independent.co.uk/news/science/stephen-hawking-transcendence-looks-at-the-implications-of-artificial-intelligence-but-are-we-taking-ai-seriously-enough-9313474.html. Accessed 12 Feb 2017

Hawking S, Musk E, Wozniak S et al (2015) Autonomous weapons: an open letter from AI & Robotics researchers. future of life institute. http://futureoflife.org/AI/open_letter_autonomous_weapons. Accessed 12 Feb 2017

Heath M (2008) Aristotle on natural slavery. Phronesis 53:243–270. http://eprints.whiterose.ac.uk/4463/1/Heathm1.pdf. Accessed 12 Feb 2017

Hughes M (1975) Egocentrism in preschool children. Doctoral dissertation. Edinburgh University, Edinburgh

Humans (2015) [TV Series]. Kudos, Channel 4. http://www.imdb.com/title/tt4122068/. Accessed 12 Feb 2017

James W (1884) What is an emotion? Mind 9:188–205. http://psychclassics.yorku.ca/James/emotion.htm. Accessed 12 Feb 2017

Keller H, Sullivan A, Macy J (1905) The Story of My Life. New York, NY: Doubleday, Page & Co. http://digital.library.upenn.edu/women/keller/life/life.html. Accessed 12 Feb 2017

Kirk E, Pine K, Wheatley L, Howlett N, Schulz J, Fletcher B (2015) A longitudinal investigation of the relationship between maternal mind-mindedness and theory of mind. Br J Dev Psychol 33(3) doi:10.1111/bjdp.12104 **(published online: 27 July 2015)**

Kirkpatrick L (2004) Attachment, Evolution, and the Psychology of Religion. Guilford Press, New York

Lantz J (2002) Theory of mind in autism: development, implications, and interventions. Reporter 7(3), 18–25. http://www.iidc.indiana.edu/pages/Theory-of-Mind-in-Autism-Development-Implications-and-Intervention. Accessed 12 Feb 2017

LeDoux J (1998) The emotional brain. Phoenix

Leviticus A Hebrew—English Bible according to the Masoretic Text and the JPS 1917 Edition. Text of Lev. 19 at http://www.mechon-mamre.org/p/pt/pt0319.htm. Accessed 12 Feb 2017

Lewis M, Sullivan M, Stanger C, Weiss M (1989) Self development and self-conscious emotions. Child Dev 60(1):146–156. doi:10.2307/1131080

Libet B (2005) Mind time: the temporal factor in consciousness. Harvard University Press, Cambridge

Liddle B, Nettle D (2006) Higher-order theory of mind and social competence in school-age children. J Cult Evol Psychol 4(3–4):231–246

Luke, New Testament, New International Version. Ed. E H. Palmer et al., 1973. Full text of Luke 6: https://www.biblegateway.com/passage/?search=Luke%206&version=NIV. Accessed 12 Feb 2017

Marraffa M (2011) Theory of Mind. Internet Encyclopedia of Philosophy. http://www.iep.utm.edu/theomind/. Accessed 29 Aug 2014

Mahābhārata (900–400BC) The Mahābhārata of Krishna-Dwaipayana Vyasa, Book 12: Shānti-Parva, Section CLXVII. Tr. Kisari Mohan Ganguli (1883–1896) Full text at: http://www.sacred-texts.com/hin/maha/. Accessed 12 Feb 2017

McIntyre M (pre 1923) The cave boy of the age of stone. https://archive.org/details/caveboyofageofst00mcinuoft Illustrations accessed from: https://commons.wikimedia.org/wiki/File:Caveman_1.jpg http://commons.wikimedia.org/wiki/File:Caveman_2.jpg

McNamara P (2014) The neuroscience of religious experience. Cambridge University Press, Cambridge

Meltzoff A (1995) Understanding the intentions of others: Re-enactment of intended acts by 18-month-old children. Dev Psychol 31(5):838–850. doi:10.1037/0012-1649.31.5.838

Metta G, Natale L, Nori F, Sandini G, Vernon D, Fadiga L, Montesano L (2010) The iCub humanoid robot: an open-systems platform for research in cognitive development. Neural Netw 23(8–9):1125–1134. doi:10.1016/j.neunet.2010.08.010

Metzinger T (2009) The ego tunnel—the science of the mind and the myth of the self. Basic Books, New York ISBN 0-465-04567-7

Mithen S (2007) The singing Neanderthals. The origins of music, language, mind, and body. Harvard University Press, Cambridge ISBN 9780674025592

Mithin S (1996) The prehistory of the mind. Thames and Hudson, New York

Morin R (2016) Can child dolls keep pedophiles from offending? The Atlantic (11 Jan 2016). http://www.theatlantic.com/health/archive/2016/01/can-child-dolls-keep-pedophiles-from-offending/423324/. Accessed 12 Feb 2017

Nadella S (2016) The partnership of the future. Slate, 28 June 2016. http://www.slate.com/articles/technology/future_tense/2016/06/microsoft_ceo_satya_nadella_humans_and_a_i_can_work_together_to_solve_society.html. Accessed 12 Feb 2017

Nagel T (1974) What is it like to be a bat? Philos Rev 83(4):435–450
Nijholt, A. (2011) No Grice: Computers that Lie, Deceive and Conceal. In: 12th International Symposium on Social Communication, 17–21 January 2011, Santiago de Cuba, Cuba (pp 889–895). http://doc.utwente.nl/75827/. Accessed 12 Feb 2017
Parker A (2003) In the blink of an eye: how vision sparked the big bang of evolution. Perseus Publisher. Cambridge
Piaget J (1932) The moral judgment of the child. Kegan Paul, London
Piaget J, Inhelder B (1956) The Child's conception of space. Routledge & Kegan Paul, London
Premack D, Woodruff G (1978) Does the chimpanzee have a theory of mind? Behav Brain Sci 1:515–526
Prescott T (2015) Me in the Machine. New Sci 225(3013):36–39
Psalm 139 In the holy bible, new international version 1982. Hodder & Stoughton, London. Full text of Psalm 139.https://www.biblegateway.com/passage/?search=psalm+139&version=NIV. Accessed 2 Feb 2017
Ramachandran V (2004) A brief tour of human consciousness. Pi Press, New York
Rawls J (1971) A theory of justice. Belknap Press, Cambridge ISBN 0-674-00078-1
Renfrew C (2007). Prehistory: the making of the human mind. Phoenix, London
Rosenthal D (1986) Two Concepts of Consciousness. Philos Stud 94(3):329–359
Rutkin A (2016) Curbing dark desires. New Sci 231(3086):20. Version at: https://www.newscientist.com/article/2099607/. Accessed 12 Feb 2017
Schön D (1984) The reflective practitioner. Basic Books, London
Shanahan M (2010). Embodiment and the inner life: cognition and consciousness in the space of possible minds. Oxford University Press, Oxford
Shukman D (2015) Being comfortable in robotics' uncanny valley, BBC News
Singer P (2009) Isaac Asimov's laws of robotics are wrong. The Brookings Institution. May 18, 2009. https://www.brookings.edu/opinions/isaac-asimovs-laws-of-robotics-are-wrong/. Accessed 12 Feb 2017
The Kursaal Flyers (1976) Little does she know. Golden Mile [LP record]. CBS Records. Lyrics: http://lyrics.wikia.com/wiki/Kursaal_Flyers:Little_Does_She_Know YouTube: https://www.youtube.com/watch?v=5ZnQQF7ucdM. Accessed 12 Feb 2017
Tomasello M (1999) The Cultural Origins of Human Cognition. Harvard University Press, New York ISBN 0-674-00582-1
Tooby J, Cosmides L (1997) Evolutionary psychology: a primer. online at: http://www.psych.ucsb.edu/research/cep/primer.html. Accessed 12 Feb 2017
Udâna (date uncertain) (1902) The Udâna, Chapter V. Sona Thera. Tr. Dawsonne Melanchthon Strong Full text at: http://www.sacred-texts.com/bud/udn/. Accessed 12 Feb 2017
Watercutter A (2015) Ex Machina Has A Serious Fembot Problem. Wired, 9 April 2015. http://www.wired.com/2015/04/ex-machina-turing-bechdel-test/. Accessed 12 Feb 2017
Westminster Shorter Catechism (1642–1647) Question 1. Text at: http://www.shortercatechism.com/resources/wsc/wsc_001.html. Accessed 12 Feb 2017
Whitby B (2008) Sometimes it's hard to be a robot: a call for action on the ethics of abusing artificial agents. Interact Comput 20:326–333
Wikipedia (2015) Venus von Willendorf, image from Mother Goddess. https://en.wikipedia.org/wiki/Mother_goddess https://commons.wikimedia.org/wiki/File:Venus_von_Willendorf_01.jpg. Accessed 6 Aug 2015
Wittgenstein L (1958) Philosophical investigations, (extract "Meaning as use", Chap. 12. In: Nye A (ed) Philosophy of language: the big questions. Blackwell, New Jersey **(1998)**
Wittgenstein L (1980) Remarks of the philosophy of psychology II. Blackwell, New Jersey
Zahn-Waxler C, Robinson J, Emde R (1992) The development of empathy in twins. Dev Psychol 28(6):1038–1047. doi:10.1037/0012-1649.28.6.1038