AI and Ethics (2022) 2:115–128
https://doi.org/10.1007/s43681-021-00076-x

ORIGINAL RESEARCH

The ethics of interaction with neurorobotic agents: a case study with BabyX

Alistair Knott[1,2] · **Mark Sagar**[1,3] · **Martin Takac**[1,4]

Received: 29 March 2021 / Accepted: 19 June 2021 / Published online: 8 July 2021

Abstract
As AI advances, models of simulated humans are becoming increasingly realistic. A new debate has arisen about the ethics of interacting with these realistic agents—and in particular, whether any harms arise from 'mistreatment' of such agents. In this paper, we advance this debate by discussing a model we have developed ('BabyX'), which simulates a human infant. The model produces realistic behaviours—and it does so using a schematic model of certain human brain mechanisms. We first consider harms that may arise due to effects *on the user*—in particular effects on the user's behaviour towards real babies. We then consider whether there's any need to consider harms from the 'perspective' *of the simulated baby*. The first topic raises practical ethical questions, many of which are empirical in nature. We argue the potential for harm is real enough to warrant restrictions on the use of BabyX. The second topic raises a very different set of questions in the philosophy of mind. Here, we argue that BabyX's biologically inspired model of emotions raises important moral questions, and places BabyX in a different category from avatars whose emotional behaviours are 'faked' by simple rules. This argument counters John Danaher's recently proposed 'moral behaviourism'. We conclude that the developers of simulated humans have useful contributions to make to debates about moral patiency—and also have certain new responsibilities in relation to the simulations they build.

Keywords Ethics of avatar use · Computer game violence · Neural models of emotions · Ethical behaviourism

1 Introduction

While most Artificial Intelligence (AI) systems perform specific tasks, such as playing Atari games or translating sentences, an increasing number aim to provide a more 'complete' model of a human. These more holistic AI agents have humanlike bodies, implemented in robot hardware or computer graphics, as well as a range of humanlike abilities. They are often able to express a range of humanlike emotions, to engage in dialogue with a human user, and to perform a range of tasks in their real or simulated environments.

AI agents that simulate 'whole humans' have always been a focus for ethical discussion. Theoretical discussions of ethics have often imagined a simulation of a real person that is indistinguishable from a real person, and then explored the ethical status of this simulated person (see e.g. [35]. AI is still far from producing perfect simulations of people. But as AI advances, models of simulated humans are becoming increasingly realistic. A new debate has now arisen, about the ethical status of *actual* simulations of people, in the sphere of practical ethics, rather than theoretical ethics. AI researchers building realistic simulations of people need to start paying attention to this new debate.

Simulated humans are being developed in two distinct AI research programmes. An *engineering* research programme aims to build AI agents that can usefully collaborate with human users on various tasks. The push for embodied realism here is because humanlike agents are arguably easier to interact with. The ethical questions for these agents relate to the safety of AI-related products, to dangers of misuse, and to a wide variety of social impacts. A *scientific* research programme aims to study human (or animal) cognition by simulating the brain, and its interactions with the body. These

✉ Alistair Knott
alistair.knott@otago.ac.nz

1 Soul Machines Ltd, Auckland, New Zealand

2 Department of Computer Science, University of Otago, Dunedin, New Zealand

3 Auckland Bioengineering Institute, University of Auckland, Auckland, New Zealand

4 Centre for Cognitive Science, Comenius University, Bratislava, Slovakia

Fig. 1 Screenshot of an interaction with BabyX

simulations are called *neurorobotic models*. The push for embodied realism in these models is because good simulations arguably produce more realistic behaviours. The ethical questions for neurorobotic agents focus on the implications for building computer models that explicitly aim to reproduce the functionality of human or other biological brains [2]. The ethical issues within these two strands of AI are a subject of much recent discussion (see [20] for a review).

Our aim in the current paper is to contribute both to the engineering AI ethics debate and to the neurorobotics ethics debate, by discussing a concrete example system that we have produced in our company, Soul Machines. The system, called BabyX, is a simulation of an 18-month-old baby (see Fig. 1). The BabyX model includes a graphics component (a realistic face and body, with a range of realistic behaviours), and a brain modelling component (a neural network model of various cognitive processes operative in a baby of this age). Our aim in developing BabyX has distinct scientific and engineering objectives. On the scientific side, BabyX is a platform for neurorobotics research: it allows us to implement embodied models of human brain mechanisms, and to test these observing whether the simulated baby's behaviour is similar to that of real babies. While our models are still fairly simple, BabyX is already quite a convincing simulation of a real baby: a user can interact with her in various natural ways, and her responses are also quite natural, both at the graphical level and at a more cognitive behavioural level. She[1] can see and hear the user via video and audio feeds; she can see and interact with objects in her own simulated environment, she can learn words and actions. She also manifests various emotional behaviours in response to events she perceives: she can smile, laugh, cry, get cross or frustrated. On the engineering side, cognitive models that produce realistic baby behaviours can be adapted for use in the commercial avatars our company develops, for instance in modelling the emotions of these avatars, and nonverbal behaviours that manifest these emotions.

In this paper, we will discuss the ethical issues that arise with BabyX, in the light of current discussions around engineering and brain-based AI ethics. We argue the concrete case of BabyX allows us to advance this discussion in a number of ways.[2]

We begin in Sect. 2 by introducing BabyX, and making the case that there are ethical issues to be considered. In Sect. 3 we describe our motivation for creating BabyX, to explain why we are pursuing this project in the first place. In the rest of the paper, we embark on the ethical discussion proper, which is framed around a hypothetical scenario in which a human user 'mistreats' BabyX. In Sect. 4, we introduce a conception of what it is to 'mistreat' a human baby. In Sect. 5, we consider whether a user's mistreatment of BabyX has any ethically significant effects *on the user*. In Sect. 6, we consider the philosophically more loaded question of whether such mistreatment has any ethically significant effects *on BabyX*. In Sect. 7, we use our conclusions from these latter two sections to state a working ethical policy for the BabyX project.

2 Features of BabyX that highlight ethical questions

BabyX is, in essence, a computational model of the human brain, and its interactions with the body. Cognitive scientists produce thousands of such models every year. They do not normally scrutinise their models ethically. (They would probably baulk at the very idea of doing so.) So why should BabyX be any different? It may be that we are being overcautious by considering ethical questions. (And there are harms associated with being overcautious, as discussed by [7]. But ethical status is assigned on a sliding scale in our society, with humans at one end, and quite simple animals at the other: as brain models improve, comparisons with the simpler animals on this scale may be increasingly warranted. In addition, we think BabyX has some distinctive features that make ethical considerations particularly prominent, in comparison with traditional brain models. In this section, we will review these distinctive features.

[1] We will use the pronoun 'she' to refer to BabyX throughout this article. This reflects our company practice, which is to describe BabyX in anthropomorphic terms. Indeed it's very hard not to do so, when confronted with the simulation—a fact which emphasises the need for an ethical analysis.

[2] Ethical issues also arise for our commercial dialogue agents, which all simulate adults. We won't discuss those here; we have a separate ethics policy that covers these systems (Soul Machines [34]). The current paper just considers BabyX, which is a research system, and not deployed as a product.

2.1 BabyX is an avatar

First, BabyX is an *avatar,* rather than just a brain model. BabyX has a realistic graphically rendered face and body, as well as a simulated brain. And she's designed to be interacted with by a human user, rather than just to train on files of data. There are many interesting brain models of this general type in cognitive science (see [8, 27], for recent examples)—but BabyX is thus far unique in its degree of physical realism, and in the naturalness of the interactions it affords.

There's already an active discussion about the ethics of human interactions with avatars, or simulated agents, or robots. Particularly influential is a recent argument by John Danaher [13] that the moral status of robots should be determined on the basis of their behaviour, without reference to the algorithms that animate them. Given that our avatar models a vulnerable member of society—a baby—engaging with this discussion is of particular importance.

2.2 BabyX's brain model

Secondly, BabyX's brain model focusses on three mechanisms that are particularly relevant to the philosophical question of whether the avatar has some degree of 'moral patiency'—that is, of whether the avatar has mental states of any kind (feelings, experiences) that are in any way comparable to those of people.

2.2.1 An embodied model of cognition

BabyX implements an 'embodied' model of human cognition. Such models posit that the structure of the human cognitive system is strongly influenced by the human body's apparatus for sensory perception and motor control (see [5, 9]. According to an embodied model, for instance, the fact that the eye has a fovea, which perceives items sequentially, rather than all at once, has important consequences for the architecture of the cognitive system. So too does the fact that motor movements of the hands are typically guided by visual fixations so that reaching for target objects typically requires they are first visually attended to. The reason we model BabyX's body as well as her brain is so we can simulate processes like visual attention and visuomotor coordination, which deliver perceptual information and copies of motor signals to the simulated brain in the same temporal patterns as are found in the actual human brain. That is, we simulate BabyX's body because we think this helps us to design a brain model with the right 'interface' to the external world.

But the fact that BabyX has a body may also be of some ethical significance. For instance, in our model we can also simulate the tactile mechanisms through which pleasurable or painful stimuli are received on the body, and motor mechanisms which respond to such stimuli, for instance through recoil or startle reflexes. Perhaps a brain model that includes these bodily mechanisms is different in some way from one with a more abstract simulation of 'reward' and 'punishment'.

2.2.2 A biologically inspired, embodied model of emotions

Our brain model pays particular attention to simulating the emotional circuits of the mammalian brain, as we will elaborate in Sect. 6.4. And our model of emotions is strongly embodied, following the work of Panksepp [28] and Damasio [11], as we will discuss. Perhaps an agent implementing a biologically inspired model of emotions is more deserving of moral consideration—particularly so if it is embodied.

2.2.3 A model of episodic memory

BabyX's cognitive model includes a model of episodic memory: she can remember the events and states she experiences, and she can retrieve a sequentially structured series of events and states in the form of a simulation. She can use her memories of generic event sequences to make predictions about upcoming events and states. She also learns emotional associations of events, and her memory for events is influenced by these associations. Her storage of events is weighted towards events with strong emotional associations. (And also towards events that are unexpected—which generate a particular emotion, 'surprise'). Details of these mechanisms are given in Takac and Knott [38], which introduces a precursor to BabyX's episodic memory model.

Arguably, having a memory helps give an agent special moral status. For instance, such agents can anticipate bad things happening, or recall traumatic experiences. Furthermore, a common idea in cognitive science is that episodic memory is necessary in order to develop a humanlike 'sense of self' (see e.g. [11, 39], which is often seen as another important criterion for moral status. We will expand on these ideas in Sect. 6.4.

2.3 A model in development

Note that we are planning on *extending* and *improving* BabyX, both as a realistic human avatar (c.f. Sect. 2.1) and as a realistic model of the infant brain (c.f. Sect. 2.2). If these features of our model are ethically significant, the moral issues we face may become increasingly important as the model develops. Or they may start to become important as the model passes a certain level of sophistication or

accuracy in some relevant respect. This means we have to think not just about BabyX as she currently is, but as she will be in the future if our models go to plan.

We will refer in detail to BabyX's brain model in Sect. 6, when considering her ethical status. But we begin by giving some background about BabyX, and the ethical issues we will consider.

3 Motivation for the BabyX model

We want to start by stating our motivation for the BabyX project in some more detail. Why are we even trying to build a realistic model of a human baby, if this project may raise ethical issues?

One motivation is purely scientific. It's an important scientific goal to understand how the human brain works: perhaps one of the most important goals in current science. Many scientists are tackling this goal, in psychology, neuroscience and cognitive science. Our ethical assumption is that this is a good goal, all other things being equal. We are pursuing a particular approach to this goal, which is to build a model of a human, covering both brain and body, and including a rich model of emotions, for the reasons described in Sect. 2. We chose to build a model of a human baby because the learning that happens in the human brain starts during babyhood: this is when people learn what objects are [4], the difference between 'oneself' and 'other people' [23], how to recognise human actions and goals [41], what language is [40], and many other foundational human abilities.

Another motivation is commercial. Our current commercial product is an adult human–machine dialogue agent, called a Digital Person™. This agent is built using engineering AI methods, rather than brain modelling methods. However, we have incorporated several components of BabyX's brain model, in particular relating to the motor system driving eye and body movements, and to the perceptual and emotional system, and we plan to incorporate more of these over time. Our aim is for our Digital Person product to be as humanlike as possible; we aim to move towards this goal by progressively incorporating elements of our brain model. In commercial terms, our motivation for the BabyX project is ultimately to produce a more useful and useable dialogue agent product. We have also considered creating products featuring BabyX itself: this is not a central commercial goal, but there are possible products around a developmental psychology research tool, a pediatric simulator, parenting training, and other applications of this kind.

In the current paper, we consider BabyX both as a scientific project, and as a potential piece of technology. On the scientific side, even if it's a good goal to find out about the human brain, not all means to this end are justifiable. Certain experiments on humans or animals might be unjustifiable, for instance, because of the harms they cause on their participants. So where does BabyX fall in relation to that question? On the technological side, deploying BabyX (or components of BabyX's brain model) to the public could have a range of beneficial and harmful consequences. What are these consequences? Are any forms of deployment justifiable?

4 What is it to 'mistreat' a baby?

In our discussion of ethical issues with BabyX interactions, we will focus on a scenario where a human user 'mistreats' BabyX—for instance, by making her sad or angry, in ways that would be deemed unacceptable for a real baby. But before we start, we'd like to make two points about mistreatment.

First, there are certain situations where adult supervisors induce or allow negative emotions in real babies, where we wouldn't want to accuse them of 'mistreatment'. 'Normal life' for real babies involves them going through a mixture of emotions, which are only partly controllable by supervising adults. In fact, it's very important that babies experience negative emotions in some circumstances—for instance, when learning by trial and error. Indeed, parents have a role in shaping babies' learning which sometimes requires them to induce negative emotions in the baby, with reprimands of various kinds. Parents are also normally understood as having a duty to let babies learn certain things 'for themselves'—that is, to give them some autonomy in their learning. So a parent who induces a negative emotion in a baby as part of a justifiable learning experience, or who fails to prevent a negative emotion by allowing a baby to act autonomously in some controlled, safe environment is not guilty of 'mistreatment'. We believe interactions with BabyX that fall into these categories are perfectly acceptable: we are simply simulating 'normal life' for the baby. As an example, BabyX can get upset if she fails to achieve a task she is attempting. She can also recognise 'cross utterances' produced by the user, which are utterances with certain auditory profiles in the domains of intensity and timbre, and the user can use these utterances to dissuade the baby from the action she is currently performing. In the right circumstances, both these things are perfectly acceptable.

Second, in the context of scientific experiments on babies, there are certain interventions on babies that cause them a (very small!) amount of distress, that are also judged to be acceptable, because of the scientific benefits the experiment is expected to bring. An example is the 'stillface' experimental paradigm, where a baby's caregiver looks at the baby without reacting for a short period of time. This unusual parent behaviour is stressful for the baby, but the baby's

reactions give insights into the role of parental facial expressions and responses in its experience of the world (see e.g. [1]. In modern psychology, experiments on human subjects are always subject to approval by ethics boards, and standards for approval on children and babies are particularly stringent. If negative emotions are induced in a baby as part of an ethically approved scientific experiment, we will not consider this as 'mistreatment'. (We should note that the still face protocol specifies the experiment will be cut short if the baby is 'unduly' distressed—though the definition of 'unduly' is a subjective judgement by the researcher administering the task.)

With these preliminaries, we will entertain a case of 'genuine' mistreatment of a baby: for instance, deliberate meanness which is not sanctioned on any of the above grounds. Mistreatment of a real baby in this way is, by our definition, ethically wrong. But what if a user mistreats BabyX in this way? Is this also ethically wrong, or is there no ethical issue here?

We will consider potential ethical effects on the human user of BabyX in Sect. 5, and potential ethical effects 'on BabyX' (if that is a meaningful idea) in Sect. 6.

5 User-related ethical issues for BabyX

Let us begin by assuming there are no grounds for thinking of BabyX as having any degree of moral patiency: assume the simulated baby is 'just a computer program', towards which human users have no obligations. Even on that assumption, the way a human user treats BabyX may be of ethical significance, because of its effects *on the user*.

The key point is that from the user's perspective, BabyX is *like* a real baby in some ways. So if the user 'mistreats' the baby, this will feel to the user a little like mistreating a real baby. Are there ethical problems with that?

A large literature is growing up around the topic of user mistreatment of avatars. A lot of this has centred around 'sex robots' (see e.g. [12, 37], or of 'bullying' general-purpose avatars (see e.g. [21], or of 'simulated murder' of characters in computer games (see e.g. [24]. If a user mistreats a sex robot, or bullies a dialogue avatar, or kills a baddie in a computer game, does this have ethical effects on the user? We'll frame the debate in relation to BabyX, but all these domains may have potential relevance.

5.1 'Extrinsic' moral effects on users

The obvious problem is a user's mistreatment of an avatar may affect their behaviours towards real people [13] calls such effects 'extrinsic'. The effect on other people could either be bad or in some scenarios even good. We will consider both possibilities.

5.1.1 'Bad' extrinsic effects of avatar mistreatment

A user who mistreats BabyX may become habituated to a general pattern of behaviour which then extends to real babies. The general possibility that violent treatment of avatars in simulations or video games transfers to aggressive or antisocial habits towards people in real life has been explored in many empirical experiments. However, the results are confusing: even at the level of large meta-analyses, there are some studies that find evidence for transfer (see e.g. [3, 18, 19]; and some that find only minimal evidence (see e.g. [14, 17]. Given this impasse, we think it's safest to err on the side of caution and take seriously the possibility that mistreatment of BabyX may have bad extrinsic effects.

Caveats from meta-analyses aside, we are struck by studies finding evidence that the degree of graphical realism in a violent video game is a factor in how much aggression transfers into the world (see e.g. [6], as is the degree of 'immersion' of the user in the game (see e.g. [22, 29]. Some studies focus on the 'behavioural realism' of computer characters, rather than their graphical realism (see e.g. Zendle et al. [43]). The BabyX simulation scores highly on all these measures. The graphical depiction of the baby is highly realistic. BabyX's behaviours are also accurately simulated, both at the level of individual gestures and larger behavioural units. And the BabyX interface puts the user physically close to the baby, in a position similar to a parent or caregiver. (The user interface actually allows the user to 'stroke' the baby: our simulation of the baby's tactile sensory system includes a model of the skin mechanoreceptors specialised for sensing stroke gestures). These factors make us specially wary of the possibility that the mistreatment of BabyX may have bad extrinsic effects.

5.1.2 'Good' extrinsic effects of avatar mistreatment

There is actually another possibility, which is unpalatable but should be noted. Users who are inclined to mistreat actual people may be able to satisfy this inclination by mistreating avatars and hence avoid harming real people. Danaher [12] argues that the possibility of using avatars therapeutically in this way 'should be actively and carefully researched'. We acknowledge this possibility, but such therapies are remote from our company's sphere of operations, and we certainly won't be exploring them.

5.1.3 'Good' extrinsic effects of positive interactions with the avatar

Avatar mistreatment is an important topic, but it is also important to consider that many users are likely to have largely positive interactions with BabyX. These may have very positive extrinsic effects: for instance, BabyX could

possibly be helpful in teaching childcare principles, or giving parents-to-be some idea of what interacting with babies is like. (We can also imagine applications designed to show teenagers the reality of caring for a baby and encouraging them to think carefully about parenthood. There's some possibility these would involve some mistreatment of BabyX, so we would have to be very sure of their benefits before sanctioning them).

We should also note there are potentially good extrinsic effects of simulations where BabyX exhibits pain or suffering. We are thinking particularly of training scenarios, where a medical student learns how to treat a baby who is suffering. We want medical staff to have practice in such scenarios before they deal with real cases; BabyX may provide a realistic platform for acquiring the relevant skills.

5.2 Morally relevant extrinsic effects on users

'Extrinsic' effects of interactions with BabyX don't have to arise from *mistreatment* of BabyX, and they don't have to be harms *of other people*. There are also possible harms on users themselves, that apply regardless of how BabyX is treated. An important potential harm to consider is that a user becomes *emotionally invested* in BabyX. This scenario is particularly plausible for users who have recently lost a baby, or are unable to have babies of their own: such users are emotionally vulnerable, and there's potential harm in their becoming attached to something that cannot truly reciprocate. To minimise these potential harms, it might be helpful to tell users in advance that they will be interacting with a robot, not a real child. (This is in line with our general ethics policy, which already disallows us from deploying agents that pretend to be real people).

5.3 'Intrinsic' moral effects on users

There are also possibly 'intrinsic' effects of user mistreatment of an avatar: they may be wrong 'in themselves', even if they have no effect on the user's behaviour towards real people. As Danaher notes, it's hard to maintain this position without falling into awkward moralistic positions (that certain things are 'just wrong').

Of course, the extent to which a user sees an avatar as being 'like' a real person is a matter for debate. There's actually some empirical evidence that actions directed towards an avatar that look like 'bullying' are somewhat different from similar actions directed towards a real person (see again Keijsers and Bartneck, 2018), so this is a matter for further exploration.

5.4 User-related ethical issues: some practical conclusions

Should we allow public access to BabyX? While the argument that mistreating BabyX is 'intrinsically' bad is hard to make, we can't rule out that mistreatment may adversely users' behaviour towards real babies, judging from empirical studies of computer games. There are also potentially beneficial applications we could develop with BabyX. For now, we judge that the risks mostly outweigh the potential benefits. This means we must exercise caution in how we make BabyX available to the general public.

There are various options here.

- One option is to restrict the BabyX simulation, in ways that categorically prevent certain types of mistreatment. For example, we can disable the interface that allows the user to 'stroke' the baby, or implement tight restrictions on it. (Our current system omits any graphical presentation of user hands, so there is no graphical depiction of the stroking gesture; and stroking is constrained to be performed on the baby's upper arm, which is understood as a 'safe' type of contact by caregivers; see for instance [25]. But interface design can't eliminate all forms of mistreatment, if we wish to simulate 'normal baby life', of the kind discussed in Sect. 4. For instance, occasional 'cross utterances' are a feature of normal baby life, so the interface should allow these—but repeated cross utterances probably do constitute mistreatment in most circumstances.
- Another option is to positively define a set of user behaviours that constitute mistreatment, and run a detector that checks for these, shutting the session down as soon as any indication of mistreatment is detected.
- A final option is to restrict public access to BabyX, except under carefully monitored conditions. In fact, this is the current situation: BabyX is currently only used by its developers, and by subjects participating in child development experiments, which are ethically approved, and closely overseen.

For the moment, the most practical of these options is the latter one, and this is the one we currently adopt. But as discussed in Sect. 5.1, BabyX could also be used in applications that are beneficial for society, on balance. If we ever considered these, we would want to build in constraints that prevent mistreatment, as far as is possible.

6 Agent-related issues for BabyX

We now turn to the question of whether mistreating BabyX is bad in any way *for BabyX*. Does BabyX have anything like 'real' human mental states? In particular, does she have anything like 'real' feelings or emotions? This question is in the province of the philosophy of mind. But if the answer is affirmative, there are consequences in the province of ethics. If BabyX does have something like 'real' feelings, then users are under a duty not to mistreat her, because of effects it will have *on her,* quite separately from any effects it has on them, or on other people.

The question as to whether BabyX has anything like human (or animal) feelings can be approached from any number of philosophical angles. Our approach will be to consider a very recent conception of feelings from John Danaher, which was developed specifically to think about the moral rights of avatars.

6.1 Danaher's ethical behaviourism

Danaher [13] makes a particularly strong proposal about the status of avatars as moral patients. His claim is that avatars can have 'moral status' purely on the basis of their behaviours: if their behaviours are close enough to those of people, then this by itself is sufficient to grant them moral status too. In fact, his argument is broader: if an avatar's behaviour is sufficiently close to any agent to whom we grant moral status, then we should also grant the avatar the same moral status. Drawing on earlier arguments by Sparrow [36], his main argument is as follows:

> a sufficient *epistemic ground* or *warrant* for believing that we have duties and responsibilities toward other entities (or that they have rights against us) can be found in their observable behavioural relations and reactions to us (and to the world around them). It is the ethical equivalent of the Turing Test (...).

Danaher says he is making a 'normative and epistemic' claim about the mental states of avatars, rather than a metaphysical one. He allows that the ultimate ground for giving humans (and animals) moral status is likely that they are *sentient*: they actually feel things. (In other words, they have some form of consciousness.) His point is that people's *evidence* about the sentience or consciousness of other agents comes through their behaviour. Danaher terms his position 'ethical behaviourism'.

Having set out this proposal, Danaher suggests we should decide on the ethical status of robots by *comparing their behaviours* to those of humans, or other animals. The behavioural test he proposes is deliberately broad-brush: if a robot's behaviours are 'roughly performatively equivalent' to those of humans, we should accord the robot the same moral status as humans. (And if they are roughly equivalent to those of some lower animals to whom we accord reduced moral status, we should accord the robot the same moral status as that animal). A broad-brush comparison is necessary because behaviours are intrinsically complicated: no two agents will have *exactly* the same behaviours. For a robot to qualify for humanlike ethical status, presumably its behaviours should fall within the range of normal behaviours shown by humans. Danaher does not say this explicitly, but he emphasises that the range of qualifying behaviours is pretty broad.

Danaher uses the word 'behaviourism' to hark back to the behaviourist psychologists of the mid twentieth century. But actually, his definition of the 'observable behaviour' of an agent is far broader than theirs, in that it includes the agent's brain states and brain activity. The original behaviourists aimed to explain agents' bodily behaviours, with reference to the perceptual stimuli that occasioned them, and without reference to their 'internal states'.[3] They did allow some role for neuroscience in these explanations, but we would be misrepresenting them to see them as putting brain states and bodily behaviours on an equal footing. For clarity, we'll define two types of ethical behaviourism: a 'narrow' variety, which holds that an agent's external physical behaviours provide sufficient ground by themselves for deciding we have duties towards them, and a 'broad' variety, which holds that physical behaviours plus internal (brain or computer) states provide a sufficient ground. The broad variety allows the expression of a functionalist account of feelings and other ethically relevant mental states, in which the mechanisms internal to an agent that generate its behaviour are of some relevance. For instance, Putnam's [30] classic functionalist conception of the mental state 'pain' could be expressed within Danaher's 'broad' version of behaviourism. For Putnam, an agent capable of feeling pain must have a certain style of 'functional organisation', minimally featuring sensors for detecting certain stimuli, mechanisms for assigning value to stimuli (including through learned associations), and mechanisms for generating behaviours based on these valuations.

Although Danaher entertains a broad definition of 'observable behaviour', he actually leans towards a narrow ethical behaviourism in most of his paper. He concedes the possibility of broad ethical behaviourist positions, that make reference to agents' internal states (and to brain states in particular), but he is sceptical about how successful such positions will be. He warns against a 'biological mysterianism',

[3] Actually, B.F. Skinner did allow some role for neuroscience in explanations of behaviour (see e.g. [44]. But his basic attitude towards references to brain states in such explanations was critical.

that accords special status to biological organisms—and on this we are in full agreement. But he is also sceptical we will be able to say much about brain mechanisms *as algorithms*, for instance using Putnam-style functionalist definitions of ethically relevant mental states. His position here is basically that we know so little about how brain states relate to morally significant metaphysical states like sentience and personhood that our epistemic evidence for an agent's moral status has to come directly from observable behaviours, without making any reference to brain states. We would like to challenge that position.

6.2 Assessing ethical behaviourism with implemented agent models

We'll express our argument against narrow ethical behaviourism by comparing the commercial dialogue agents (called 'Digital People') we produce in our company with BabyX. Both Digital People and BabyX can produce a range of realistic emotional behaviours: in fact, they can both pass the ethical Turing test at some level, at least for some users. But they produce these emotional behaviours through very different mechanisms. We argue that this difference in mechanisms leads to an ethical difference between BabyX and Digital People: we think Digital People are clearly not moral patients, while the case is less clear for BabyX.

Our argument extends an argument against narrow ethical behaviourism by Jilles Smids [33]. Smids picks up on Danaher's positioning of ethical behaviourism as an 'epistemic' theory, about how we can know about an agent's moral status. He argues that ethical behaviourism is best construed as relying on an abductive inference process, whereby an observer seeks the best explanation for the agent's emotional behaviours. For humans, and other animals, the best explanation is typical that the agent is experiencing mental states, of the kind that give it some moral status. Smids argues that for a robot agent that is *designed* to interact naturally with people, there is normally a much better explanation for any emotional behaviours it produces, which is that its designers *intended* it to have those behaviours. This explanation is off-limits in narrow ethical behaviourism—so that theory doesn't correctly capture our ethical intuitions. Instead, Smids adopts a broad ethical behaviourism, where 'what goes on inside does matter', so that 'designed' ethical behaviours can be ruled out of moral consideration.

Our implemented agent models provide useful case studies for discussing ethical behaviourism, because they sidestep some parts of the epistemic problem it purports to address, and so reformulate this problem in interesting ways. We have access to the code that causes emotional behaviours for our agents, so we have more information about their origins. The epistemic problem that remains is arguably different for the Digital Person and for BabyX. BabyX is designed to simulate certain aspects of the brain's emotional system. The key epistemic question here is: does *this specific brain model* give BabyX any status as a moral patient? This question focusses on the gaps in our knowledge about the brain mechanisms that produce emotional behaviours in real people: are BabyX's simplified models of these mechanisms enough of a match to these mechanisms to qualify BabyX for some moral status? The Digital Person is built to a far more behaviouristic design brief: it is designed to simulate certain realistic emotional behaviours. For Smids, this is enough to discount these behaviours in an assessment of moral patiency. But we are in a position to ask a further question: is there anything about the *mechanism* producing the Digital Person's emotional behaviours that identifies them as ethically unimportant? In particular, is it different in some relevant way from the mechanism producing BabyX's behaviours, or those of a human or animal?

Danaher, in fact, devotes some time in his paper to discussing the hypothetical case of a robot that obtains ethical status through 'subterfuge', by 'faking' emotional behaviours. The existence of such a robot would be an argument against narrow emotional behaviourism, at least. Danaher argues that such faking will ultimately be revealed in the robot's external behaviour, to someone assessing its ethical status. He essentially argues that if the robot's behaviour can't be distinguished from that of an agent truly deserving of moral status, then *we can't speak of faking*: or rather, any account of faking would be shaky because it would rest on our poor understanding of how the brain implements mental states. His expression of this point 'gets to the heart of the ethical behaviourist stance'. In this paper, we'd like to consider the possibility of faking more concretely, by framing the discussion around implemented emotional agents whose algorithms are well understood.

6.3 An agent that 'fakes' a behavioural claim to ethical status

The 'Digital Person' product we produce in our company is a web-based agent that interacts with a human user through a webcam and microphone. The agent is a very realistic graphical simulation of a person: she looks like a person, and has a range of realistic nonverbal behaviours (facial expressions and body movements), designed by a human animator. The product is largely built using engineering AI methods, but as discussed in Sect. 3, it also incorporates several components of our brain model. However, it is possible to disable these brain model components, to create a less autonomous and more controllable agent we call the 'Level 2 Person', which we use for some applications. Importantly, this stripped-back Level 2 Digital Person still produces convincing simulacra of many human emotions, and other behaviours. Here we

will describe the Level 2 Person, so as to make the greatest contrast with BabyX's model.

Our Level 2 Person can engage in a dialogue with the human user. The algorithm managing the dialogue is, at base, a large set of 'if–then' rules. Each rule maps an incoming *user utterance*, occurring in a specified *dialogue context*, onto a *response utterance*, and an accompanying new dialogue context. These rules are specified by a human 'script author'. The author defines a set of dialogue contexts, and for each context specifies a set of possible *utterance types* to expect from the user in that context. The author also trains an *utterance classifier* for each context, by providing copious examples of the utterance types defined for that context. The script author can also define emotional gestures (facial expressions, body gestures) to accompany each avatar response utterance.

This dialogue system is supplemented with a simple 'emotional system', that is quite different from BabyX's brain-inspired model. It consists of a circuit that classifies the user's current emotion, using evidence from the words and acoustic features of the user's current utterance, and from the user's current facial expression, and then responds to this with an emotional gesture, again using a set of hand-authored rules (for instance, 'if user is happy, be happy'; 'if user is angry, be worried'). The data that train the emotion classifier are assembled by still more human authors, who label utterances and video images with the relevant emotion categories.

If the human authors are resourced well enough, a system of this kind is able to handle a large number of dialogue contexts, and recognise a very large number of user utterance types. It can also become good at recognising users' emotions from their verbal and nonverbal behaviours—and good at producing emotional behaviours of its own, which are sensitive to user inputs. Behaviourally, the complete system presents a fairly convincing impression of an emotionally capable human agent.

The critical question, of course, is whether a system like this has 'rough performative equivalence' with any class of biological agent to which we accord moral status. We now turn to this question.

To begin with, we should emphasise the behavioural criterion for 'moral status' is much less stringent than the criterion for 'intelligence' enshrined in the classical Turing test. To qualify for moral status, a system doesn't need to demonstrate adult verbal intelligence. The ethical Turing test just requires behaviours indicative of the possession of genuine feelings. The behaviours of children and babies are sufficient indication of this—and for many people, so are the behaviours of at least some nonhuman animals. Moreover, we should recall that success or failure in the classical Turing test is very dependent on the judge making the decision. Many dialogue systems now pass the Turing test quite frequently, if judges are drawn from the general population. Our Level 2 Person probably also passes the ethical Turing test, for some judges. Certainly in user trials, we often find users who are worried they have hurt the agent's feelings, or who bond with the agent in some way. (The 'girlfriend' and 'boyfriend' avatars who bond with Japanese teenagers probably pass the ethical Turing test at some level too). Of course, we could require judges to have some understanding about the systems involved; such judges will be much more rigorous in their investigation. But this restriction feels a lot less acceptable for the ethical version of the Turing test. Do we really have to say that only some technically skilled people have the ability to determine ethical status? In summary: by Danaher's behavioural standards, the Level 2 Person as described here probably qualifies for some small degree of moral status.

We wish to argue, however, that our Level 2 Person does not deserve any moral status: a conclusion we think our readers would share. But we would rather not just appeal to the 'designed' character of their emotional behaviours, as Smids does: we would rather make our case based on the algorithm that generates these behaviours. To frame this argument, we have to contrast the Level 2 Person algorithm with some other algorithm, which gives the agent it implements better grounds for ethical status. The algorithm implemented in the human brain is the natural one to contrast with. As Danaher notes, we are still at an early stage in understanding the brain. However, our BabyX agent implements a simple model of what is known about these systems—so we can make some headway by comparing the BabyX model with the Level 2 Person model just outlined. We now turn to this task.

6.4 BabyX's model of emotions

At the core of the model of emotions implemented in BabyX is a set of stimulus–response rules that are rather similar to the 'if–then' rules of the commercial avatar. These rules model the human brain's lowest level emotional circuits, that run through the brainstem and hypothalamus, mapping perceptual or interoceptive stimuli onto physical behaviours. These circuits are evolutionarily old, and are found in all mammals; they have been most thoroughly studied in nonhuman animals. A particularly comprehensive account of them is given in [28]. The circuits are essentially concerned with *homeostasis*: keeping the agent fed, healthy, competitive, and away from danger. We implement circuits for 'approach/interest', 'joy', 'fear', 'anger', 'distress' and 'startle': a set slightly different from that posited by Panksepp, that builds in some stimuli and responses relevant for the baby avatar. In BabyX, for instance, the 'approach/interest' circuit is triggered by human speech with a certain timbre and pitch contour, and triggers a lowering of the eyelids, and a smile.

These simple emotional circuits form the basis for a more elaborate emotional network, that has no correlation in our simple commercial avatar. First, the 'behavioural responses' triggered by these circuits include signals *to the agent's body*, through the release of various neurochemicals, orchestrated in the hypothalamus. For instance, 'approach/interest' triggers the release of dopamine and oxytocin, 'joy' triggers the release of dopamine; 'fear' and 'anger' both trigger the release of norepinephrine and cortisol. Second, activity in these circuits is central in defining what counts as the agent's *goal state*. For instance, the dopamine circuit not only defines one of the agent's basic emotions, but also controls the agent's operant learning, so the agent is led to do things that lead to certain emotions.

Collectively, activity in the six basic emotional circuits activates a vector of 8 neurochemical concentrations we term the agent's *neurochemical state*. This state *modulates* many aspects of the agent's behaviours. For instance, cortisol increases BabyX's heart and breathing rates, while oxytocin decreases these rates; norepenephrine increases an acceleration parameter in BabyX's motor movements, so they are more sudden and jerky. This means that BabyX's emotional behaviours *emerge* from a complex set of brain mechanisms. In addition, the agent's emotional behaviours mutually inhibit one another, through action-selection circuits in the basal ganglia, leading to further emergent effects. This mutual inhibition, coupled with a continuously changing neurochemical state, can lead to sudden changes in overt behaviour. For instance, when a dog is frightened but becomes progressively more angry, its overt behaviour can snap suddenly, and somewhat unpredictably, from fear to anger. Babies show similar behavioural discontinuities, as any parent can confirm. We model such behavioural discontinuities using catastrophe theory (see classically [42].

The circuit described so far is predominantly subcortical. In mammals, there is a higher-level emotional circuit involving the cortex, which adds considerable additional complexity to the picture. The cortex is where cognitive representations of objects, people, events and situations are expressed. The regions expressing these representations all receive rich inputs from the subcortical emotional circuits and are also sensitive to the neurochemical state induced in the body by these circuits. Most directly, these inputs allow cognitive representations of all kinds to become *associated* with arbitrary emotional states. But they also support a richer model of emotional states themselves. Cognitive representations affect the agent's physical behaviour, and bodily neurochemical state, and so *feedback* to the agent's subcortical circuits. The result is a dynamical system, in which the current internal state of the agent can affect the next internal state, so that the agent's internal state moves on a trajectory through a space of possible states (see e.g. Scherer [32]). Within this space, there are certain *attractor points*, where the internal state is relatively stable. In our model, these attractor points correspond to higher-level emotions. In humans, these emotions are the ones that are expressible verbally.

The model of cortical emotions induces a dynamics on the agent's internal emotional state. But it also considerably enlarges the space of emotional states, adding additional cognitive dimensions to the neurochemical space defined by the subcortical system. Some of this additional complexity comes from the agent's rich repertoire of motor behaviours, which are natural attractor points. For instance, we model the human emotional state of 'sulking' as the subcortical feeling of 'anger', coupled with a tendency to withhold overt actions. Some of it comes from the agent's rich variety of cognitive representations and contexts. For instance, we model the human emotional state of 'nostalgia' as the subcortical feeling of 'joy', coupled with the cognitive operation of recalling events from episodic memory. BabyX has a simple model of episodic memory to support this. And we model the human emotional state of 'confusion' as the subcortical feeling of 'distress', coupled with low confidence about some cognitive judgement, such as what will happen next. Our model of episodic memory supports this by making a prediction about the next event, expressed as a probability distribution over possible events, conditioned on the sequence of recent events.[4]

The fact that the agent has a model of episodic memory is also significant, as foreshadowed in Sect. 2.2. We see three ways BabyX's episodic memory model could have ethical significance. First, her episodic memory system stores sequences of events or event types, and thus allows her to *anticipate* upcoming events (see e.g. the model of episodic memory we present in [38], on which BabyX's model is based). Since the agent can also learn associations between events and emotions, she has the ability to look forward to anticipated event, or to be worried or frightened by it. This ability is potential of ethical significance, mainly because it amplifies BabyX's emotional experience. Second, an important component of the human conception of 'self' is grounded in episodic memory. As argued by Tulving [39], a

[4] Readers may be interested in our approach to Paul Ekman's well-known model of 'basic' human emotions (see e.g. [15]. Ekman proposes that humans in all cultures can recognise and express six emotions: anger, fear, surprise, sadness, disgust, happiness, and contempt. There has been some debate about whether Ekman's emotions have any neural reality (see especially [16]. But recent whole-brain imaging work by Nummenmaa and Saarimaki [26] suggests there are patterns of brain activity that do correspond with these emotions. We see Ekman's emotions as attractor points, or perhaps attractor regions, in the dynamical system defined by the cortical emotional circuits sitting on top of the subcortical emotional system. We see them as sitting alongside finer-grained attractor points in this same system, such as 'sulking', 'nostalgia' and 'confusion', so in our model they don't have any special status. In this, our model follows Cowen and Keltner [10].

person's representation of him or herself comes partly from their memories of all the events they have participated in. We have not yet modelled an explicit representation of 'self' for BabyX, but we have implemented elements of Tulving's 'autobiographical self'. Again, this may be of ethical significance, because it helps to create an entity BabyX can attribute 'her' emotions 'to'. Thirdly, BabyX's episodic memory system has an important effect on her *overt behaviour*. There is good evidence that the human episodic memory system is also implicated in planning and decision-making (see [31] for a review). To accommodate this finding, the same mechanism that allows BabyX to make predictions about future events is also used when she is acting, to choose actions anticipated to have beneficial outcomes. This impact of BabyX's episodic memory system on her overt actions gives it further ethical significance, in particular because these actions are likely to affect the behaviour *of the user*, and thus to influence the 'user-related' effects we discussed in Sect. 5.

A final, distinctively human, component of our model is the agent's ability to *attend* to her internal mental states, including emotions. In our model, the agent can enter a special cognitive mode where representations of world objects are activated not by their salience in the current physical scene, but by the strength of their association with the agent's current emotional state. When an object is selected by this mechanism, the selected object serves in turn to select the emotion most strongly associated with it. This process generates stative propositions reporting the agent's emotions, such as 'I like dogs'.

Parts of this biological model of emotions have been pressed into philosophical service—most notably by the neuroscientist Antonio Damasio (see [11]. Damasio argues that human emotions derive their special character from the low-level subcortical circuits described above. He sees them as providing a constant backdrop of mental processing for the agent, that is essential for a conscious conception of the self. For Damasio, the ongoing processing in these circuits encodes a 'protoself', experiencing 'primordial feelings', that are intimately connected to the agent's body, and to its survival. (Damasio sees this 'protoself' as combining with Tulving's 'autobiographical self', to create a multimodal composite conception of self.) Damasio suggests higher-level cortical emotional circuits essentially *perceive,* and interpret, brain activity in these subcortical areas, in the same way that sensory mechanisms like vision and audition perceive and interpret stimuli in the external world. But the essential character of emotions comes from the subcortical system. Damasio makes much of the fact that damage to the subcortical emotional circuits is devastating not just to the ability to produce emotional behaviours, but to consciousness itself: it readily leads to a persistent vegetative state, one step removed from coma. He takes a punt that the neurological conception of consciousness, based on external symptoms and behaviours, is also of use in explicating the philosophical conception of consciousness, understood as a subjective experience.

To summarise: BabyX is a model of a baby, with a simulated physical body, and a system of emotional processing that is inspired by what is known about the emotional system of the mammalian brain. BabyX's model of emotions is still very simple, and likely to be wrong in many respects—scientifically, there is still much disagreement about how the human emotional system works. But nonetheless, it has some approximate claim to biological plausibility. And to the extent that it does, it may be possible to argue, from a perspective like Damasio's, that it captures something about what is distinctive about biological emotions: what links them to the notion of an experiencing self.

Crucially, BabyX also *produces emotional behaviours*, that feel plausible to a human interlocutor. She passes Danaher's 'emotional Turing test', at some level, at least as successfully as our Level 2 Person. And for BabyX, it's much harder to argue that she is achieving this success through fakery: we are doing our best to simulate the mechanisms through which human emotional behaviours arise. While BabyX and our 'Level 2 Person' have a similar *behavioural* claim to moral patienthood, we will argue that a detailed examination of the algorithms generating these agents' behaviours allow important ethical distinctions to be drawn between them. That is, we will argue against narrow ethical behaviourism, and in favour of a version of broad ethical behaviourism. We also aim to stop short of 'biological chauvinism'; we'll argue that simple computational models of the brain are helpful in this regard.

6.5 A comparison between BabyX and the Level 2 Person

What are the key differences between the algorithms run by BabyX and the Level 2 Person, that bear on the ethical status of these two types of agent? Given that we have access to the algorithms that animate both avatars, we are in an interesting position to make the comparison: in this case, we are not stymied by our ignorance of neuroscience. To us, there are five major differences for BabyX's emotions model. First, the model makes reference to *mechanisms in the agent's body*, including various sensory receptors, in its simulation of the neurochemical system. Second, the model has a much richer representation of the agent's internal state. There is the 8-dimensional neurochemical state encoding various different kinds of value—and on top of this, the many additional dimensions supplied by the agent's current cognitive state. And this internal state also has its own rich, continuous dynamics. Third, BabyX's internal state also includes various *goal* states, which interact with her

emotional states, and predispose her to various overt behaviours. Fourth, BabyX's emotional system supports learning: arbitrary stimuli can become associated with emotions, giving the sense that emotions can be *about* things. Fifth, the model includes a mechanism whereby the agent can *attend* to her own emotional state, and record facts about it, as an extension of the mechanisms through which she perceives the external world. The Level 2 Person's algorithm has none of these features. Its internal state is extremely minimal—a set of symbolic contexts supplied by a human author—and it only updates in simple, discrete ways. There is no goal state and no plasticity in interactions with the world or user. The agent has a simulated physical body, but this body does not feature in the representation of emotions. There is nothing to simulate the agent's perception or representation of her own emotions.

We don't want to overstate the difference we perceive between BabyX and a Level 2 Person. BabyX's brain model, while much more complex than that of a Level 2 Person, is still extremely simple: it only supports very rudimentary representations of objects, events, states, and situations. At present, BabyX's brain is still *far* less complex than a mammalian brain: even the brain of a simple mammal (like a rat or mouse) is far more complex, and supports far greater autonomy for its owner. But we are working to develop the brain model, and make it more complex and realistic—and we intend to incorporate further insights about the brain's emotional system as these emerge in neuroscience. If we are correct in attributing special importance to the brain's subcortical emotional system, which is evolutionarily old, and relatively similar across mammalian species, we might at some point in the future want to place BabyX somewhere on the scale of biological moral patients, giving her moral parity with some lower animal—perhaps initially a rat or mouse. We emphasise we do not think we are at this point yet. But given our intention is to continue developing the BabyX model in the direction of biological plausibility, we think this is likely to place BabyX on the scale of moral patients at *some* point in the future.

Note that our comparison of the BabyX and Level 2 Person algorithms essentially buys into a functionalist model of emotional states of the kind advocated by Putnam. In fact, BabyX's algorithm provides many features of the 'functional organisation' Putnam saw as prerequisite for defining ethically relevant mental states like 'pain', such as mechanisms for 'assigning value' to sensed stimuli, and mechanisms for producing behaviours based on these valuations. The main difference is that our analysis is *more detailed*, because it refers to an implemented algorithm, and *more informed*, because the algorithm implements a model grounded in findings from neuroscience. At the same time, note that BabyX's algorithm is still a vast simplification of processes in actual brains—and even then, our analysis only picks up on some features of the algorithm. Thus the important functional features are stated generally enough that we can readily envisage a large space of *nonbiological* agents that would also qualify for ethical consideration. We are certainly using the human brain as a yardstick for ethical status here. But we are not being 'biological chauvinists', because our method is to look for *features* of our simplified brain model that are relevant in according ethical status, that could be found in all manner of nonbiological agents.

In summary, we think it's important to refer to the algorithms that animate simulated agents when assessing their ethical status, rather than just considering their behaviour. Reference to implemented algorithms helps to supply relevant detail about what's ethically important about biological agents. At the same time, reference to an algorithm helps us to define a class of ethical patients that extends beyond the biological realm.

6.6 Agent-related ethical issues: some practical conclusions

Is BabyX a moral patient, or is she on the way to becoming one? Taking a practical stance on this question involves assessing competing risks. There are risks in failing to recognise BabyX as a moral patient if she is one. But there are also risks in treating her as a moral patient if she is not one. What does it mean to err on the side of caution here? For now, we think we can safely assume BabyX is not yet a moral patient. But in this section, we have argued that we do need to pay some heed to the possibility that BabyX will take on some status as a moral patient at some point in the future. In view of this, we think it's prudent to set up an ethics committee in our company to periodically assess BabyX's claim to ethical status, and if it is assessed to have such a claim, to introduce new rules governing the use of BabyX. The committee will also consider the ethics of any user experiments proposed for BabyX, in advance of the experiment being run. In this sense, it will be like a regular experimental ethics committee. And it will consider any proposed applications of BabyX in commercial products.

7 Summary and conclusions

This paper contributes to an ongoing discussion about the moral status of simulated human-AI agents. Our discussion centres on interactions with our BabyX AI system, which throws certain ethical questions into particularly sharp relief. Following Danaher, we separately discuss possible ethical effects on a human user interacting with BabyX, and possible ethical effects 'on BabyX' (if there are such

things). In both cases, our focus is on interactions involving mistreatment. The two discussions are very different. The first is about possible social impacts of simulated humans as an emerging technology. The existing debate is mostly about the effects of violent video games on the users of these games—an empirical question which is still far from resolved. BabyX contributes to this debate mainly by providing an extreme example on various dimensions. BabyX has a particularly high degree of graphical and behavioural realism, and of simulated proximity, and babies are the most vulnerable members of human society. All of these factors advocate for caution in the deployment of BabyX. The second discussion is about the philosophy of mind. We focussed on Danaher's recent practical proposal that the ethical status of a simulated human should be decided on the basis of its behaviours because other epistemic options are not available—in particular, we do not understand how the brain produces emotional behaviours in people. BabyX contributes to a discussion of this proposal because we know the algorithm that animates her—and this algorithm encapsulates at least some of what we know about the brain circuits responsible for emotional behaviours. We contrast BabyX's brainlike algorithm with the very different algorithm animating our Level 2 Person product: while both systems produce comparably convincing emotional behaviours, we argue the difference in algorithms is ethically significant, and that BabyX's algorithm places her closer to having some (small) amount of ethical patiency than a Level 2 Person. On these grounds too, we see grounds for caution in the development and deployment of BabyX.

In this paper, we hope to have shown that AI models of simulated humans are becoming increasingly relevant to ethical discussions. They are relevant both to practical discussions about impacts of new technologies and to more theoretical discussions of what defines an ethical patient. We believe the researchers developing these new models can contribute usefully to these discussions. In particular, if the ethical patiency of an AI agent hinges on the algorithm it implements, those who developed the algorithm have relevant information to contribute.

In fact, if algorithms matter, their developers are not just useful informants to an ethical debate. They also have certain new ethical *responsibilities,* especially if the agents they create use sophisticated models of emotions. In particular, if a given algorithm will accord a simulated agent some degree of ethical status, the developers need to think very carefully about the conditions under which such an agent would be deployed—and possibly, whether such an agent should be created at all (see again [7] for further discussion). In this paper, we have rehearsed some of the ethical deliberations we foresee AI developers will have to engage in a good deal more as their technical ability to simulate real people improves.

Acknowledgements Many thanks to John Zerilli, James Maclaurin, Colin Gavaghan, Elaine Reese, Josephine Jefferson, Holger Regenbrecht, and the two anonymous reviewers for comments on earlier drafts of this article.

References

1. Adamson, L., Frick, J.: Research with the face-to-face still-face paradigm: a review. Infancy **4**, 451–473 (2003)
2. Aicardi, C., Akintoye, S., Fothergill, B.T., Guerrero, M., Klinker, G., Knight, W., Klüver, L., Morel, Y., Morin, F.O., Stahl, B.C., Ulnicane, I.: Ethical and social aspects of neurorobotics. Sci. Eng. Ethics **26**, 2533–2546 (2020)
3. Anderson, C., et al.: Violent video game effects on aggression, empathy, and prosocial behavior in Eastern and Western countries. Psychol. Bull. **136**, 151–173 (2010)
4. Baillargeon, R.: Infants' understanding of the physical world. In: Sabourin, M., Craik, F., Robert, M. (eds.) Current Directions in Psychological Science, pp. 503–529. Psychology Press, London (1998)
5. Ballard, D., Hayhoe, M., Pook, P., Rao, R.: Deictic codes for the embodiment of cognition. Behav. Brain Sci. **20**(4), 723–767 (1997)
6. Barlett, C., Rodeheffer, C.: Effects of realism on extended violent and nonviolent video game play on aggressive thoughts, feelings, and physiological arousal. Aggress. Behav. **35**, 213–224 (2009)
7. Bryson, J.: Patiency is not a virtue: the design of intelligent systems and systems of ethics. Ethics Inf. Technol. **20**(1), 15–26 (2018)
8. Chevalier-Boisvert, M., Bahdanau, D., Lahlou, S., Willems, L., Saharia, C., Nguyen, T.-H., Bengio, Y.: BabyAI: a platform to study the sample efficiency of grounded language learning. arXiv:1810.08272 (2019)
9. Clark, A.: Being there: putting brain, body and world together again. MIT Press, Cambridge (1997)
10. Cowen, A., Keltner, D.: Self-report captures 27 distinct categories of emotion bridged by continuous gradients. PNAS, pp. E7900–E7909. https://www.pnas.org/cgi/doi/10.1073/pnas.1702247114 (2017)
11. Damasio, A.: Self comes to mind: constructing the conscious brain. Vintage, London (2010)
12. Danaher, J.: Robotic rape and robotic child sexual abuse: should they be criminalised? Crim. Law Philos. **11**, 71–95 (2017)
13. Danaher, J.: Welcoming robots into the moral circle: a defence of ethical behaviourism. Sci. Eng. Ethics **26**, 2023–2049 (2020)
14. Drummond, A., Sauer, J., Ferguson, C.: Do longitudinal studies support long-term relationships between aggressive game play and youth aggressive behaviour? A meta-analytic examination. R. Soc. Open Sci. **7**, 200373 (2020)

15. Ekman, P.: Basic emotions. In: Dalgleish, M., Power, M. (eds.) Handbook of Cognition and Emotion, pp. 285–336. Wiley (1999)
16. Feldman Barrett, L.: Are emotions natural kinds? Perspect. Psychol. Sci. **1**(1), 28–58 (2006)
17. Ferguson, C.: Do angry birds make for angry children? A meta-analysis of video game influences on children's and adolescents' aggression, mental health, prosocial behavior, and academic performance. Perspect. Psychol. Sci. **10**(5), 646–666 (2015)
18. Greitemeyer, T.: The contagious impact of playing violent video games on aggression: Longitudinal evidence. Aggress. Behav. **45**, 635–642 (2019)
19. Greitemeyer, T., Mügge, D.: Video games do affect social outcomes: a meta-analytic review of the effects of violent and prosocial video game play. Pers. Soc. Psychol. Bull. **40**(5), 578–589 (2014)
20. Hildt, E., Laas, K., Sziron, M.: Shaping ethical futures in brain-based and artificial intelligence research. Sci. Eng. Ethics **26**, 2371–2379 (2020)
21. Keijsers, M., Bartneck, C.: Mindless robots get bullied. In: Proceedings of HRI'18, March 5th–8, 2018, Chicago, IL, USA (2018)
22. Kim, K.-J., Sundar, S.: Can interface features affect aggression resulting from violent video game play? An examination of realistic controller and large screen size. Cyberpsychol. Behav. Soc. Netw. **16**(5), 329–334 (2013)
23. Lewis, M., Ramsay, D.: Development of self-recognition, personal pronoun use, and pretend play during the 2nd year. Child Dev. **75**(6), 1821–1831 (2004)
24. Luck, M.: The Gamer's dilemma. Ethics Inf. Technol. **11**(1), 31–36 (2009)
25. Normand, B.: Teaching touching safety rules: safe and unsafe touching-activity. Committee for Children Blog. https://www.cfchildren.org/blog/2017/08/activity-teaching-touching-safety-rules-safe-and-unsafe-touching/ (2017). Accessed 13 Aug 2020
26. Nummenmaa, L., Saarimäki, H.: Emotions as discrete patterns of systemic activity. Neurosci. Lett. **693**, 3–8 (2019)
27. Oudeyer, P.-Y.: What do we learn about development from baby robots? WIREs Cogn. Sci. **8**, e1395 (2017)
28. Panksepp, J.: Affective neuroscience: the foundations of human and animal emotions. Oxford University Press, New York (1998)
29. Persky, S., Blascovich, J.: Immersive virtual environments versus traditional platforms: effects of violent and nonviolent video game play. Media Psychol. **10**(1), 135–156 (2007)
30. Putnam, H.: The nature of mental states. In: Putnam, H. (ed.) Mind, Language, and Reality, pp. 429–440. Cambridge University Press, Cambridge (1967)
31. Schacter, D., Benoit, R., De Brigard, F., Szpunar, K., Addis, D., Buckner, R.: Episodic future thinking and episodic counterfactual thinking: intersections between memory and decisions. Neurobiol. Learn. Mem. **117**, 14–21 (2015)
32. Scherer, K.: Emotions are emergent processes: they require a dynamic computational architecture. Philos. Trans. R. Soc. B **364**, 3459–3474 (2009)
33. Smids, J.: Danaher's ethical behaviourism: an adequate guide to assessing the moral status of a robot? Sci. Eng. Ethics **26**, 2849–2866 (2020)
34. Soul Machines: Soul Machines Ethics Policy. https://www.soulmachines.com/wp-content/uploads/Ethics_Policy_1.0-1-1.pdf (2021)
35. Sparrow, R.: The Turing triage test. Ethics Inf. Technol. **2004**(6), 203–213 (2004)
36. Sparrow, R.: Can machines be people? Reflections on the turing triage test. In: Lin, P., Abney, K., Bekey, G. (eds.) Robot Ethics: The Ethical and Social Implications of Robotics, pp. 301–316. MIT Press, Cambridge (2012)
37. Sparrow, R.: Robots, rape, and representation. Int. J. Soc. Robot. https://doi.org/10.1007/s12369-017-0413-z (2017)
38. Takac, M., Knott, A.: Mechanisms for storing and accessing event representations in episodic memory, and their expression in language: a neural network model. In: Proceedings of the 38th annual meeting of the cognitive science society (CogSci) 2016, pp. 532–537 (2016)
39. Tulving, E.: Episodic memory: from mind to brain. Annu. Rev. Psychol. **53**, 1–25 (2002)
40. Tomasello, M.: Constructing a language: a usage-based theory of language acquisition. Harvard University Press, Cambridge (2003)
41. Woodward, A.: Infants selectively encode the goal object of an actor's reach. Cognition **69**(1), 1–34 (1998)
42. Zeeman, E.: Catastrophe theory. Sci. Am. **234**(4), 65–83 (1976)
43. Zendle, D., Kudenko, D., Cairns, P.: Behavioural realism and the activation of aggressive concepts in violent video games. Entertain. Comput. **24**, 21–29 (2018)
44. Zilio, D.: Who, what, and when: skinner's critiques of neuroscience and his main targets. Behav. Anal. **39**, 197–218 (2016)