Being One, Being Many

Christian Kroos and Damith Herath

Abstract If the current development of robotics indicates its future, we will be soon able to create robots that are exactly identical, intentional agents—at least as far as their software is concerned. This raises questions about identity as sameness and identity in the sense of individuality/subjectivity. How will we treat a robotic agent that is precisely the same as multiple others once it left its inanimate appearance behind and by its intentionality claims to be individual and subjective? In this chapter we show how these issues emerged in the implementation of the artwork 'The Swarming Heads' by Stelarc.

I

Identity in intentional agents (humans, animals, robots) is traditionally understood in the Cartesian sense as being subject to spatial and structural coherence. The agent cannot be at two or more places at the same time or be several separate physical entities. Emotional and cognitive processing happens on the inside, within some kind of border that separates the agent from its environment. For biological agents Andy Clark has called this border the 'metabolic boundary' [1].

In the internalist view, the environment arrives in the form of sensory 'input' and the agent performs disassociated information processing to produce adaptive motor behaviour considered 'output'. Various externalist approaches, among them Clark, have put forward strong arguments against the input/output reduction,

C. Kroos (✉)
Centre for Vision, Speech and Signal Processing,
University of Surrey, Guildford GU2 7XH, UK
e-mail: chkroos@gmail.com

D. Herath
Faculty of Education, Science, Technology and Mathematics,
University of Canberra, Canberra, Australia
e-mail: damithc@gmail.com

D. Herath et al. (eds.), *Robots and Art*, Cognitive Science and Technology,
DOI 10.1007/978-981-10-0321-9_10

emphasising the fact that individual beings are embedded into their surrounding environment through a *gewebe* (web) of interactive relationships. However, even if the information processing view is not upheld in an externalist approach, the agent conventionally resides in a single location and at best extends into the environment.

According to the Cartesian tenet, identical reduplication of the agent leads to the creation of several different agents with identical properties. Our phylogenetic and (currently also still) ontogenetic experience with exclusively biological agents might have crucially shaped our intuition. The metabolic boundary convincingly and verifiably defines the perceivable boundary of any biological agent (the story might be more complex in plants though).

Technically, nearly exact reduplication of a robotic agent is straightforward, owing to the industrial production of the components in the networked way described by Gilbert Simondon as drawing out the 'technical mentality' [2]. There are remaining differences between agents; hardware components are only identical to the degree specified through set production tolerances, and more importantly, the physical extension of the robot agents always allows marking them for identification in one way or another, that is, presenting them separately, referring explicitly to individuals or even destroying a specific individual while keeping the others. In contrast, the software of the agent can be *exactly* identically reproduced and would stay this way unless unsupervised learning algorithms are used or hardware problems lead to processing failures. Thus, if one would grant current autonomous robots agency—and noted, that would be controversial—we are already capable of creating agents which are different and yet the same (Fig. 1).

Fig. 1 Swarming Heads installation (© Christian Kroos, Damith Herath and Stelarc; *photo* Christian Kroos)

Identical robotic agents are likely to be readily accepted in the (post-)industrial culture, owing to their perception as mere machines (lacking 'feelings', 'consciousness', a 'soul', etc.). Combined with a still prevalent mind-body dualism, the mechanistic perspective prevents the dilemma of split identities our human thinking would otherwise face. If there is no mind in the machine, having several identical agents is not more problematic than a collection of e.g. identical mobile phones in a store. It becomes more complicated if the mind cannot be thought any longer as an entity independent of its physical implementation or—alternatively and currently only in fiction—if the absence of an artificial mind in a machine cannot be any longer assumed beyond doubt. In popular culture, the latter is often construed as a scenario in which the information-based mind/consciousness of an agent can be transferred to different physical implementations. The information-based mind is considered unique while the physical implementation can be identically replicated—rather the opposite of the technical reality of software and hardware today. From the tension between the fictional account and the current reality of computational programs often the fundamental conflict in these narratives arises.

Moreover, the scenario of the unique mind and the replaceable body of the machine frequently leads to the reverse inference that it will become possible at one point in the foreseeable future to transfer ('upload') the human mind using technology not yet developed but conceivable. Typically and without further explanation, the transfer can be only accomplished in the moment of dying, presumably to avoid the problematic topic of identical agents—the prospect of creating identical agent copies might be too challenging.

In the Western industrialised nations, a tradition of fearing the 'Doppelgänger' appears to be deeply engrained into society, from the German silent movie 'Der Student von Prag' (1913, directed by Stellan Rye and Paul Wegener, written by Hanns Heinz Ewers) to José Saramago's novel 'O Homem Duplicado' (2002) to the Hollywood movies 'Matrix Reloaded' and 'Matrix Revolutions' (both 2003, written and directed by the Wachowski brothers), to mention only a few. Note, however, that most of these depiction only refer to appearance while the 'mind' is always unique, including in the case when robotic technology is used as in Fritz Lang's classic silent movie 'Metropolis' (1927), in which an indistinguishable robotic copy of working class activist Maria is created.

It appears to be excruciatingly difficult or outright paradoxical to consider identical conscious agents, that are not—in some way or another—a single entity. This difficulty is also reflected in the widely unchallenged acceptance of the idea that storing all the information of the brain (whatever that exactly would mean) in an external device would constitute a continuation of this one person and not a new individual. If it would be indeed continuation, however, that is, if the person, whose brain information is transferred, is the same as the newly created recipient of this information, any additional copying of those constitutive data would create a serious predicament: Either the copies would create new individuals leading to the paradox that the process could not have been continuation in the first place (even in the case when only one new agent is created) or a single mind would split in several entities. For the latter we appear to have few concepts to apprehend its

meaning, both intellectually and emotionally. Typically it would be framed retrospectively, in which case its defining characteristics can be reduced to identical memories of a shared past. But this ignores the transition process, in which a person changes from being one to being many, regardless of how quickly the new instantiations diverge afterwards. Admittedly, one could question whether there is continuation in the first place or whether the perceived continuation is always constructed retrospectively since any period of unconsciousness disrupts experienced continuation nevertheless.

These issues sometimes surface in the discussion of human cloning, too. Despite lacking any basis here, since only DNA is replicated and since even monozygotic twins are not genetically exactly identical [3] and the differences can be assumed to be even more pronounced in clones. Most importantly, clones would go through their own biological, in particular neural, and mental development, shaped by individual experiences. Accordingly, there are few if any justifications to question the individuality of the clone. The life experiences of the clone would always be different from the ones of the source individual and if it was only because of the different 'parent' situation. Still, even a contemporary artist with a Ph.D. in Genetics appeared compelled to have to point out explicitly in a public presentation that human clones should have human rights and should be considered individuals: As if a certain degree of congruence would inevitably have to be thought as complete unity and thus the seemingly identical make-up, but multiple physical instances would require re-asserting the foundations of what makes a person a person. Interestingly, it seems to be never the source human that was (hypothetically) cloned, whose individuality and personhood is in doubt as a consequence of the cloning process.

The aforesaid evokes an alternative solution, one which is again conjured frequently in popular culture—especially, if intelligent robots are involved—and which emerges as a trend in current robot development: All identical individuals are connected into one comprising 'organism'. If taken seriously, this amounts to more than a hidden communication channel among the agents. In its simpler shape, there would be a remote central controlling entity, a master mind, so to speak, and identical replication of individual semi-autonomous agents would resemble adding an additional eye or leg within the animal analogy. After all, humans are not alarmed by having two very similar and functionally nearly identical eyes, ears, legs or arms and adding another one would create practical but not philosophical problems—see e.g., Stelarc's Third Hand [4]. In its more complex shape, there would be no such central control and although things start to get messy in terms of imagining the inner working of such an organism, no paradoxical or unimaginable situation would present itself. Some schools of thought in contemporary neuropsychology are already trying to get us used to the idea that there might be no single location in the brain, where consciousness or awareness resides [5]. Any set-up of a cohesive, but dispersed technological organism without central control is at present still beyond current robotic technology and artificial cognition, with the exception of the most basic levels, e.g., ad hoc networks. Artificial swarm behaviour in robot collectives [6] seems to come close, but differs in an essential aspect: The

individual robot is seen as an individual agent and is recruited to solve a common task. The biological models used are often ant or bee colonies, in which agency resides within the individual animal and is not taken over by the colony. Thus, it looks as if with future technological progress we will be first faced with the more confusing and challenging situation of identical, individual agents within in the domain of autonomous machines. It appears to be about time to explore this future.

II

In 2012 the Thinking Head project came to an end. The multi-university, interdisciplinary research undertaking funded by the Australian Research Council and the National Health and Medical Research Council had the aim to develop a sophisticated embodied conversational agent, a 'talking head' that would venture beyond uttering only pre-defined phrases and would pass for being intelligent. The project's starting point was the *Prosthetic Head* by Australian performance artist Stelarc, a convincing virtual 3D representation of the artist, created using a laser scan of the artist's head and animated using computer graphics. People were able to interact with the *Prosthetic Head* by submitting questions or comments through a computer keyboard. A modified version of the A.L.I.C.E. chatbot [7], a widely used conversational artificial intelligence computer program, generated the responses.

The research-and-art track of the *Thinking Head* project had produced a robotic embodiment of the *Prosthetic Head*, an art installation initiated and conceived by Stelarc and built by a small team of two robotics engineers (one of them the second author of this chapter) and a cognitive scientist (the first author). The robot (Fig. 2), named *Articulated Head*, exceeded the original aims of the Thinking Head project, in which the agent was never meant to become a part of the physical world. The artist's vision of an LCD monitor displaying the *Prosthetic Head* as the end-effector of a six-degree-of-freedom industrial robot arm stimulated extensive further research. After all, here was a powerful machine with a vast range of movement possibilities: A potential waiting to be utilised and—not surprisingly—at the same time a potentiality posing deep challenges. Each of the six sequential joints allowed the rotation of the connected limb with rotational speeds ranging from 0 to 360 degree/s, enabling a rich continuum of motor behaviour that could be harnessed in order to realise the artistic and scientific aim of creating the impression of the *Articulated Head* as an intentional agent. The research resulted in a complex control system that used the advanced sensing capabilities empowering the *Articulated Head* and included a software-based attention model to let seemingly meaningful behaviour arise from the interaction with the visitor.

As the *Thinking Head* project drew to an end, Stelarc suggested another rather different robotic embodiment of the *Prosthetic Head*: A swarm of small mobile (wheeled) robots, which again would show the *Prosthetic Head* on their individual LCD monitors, but move around on their own accord.

Fig. 2 The Articulated Head in the Powerhouse Museum, Sydney, Australia (© Christian Kroos, Damith Herath and Stelarc; *photo* Christian Kroos)

The transition between these very different embodiments and an analysis of the technical, scientific and conceptual implications will be the subject of the next sections. Our focus will be on emerging behaviour as a consequence of design and implementation choices and the resulting differences in the structuring of the interaction with humans. We will finally revisit the fundamental questions that arise from identical replication of a robotic agent and touch on issues of sameness and individuality on a more concrete basis.

III

The *Articulated Head* consisted of a robotic platform that was not able to change its location. Although fully flexible where to orient its 'face' and focus its attention, resting on a static tripod, the *Articulated Head* could not leave its safety enclosure or move its entire 'body' toward or away from an interaction partner (it could turn, though, and face the other direction). There was also only one mobile sensor, a camera, used for visitors' face detection, attached to the top of the LCD monitor. The remaining sensors were fixed: An acoustic localisation system employed two microphones clipped to the top of the back wall of the enclosure. A short-range sonar proximity sensor was integrated in an information kiosk,

which housed the similarly unmoveable keyboard. Most importantly, the main stereo camera for software-based people tracking was mounted at a museum wall opposite the enclosure, amounting to a third-person perspective. Visitors were not aware of the locations of the sensors and appeared to assume all sensing devices were attached to the computer monitor displaying the virtual face: Attempts to attract the attention of the *Articulated Head* through e.g. gestures, jumping up and down, and vocalisations were always directed toward its 'head'. Furthermore, the conceptual framework, the technical implementation and the control system incorporated the assumption of a static base location and a third-person perspective from the beginning. In some ways the *Articulated Head* resembled more a coral polyp than a mammal.

The control system of the *Articulated Head*, the Thinking Head Attention Model and Behavioural System (THAMBS), is described elsewhere [8, 9], therefore we will give only a brief overview here, going as far into the details as is needed for later sections.

THAMBS (Fig. 3) employs a primary processing cycle, which sequentially runs through all the necessary tasks to maintain its situational knowledge and generate its response behaviour. In a single processing cycle, sensory information arriving from low-level processing routines such as acoustic localisation or people tracking is turned into standardised *perceptual events* by a perception subsystem. The properties of the events are subjected to threshold tests, introduced

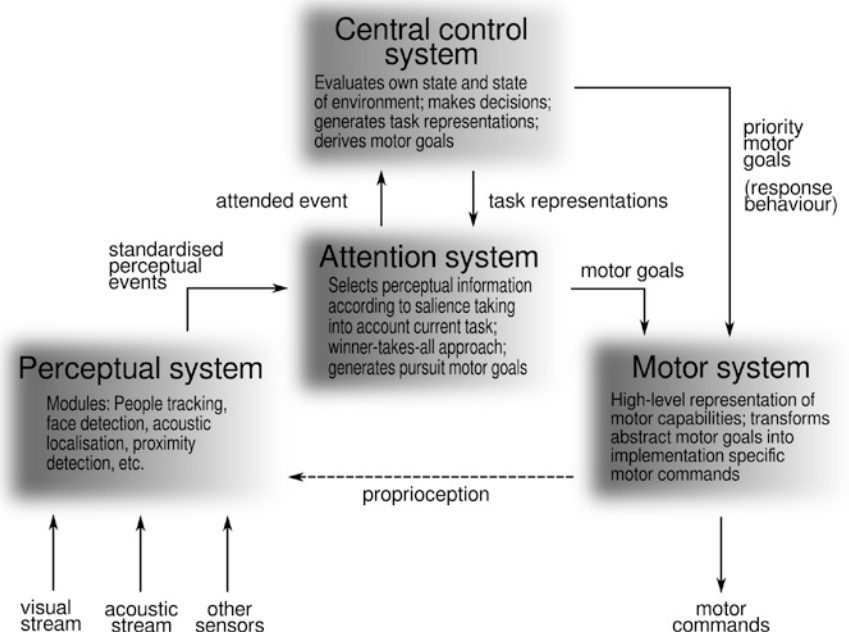

Fig. 3 Thinking Head Attention Model and Behavioural System (THAMBS)

to remove e.g. unreliable tracking values from further consideration. Surviving events are passed on to the attention subsystem, which subjects the events to its own thresholding based on the perceptual event type and dependent on the state of the overall system and its current task. For instance, THAMBS might be in a vision-based interaction with a visitor and thus change its setting to make it more difficult to divert its 'attention' through an unrelated acoustic event. This thresholding, however, constitutes only a basic, brute-force mechanism to manage the system's attentional behaviour. The primary mechanism employs attention weights and attention decay profiles assigned to the attention foci created from the perceptual events after passing the initial threshold test.

Since unconstrained object recognition in real-world environments is an unsolved problem [10] and even robust tracking poses serious challenges [11], attention foci are spatially defined: THAMBS pays attention to a specific confined three-dimensional region in the space surrounding the *Articulated Head*. The attention weights are determined in relation to current system preferences, task requirements and the event type. The weight values are decisive in the final selection of an attention focus as the single attended event (winner-takes-all strategy). An active focus persists for a certain duration, but its weight decays exponentially over time, though with a relatively flat curve. Persistence and decay enable short-lived, but prominent events, say, a loud noise burst, to attract THAMBS attention beyond the lifetime of the event, but also guarantees that the attention to static or repetitive attractors wanes over time (habituation). Outdated attention foci are able to bind the system's attention only if nothing else of interest happens in the robot's environment and even then only for limited time.

The attended event is forwarded to the behavioural control system, the transfer realising selection-for-action, the second primary function identified for biological attention systems besides binding different events to a single focus: Attention is guided by the actions available to the individual relative to the affordances of its environment and prioritises stimuli that have particular relevance for those potential actions. In THAMBS, its behavioural system invokes then a behavioural response, which includes the option to ignore the event. If the response involves a motor action (movements of the robot, facial expressions of the virtual avatar, speaking), a dedicated motor system generates the action command, filling in context-specific parameters where needed.

In the onsite implementation it was attempted to uphold an operating speed of 10 Hz, but the system often slowed down to speeds as low as 5 Hz due to processing bottlenecks. Nevertheless the control system served its purpose well, once it was protected against information overflow. Before that, in the opening night of the first short-term exhibition at the University of Technology, Sydney, Australia, as part of the 2010 NIME conference (New Interfaces for Musical Expression++, 15–18th June 2010) the *Articulated Head* was faced with a crowd of several dozen people instead of the usual handful during development and testing. THAMBS was utterly overwhelmed by the influx of potential attention foci, all the people standing around the robot's enclosure within its field of interaction, and the system, after briefly switching helplessly from one visitor to the next, froze and all

movements came to a halt. Though not intended and slightly embarrassing, we could not help finding the behaviour of the *Articulated Head* appropriate, simulating successfully an intentional agent that experienced a sudden unexpected large crowd of relevant other agents. The reaction of many animals would not have been so different.

From the description above it might have already become clear that the unmodified control system of the *Articulated Head* would be in a permanent crisis when 'inserted' in a small mobile robot. Instead of a fixed world entering from the outside through selected events of interest, it would now encounter a world which would change with every rotational and translational movement: THAMBS would be subjected to an inescapable first-person perspective.

IV

The *Swarming Heads* [12] were designed as small mobile robots that similarly to the *Articulated Head* would display a virtual representation of the artist's face on an LCD monitor. They were meant to be fully autonomous, although not acting to fulfil any utilitarian function, but to explore their world in a playful manner. They were built around the commercially available robot platform *Create* developed by *iRobot,* which resembles closely the original vacuum cleaning robot *Roomba* of the same company (but unfortunately lacking the useful vacuuming function). The base robot is a differential drive platform supported by front and rear castor wheels. A custom designed Perspex frame was added to hold a tablet computer that drove and displayed the *Prosthetic Head* on a 12.1 inch screen. A separate Linux computer was housed behind the tablet in a transparent casing, providing the computational power to run the sensing algorithms and THAMBS. The front Perspex frame also accommodated a skinned version of a *Microsoft Kinect* sensor. The robot used two sets of power sources, one to drive the motor mechanisms and other internal hardware of the robot base, a second one tucked underneath the Linux PC to power the computer and sensors.

The Kinect sensor returns rich 3D depth information of the environment in its field of view. It replaced the stereo camera system used with the *Articulated Head*; the acoustic localisation, however, was not transferred to the *Swarming Heads*. The robot base has an in-built four-way split cliff sensor that can detect sudden discontinuities on the ground, identifying the location of the drop ahead (whether it is to the left or right of the robot or directly in front, but again divided into left and right hand side). The wheels of the base contain odometry sensors, providing local translational information. The wheels are also connected to a lift sensor that gets activated when the robot is lifted up from the floor. A frontal bumper sensor, integrated into the robot base as well, generates left/right bumper activation signals when coming into contact with obstacles. All low-level sensory data were accessed through the Robot Operating System (ROS)—an open sources robotics-specific operating system—to be further processed and manipulated.

A new version of THAMBS was instantiated, called mTHAMBS. It included the four new sensors (cliff detection, lift sensor, bumper sensor, odometry). Due to the flexible core architecture of THAMBS the integration required only minor changes. A fundamental alteration, however, followed from the loss of the static world coordinate system. The term 'world coordinate system' is used in computer vision, robotics and related disciplines to describe a reference system anchored in the physical environment, which typically is not influenced by the robot's location and orientation or sensing parameters, e.g. perspective distortion caused by camera lenses. In the *Swarming Heads*, however, the entire visual field covered by the Kinect sensor was likely to change with any significant movement, for instance, as the consequence of the reaction to a peripheral stimulus that caught mTHAMBS's attention such as turning toward a person. In addition, any movement of the robot, but in particular rotational movements, would cause apparent motion in the visual field, and this apparent motion would mix with the real motion of external entities. New potential attention foci would be brought constantly into play, since mTHAMBS did not comprise any kind of episodic memory of its environment. As mentioned above, mTHAMBS was not able to 'lock' on objects, only people could be tracked and only for so long as they stayed within the field of view of the Kinect. As a consequence, the initial *Swarming Head* became very fixated on people, but also constantly distracted by its own exploration of the world that in a Heraclitian sense (Plato's view of it, to be precise) appeared to be in a permanent flux. Fine tuning of the attention weights, in particular re-evaluating the impact of apparent velocity, alleviated these behavioural problems to a degree that made uninterrupted interaction between a human and the robot possible and mTHAMBS was no longer producing behaviour akin to an attention disorder syndrome.

A more essential technical problem remained though. The tablet PC, which was running mTHAMBS but also the software generating and rendering the virtual head, was not able to maintain the central mTHAMBS loop at even the reduced rate of 5 Hz. It dropped regularly to 1 Hz, occasionally to half of that and sometimes even further. A perception-action cycle of 0.5 Hz meant that it took mTHAMBS 2 s to update its perception and attention system and modify any active motor command. This did not only severely impact on its capability of a timely response in human-robot interactions, but caused the *Swarming Head* to shoot straight over any cliff in its path. To avoid catastrophic damage to the robot, both the cliff and the bumper sensor were integrated into a reflex loop that bypassed mTHAMBS and secured an immediate stop of the motor. The information about the emergency motor stop was then forwarded together with the original cliff or collision detection information to mTHAMBS to 'deliberate' on the action to be taken, now that a response was no longer time-critical.

The general problem of delayed processing, however, could not be remedied. THAMBS had been already optimised for execution speed as much as was possible without compromising its flexibility. It became clear that the usual path planning strategy for a robot with two wheels driven by independent motors and a castor wheel could not be used. This conventional way separates rotational movements (turns) from translational forward movements. The strategy consists of a

two-step sequence: First turning towards the target location on the spot and then moving forward in a straight line until the target location is reached [13]. This can be followed by a potential adjustment of the orientation of the robot through a second turn. Given the slow processing, pursuit movements using this strategy would have in most cases resulted in the robot only turning on the spot, trapped in a constant adjustment of the orientation. If the robot would indeed have progressed to the stage of forward movement, it would likely have stopped shortly afterwards to re-adjust its orientation. Therefore, we implemented an alternative path planning strategy that uses curved trajectories when the target was not strictly straight ahead. To keep orientation changes and forward movements incremental and smooth, a circular trajectory between the current location of the robot and the target is computed. The current orientation of the robot relative to the target determines the curvature of the arc: It is more strongly curved if the target is located in the periphery of the robot's visual field and less curved if the target is closer to the centre of the visual field, diminishing to zero curvature (a straight line) if the target is straight ahead. If a new arc has to be computed while the robot is in motion triggered by a changed target location, it is guaranteed that only minor adjustments to the robots orientation are required, since the overall adjustment is spread out over the entire trajectory. In this way orientation angle and radial distance were gradually and simultaneously adjusted by continuously minimising the difference between actual and target orientation and location.

The procedure enabled a kind of sluggish pursuit behaviour. The price to pay were slightly awkward looking initial trajectories if the target was located in the horizontal periphery of the visual field of the robot. The robot seemed at first to move in the direction in which it was already oriented, ignoring the target, before gradually zeroing in on the target as if the robot wanted to avoid a direct 'confrontational' course.

Of course, none of the measures taken amounted to much more than control 'band aid' of the processing speed shortfalls, they could not solve, but would merely mask the fundamental problem that the robot's higher level processing was occasionally operating on a time frame not suitable for interactions with humans. Surprisingly, reasonable robot behaviour was achieved resulting in the impression of an engaging and accommodating machine. It is difficult to say whether this was due to the robot just delivering the right cues to evoke the impression of agency [9] combined with a forgiving patience of the human interaction partner or whether it was due to (approximately) smooth interaction occurring despite the robot's shortcomings.

Evidence for the former came from the experience with a gesture-based control that was implemented as part of a more traditional scientific longitudinal human-robot interaction study into bonding behaviour with a robot [14]. The gesture control used so-called skeleton tracking routines implemented in the open source Natural Interface algorithms (OpenNI) for the Kinect sensor (http://structure.io/openni). In the Swarming Heads, it allowed any person within the visual field of the Kinect sensor to directly steer the robot with a set of fixed gesture commands. There was a kick-off gesture that corresponded to a 'pay attention' command. It caused a change in the attention-related parameters of mTHAMBS to strongly

prioritise gesture recognition and associated behaviours, e.g., motor commands linked to specific gestures. Distracting the robot from following the gesture commands was made difficult, but was still possible. The remaining gesture commands can be paraphrased as 'come to me', 'turn right' (−90°), 'turn left' (90°), 'turn around' (180°) and 'stop'. Note that all the turn commands changed the robot's orientation sufficiently to move the gesturing human out of sight of the robot and consequently required new positioning of the human in the robot's visual field, thus, weakening the dominating role of the human in the interaction by requiring human adjustments to the robot's behaviour. If accommodations to the robot's new location and orientation were neglected, the robot would lose its prioritisation of the gesture recognition input after a short while and would happily continue with its normal exploratory behaviour.

Obviously, the gesture control was not spared by the processing delays and could render the robot unresponsive for new commands for the duration of two seconds and more while being occupied with the outdated execution of a previous gesture command or still following its internal behaviour preferences. These black-out durations were far too extended to be accepted in typical human interactions (see teleconferencing latencies, e.g., [15]) and were potentially beyond the limits of interpersonal or human-machine synchrony requirements, too [16]. However, as observed in several trials in the lab with university staff not part of the project and in a public event at the Powerhouse Museum (Sydney, Australia) people adjusted to the robot's occasional unresponsiveness. Instead of blaming failing technology, they interpreted the behaviour of the robot as inattentive, stubborn or outright mischievous. But this made them try even harder to establish a successful relationship with the *Swarming Head*.

Additional subjective anecdotal support came from the experience of the first author during early lab tests with the *Swarming Heads*. To examine mTHAMBS' working and the resulting behaviour of the *Swarming Heads* in the wild, individual robots were often set free in the HRI lab at the MARCS Institute (Western Sydney University), a spacious windowless room with a single door to a corridor leading to a public foyer and the building's exits. The door was usually left open and one day one of the *Swarming Heads* was heading straight for the exit. It happened at a stage in the development when the hardware built was finished and mTHAMBS working, but no sensing activated except for the reflex-like bumper sensors and the cliff detection. In this situation, that is, when mTHAMBS receives almost no environmental input, it switches to an exploratory 'idle' mode. It generates single movement targets or short sequences of movement targets using a constraint pseudo-random procedure applied to robot location, orientation and timing of the movement.

The robot could not see the location of the door or anything else, yet it went for the door, stopped, turned around as if to check with the experimenter, turned back and moved about a meter straightforward. It then stopped again, turned a second time, not quite far as the first time, as if pretending to have changed its intention and path, after which it rotated back to its original orientation and left the room. At this point the experimenter had to go and get it, since the busy foyer was not

a suitable environment for a small blind robot. Despite knowing better than everyone else that there was nothing going on in the robot other than a simple, but appropriately fine-tuned random procedure, the first author could not help himself from perceiving the episode in terms of an intentional robotic agent attempting to sneak out of its designated area. The series of serendipitously structured events evoked a strong sense of agency that was—at least for a brief moment—powerful enough to overcome the certainty of the developer's knowledge.

When the sensing was activated and the *Swarming Head* could detect people in its surroundings, the behaviour of the robot evoked the impression of agency convincingly without relying on serendipitous movement sequences. The responsive and exploratory conduct of the robot changed the behaviour of the human interaction partners as they started to adapt their behaviour to the robot and its perceived intentions. As a consequence, processing delays were reliably interpreted as lack of social ability or lack of willingness of the robot to cooperate or as outright defiance, but not as failures of technology. Therefore, for most people the motivation to make the robot-human relationship work increased and they put in an extra effort to compensate for the cognitive shortcomings or moods of the robot.

V

The *Swarming Heads* did not really deserve their names; they did not exhibit swarming behaviour as there were no routines implemented that triggered mimicking the behaviour of compatriots or allowed them to set their behaviour in relationship to that of another robot. They were also not entirely independent individuals, since with respect to their behavioural program they were identical copies. The use of probabilistic behaviour generation hid their lack of uniqueness on the surface, but did not alter their conceptual sameness.

The *Swarming Head* installation (Fig. 4) raised some of the questions discussed in Sect. I in a playful manner and used the anthropomorphic appearance of the *Prosthetic Head* as a reinforcement of their potentially challenging underpinnings. The installation conceived by Stelarc gathered five *Swarming Heads* robots on a circular pedestal with a diameter of 200 cm. The top side of the pedestal was flat and painted black. A six centimetres high translucent plexiglass raised rim running around the perimeter of the pedestal served as a fall-off barrier: The *Swarming Heads* could detect a cliff and avoid it, but nothing prevented a robot from pushing its colleague over the edge. The *Swarming Heads* moved freely in this area and were attracted by the presence of visitors. If visitors approached the installation with high walking speed, the *Swarming Heads* tended to avoid an interaction and turned away; if the approach speed was slow or the visitors maintained constant distance (moving in an orbit around the pedestal or standing still), the *Swarming Heads* exhibited curiosity and approached as far as possible. They then often locked on individual visitors, tracked their movements continuously and waited for gesture commands as a way to establish a robot-human relationship.

Fig. 4 Swarming Heads installation (© Christian Kroos, Damith Herath and Stelarc; *photo* Christian Kroos)

Since their area was rather limited, they frequently bumped into each other or ran into the confining outside rim. Any collision triggered an avoidance reaction in the robot—moving a few centimetres backwards and then turning (the turn angle was determined by a constrained pseudo-random procedure)—and most of the time also a verbal response. For the latter a phrase was selected out of 50 pre-scripted response phrases and uttered by the *Prosthetic Head*, both acoustically and visually (synchronised face motion). The phrases were mostly trivial such as 'Oops', 'Sorry', 'Not again' and 'Back up!', with a tendency to complain about the situation or the other ('Idiot', 'Silly', 'Are you always like this', 'Today is not my day') and occasionally putting the collision event into a larger context ('Lately I seem to run into all kind of things', 'We don't do this where I come from') or denying the problem ('I did not want to go in this direction anyway'). The intention was to pretend in a shallow way underlying intelligent behaviour that after a while would expose its repetitive character. The robots resembled each other very closely, the virtual *Prosthetic Heads* shown on the tablet screen looked exactly the same and their behaviour was revealed over time to be identical, too.

The installation was exhibited during the two days of the *Thinking Systems Initiative* Symposium on 8/9. December 2011 in the *Powerhouse Museum* (Sydney, Australia).

It was open to all museum visitors and with this to the general public. It attracted an interested crowd throughout this time and not all visitors could resist interacting with the robots in a more physical manner than just observation or gesture commands. Among the *Swarming Heads*, however, there was the notable

absence of a scenario one might have expected as the most likely based on the depiction of identical agents in popular fiction, that of all agents performing the same action at the same time. Technically, only minor algorithmic arrangements counteracted total behavioural uniformity. All decisions by the agent's central control system with regard to behaviour selection were probabilistic, albeit in a very simple manner: Stationary probabilities were assigned to the final behaviour options available after rule-based pre-selection (only within the attention system probabilities changed dynamically). But in combination with the environmental situatedness of the robot, this small intrusion of non-deterministic freedom caused constant asynchronic behaviour variation, even though over time the limited and identical behaviour repertoire of the agents became obvious through the re-appearance of similar behaviour patterns.

This is not to say, that no simultaneous collective behaviour ever emerged, but it needed a larger timeframe and specific conditions. We observed for instance the following anecdote:

During a quiet period in the museum with the conference attendees having returned to their session after a coffee break near the installation, two people (one of them the first author) remained in close proximity of the installation, absorbed in an ongoing conversation. On the pedestal the *Swarming Heads* were still bustling with movements and interjections, still 'excited' by the crowd of conference attendees present just a few seconds ago. The two people in their vicinity paid no attention to the robots, that is, they did not accommodate their behaviour in any way to that of the robots. However, the robots paid attention to the humans through mTHAMBS and continued to track their movements. Since mTHAMBS made them to attempt to approach the stationary people, the robots still constantly collided with each other—the ones in the second or third row with the robots in front of them—or the perimeter rim. However, when ending the conversation, the humans noticed with some surprise that all robots were staring at them, arranged in a cluster at the point on the pedestal closest to the chatting people, as if they were eavesdropping on the conversation. Occasionally the *Swarming Heads* still bumped into each other, but without breaking up the emerged formation: The overall pattern of activity had converged. Over a larger time period the instilled desire to approach people won over the disruptive avoidance behaviour following collisions. In the case of a single stationary people target, which was unresponsive to the robot's actions, the approach behaviour led to overall cohesion and created enough behavioural stability to overcome the disintegrative impact on synchronous behaviour patterning caused by collisions.

There seems to be little research on the relation between identical agents and emerging synchronous collective behaviour in robotics. As a striking contrast, in the field of agent-based simulations, the software-based virtual agents are almost always identical or at least resemble each other extremely closely. But they are in general at best superficially situated in their (virtual) environment. The environment is kept simple and mostly uniform since the aim is typically to uncover general mechanisms and boundary conditions of processes for which no analytical mathematical models exist or have not yet been discovered. Local variation of

the environment and a strong interaction of the agent with local specificities are not desirable since they would slow down the emergence of more general mechanisms. The simplification is acceptable if considered in the research design, but there are good reasons to assume that agents in the physical world are always engaged with the local variations of their environment. To overlook this would lead to flawed assumptions and deficient experimental research designs. If most of the employees of a firm arrive within a short time interval before 9 o'clock at the premises, it is not an indicator that the firm hires very similar people. It is the consequence of the firm's rule that regular work time starts at nine. It is the local constraint that produces the uniformity.

VI

In line with the observed behavioural diversity of our very simple identical robotic agents, we may consider two propositions by extrapolating to future more complex robotic agents:

(1) To make any judgement on the uniqueness of an intentional agent one would have to create an extended series of tightly controlled and exactly reproducible lab experiments and observe individual agents over a very long time period 'in the wild'.
(2) An intentional agent should not be assumed as an isolated entity, but as extending into the environment and into other agents. Boundaries are always only partial, differ in space and change over time. They are also conditional on the aspect under consideration.

Note that (1) is only a methodological issue in research with intentional agents (humans, animals, robots), while (2) constitutes a fundamental assumption about the interconnectedness and interdependency of agency. It goes much further than many other externalist views including Clark's external cognitive scaffolding.

But what would this interconnectedness mean concretely? Accounts in psychology that propose for instance human 'cognition beyond the brain' [17] are often clear and persuasive in their arguments against the internalist view, but slightly vague when describing what would replace the input/output information processing model. The same applies arguably to philosophical approaches. Interconnectedness is claimed and described as an all-encompassing mutual relationship between the agent and the environment. But the concrete examples given can be usually explained within an internalist view as well, requiring maybe a few more assumptions and in the worst case leading to the need of a representation of the entire world in the 'mind'. In fact, any situatedness, no matter how dominating and decisive, can always be accounted for in an internalist view by referencing mental representation and simulation. The externalist account alluded to above would be forced to go beyond the proposition of relations in which the agent is involved—no matter how deep this involvement is assumed to reach. Relations are

between entities, they have endpoints by definition and, thus, if the agent is one of the endpoints, it re-emerges as the potentially isolated, separable entity. In order to avoid this return of the encapsulated agent, one has to locate agency in the relations themselves, the relations between the body and the environment (including other bodies). It would run into the danger of creating yet another dualism, that of body/environment (the physical) and agency (the relational), but this would only be the case if the metabolic or hardware boundary is prioritised over all other boundaries and considered as defining.

At least with robots it is easy to see how the hardware boundary is simply one boundary among many: The hardware boundary dissolves already in a robot that is connected via wireless transmission to a cloud server on the Internet and via this server to other robots. In humans, robotic art that included cyborgs (defined as mixture of machine and human) and Internet connectivity such as the works of Neil Harbisson [18] and Stelarc (Chap. 20, this volume) venture out in the same direction. But as Stephens and Heffernan (Chap. 2, this volume) pointed out, this line of work of arts shows, what we already are, not something that we will become. Deteriorating mental health caused by solitary confinement [19] and drug-induced or mystic experiences of oneness [20] point in this direction, too, as do the importance of social behaviour in human evolution [21], the idea of distributed cognition enabling joint action of groups [22] and the discovery of mirror neurons in monkeys [23] and their assumed existence in humans [24].

Animals including humans are intentional agents from the onset; it is the machines which currently are lacking agency together with subjectivity. According to Roberto Marchesini referring primarily to animals but, of course, including humans '… subjectivity is arbitrariness, possibility, imagination, creativity, and partiality' [25]. These characteristics might or might not be achievable in machines, but if they are, it will happen in a still distant future. As Marchesini points out it would be a matter of machines very different from current ones and these new machines would be no longer under the control of the humans that created them.

The characteristics of subjectivity, however, might preclude identical reduplication even in machines; it might be a choice of either replicating identical agents or attaining subjectivity. These considerations are currently mere speculation since technology has not yet advanced enough to make even an educated guess. As mentioned above, the assumption of confined identifiable informational content in the brain might constitute an ill-guided perspective from the start, but even if not, we are more likely to approach tentative answers to questions of the relation between subjectivity, individuality and identity (as sameness) through research with robotic agents than in humans or other animals due to the latter's complexity.

There is a more fundamental assumption at stake here to which we already alluded above. If we cannot think of robotic agents as being one and being many at the same time, then there is even less of a chance to imagine this for humans. There appears to be no thinkable way of continuing one's life though transferring the information 'contained' in the brain because of the arising existential ambiguity (for other arguments in the same vein see [26]). Death would still take hold of

the individual despite the recreation of one or several perfectly similar but distinct new instantiations of the said individual. This is, of course, unless we are prepared to abandon the notion of seamless continuation of a person in general (or the concept of a self). Accordingly, at any moment in time the experienced presence might not have been uniquely connected to the experienced past and might not be uniquely connected to the subjective future. In doing so we would have to ignore ongoing processing in the biological body (including the brain) of humans and other animals during unconscious states. In case of the uploaded information content of the brain, we would have to assume that initial conditions do not matter or can be preserved and reproduced as well. Difficult if not impossible to imagine for biological agents, this might be acceptable for machines. These considerations are currently more in the realm of metaphysics, but—ironically—technology could make them a physical reality: If not a human or other animal, so at least a robotic agent might awake one day from sleep to find itself being more than one.

References

1. Clark A (2009) Dispersed selves. Leonardo Electronic Almanac 16(4–5)
2. Simondon G (2012) Technical mentality. In: De Boever A, Murray A, Roffe J, Woodward A (eds) Gilbert Simondon: being and technology. Edinburgh University Press, Edinburgh
3. Zwijnenburg PJ, Meijers-Heijboer H, Boomsma DI (2010) Identical but not the same: the value of discordant monozygotic twins in genetic research. Am J Med Genet Part B: Neuropsychiatric Genet 153(6):1134–1149
4. Smith M (ed) (2005) Stelarc: the monograph. The MIT Press, Cambridge, MA and London
5. Metzinger T (ed) (2000) Neural correlates of consciousness: empirical and conceptual questions. MIT press
6. Brambilla M, Ferrante E, Birattari M, Dorigo M (2013) Swarm robotics: a review from the swarm engineering perspective. Swarm Intell 7(1):1–41
7. Wallace RS (2009) The anatomy of A.L.I.C.E. In: Epstein R, Roberts G, Beber G (eds) Parsing the turing test. Springer, Netherlands, pp 181–210
8. Kroos C, Herath DC, Stelarc (2011) From robot arm to intentional agent: the articulated head. In: Goto S (ed) Robot arms. Advances in robotics, automation and control, InTech, pp 215–240
9. Kroos C, Herath D, Stelarc S (2012) Evoking agency: attention model and behaviour control in a robotic art installation. Leonardo 45(5):133–161
10. Andreopoulos A, Tsotsos JK (2013) 50 Years of object recognition: directions forward. Comput Vis Image Underst 117(8):827–891
11. Yang H, Shao L, Zheng F, Wang L, Song Z (2011) Recent advances and trends in visual tracking: a review. Neurocomputing 74(18):3823–3831
12. Herath DC, Kroos C, Stelarc (2012) Encounters: from talking heads to swarming heads. In: Proceedings of the seventh annual ACM/IEEE international conference on human-robot interaction. HRI '12, pp 415–416
13. Thrun S, Burgard W, Fox D (2005) Probabilistic robotics. MIT press
14. Herath DC, Kroos C, Stevens C, Burnham D (2013) Adopt-a-robot: a story of attachment (or the lack thereof). In: Proceedings of the Eights Annual ACM/IEEE international conference on human-robot interaction. HRI '13, Tokyo, Japan
15. Moon Y (1999) The effects of physical distance and response latency on persuasion in computer-mediated communication and human–computer communication. J Exp Psychol Appl 5(4):379–392

16. Delaherche E, Chetouani M, Mahdhaoui A, Saint-Georges C, Viaux S, Cohen D (2012) Interpersonal synchrony: a survey of evaluation methods across disciplines. IEEE Trans Affect Comput 3(3):349–365
17. Cowley SJ, Vallée-Tourangeau F (2013) Cognition beyond the brain. Computation, interactivity and human artifice. Springer, London
18. Jeffries S (2014) Neil Harbisson: the world's first cyborg artist. The Guardian. http://www.theguardian.com/artanddesign/2014/may/06/neil-harbisson-worlds-first-cyborg-artist. Accessed 03 Oct 2015
19. Grassian S (2006) Psychiatric effects of solitary confinement. Wash. UJL & Pol'y, 22, 325
20. Goodman N (2002) The serotonergic system and mysticism: could LSD and the nondrug-induced mystical experience share common neural mechanisms? J Psychoactive Drugs 34(3):263–272
21. Tomasello M, Carpenter M, Call J, Behne T, Moll H (2005) Understanding and sharing intentions: the origins of cultural cognition. Behav Brain Sci 28:675–691
22. Hutchins E (1995) Cognition in the wild. MIT press
23. Kohler E, Keysers C, Umilta MA, Fogassi L, Gallese V, Rizzolatti G (2002) Hearing sounds, understanding actions: action representation in mirror neurons. Science 297(5582):846–848
24. Fabbri-Destro M, Rizzolatti G (2008) Mirror neurons and mirror systems in monkeys and humans. Physiology 23(3):171–179
25. Marchesini R (2015) Dialogo ergo sum: subjectivity, posthuman and nonhuman alterities. In: Curtin university CCAT symposium 'what is philosophical ethology?', Perth, Australia
26. Hauskeller M (2012) My brain, my mind, and I: some philosophical assumptions of mind-uploading. Int J Mach Conscious 4(01):187–200