Ethics Inf Technol
DOI 10.1007/s10676-015-9371-z

ORIGINAL PAPER

Social robots, fiction, and sentimentality

Raffaele Rodogno[1]

Abstract I examine the nature of human-robot pet relations that appear to involve genuine affective responses on behalf of humans towards entities, such as robot pets, that, on the face of it, do not seem to be deserving of these responses. Such relations have often been thought to involve a certain degree of sentimentality, the morality of which has in turn been the object of critical attention (Sparrow in Ethics Inf Technol 78:346–359, 2002; Blackford in Ethics Inf Technol 14:41–51, 2012). In this paper, I dispel the claim that sentimentality is involved in this type of relations. My challenge draws on literature in the philosophy of art and in cognitive science that attempts to solve the so called paradox of fictional emotions, i.e., the seemingly paradoxical way in which we respond emotionally to fictional or imaginary characters and events. If sentimentality were not at issue, neither would its immorality. For the sake of argument, however, I assume in the remaining part of the paper that sentimentality is indeed at play and bring to the fore aspects of its badness or viciousness that have not yet been discussed in connection with robot pets. I conclude that not even these aspects of sentimentality are at issue here. Yet, I argue that there are other reasons to be worried about the wide-spread use of ersatz companionship technology that have to do with the potential loss of valuable, self-defining forms of life.

Keywords Social robots · Sentimentality · Paradox of fictional emotions · Ersatz companionship · Technology · Affective engagement

✉ Raffaele Rodogno
filrr@cas.au.dk

[1] Department of Philosophy and History of Ideas, Aarhus University, Jens Chr. Skous Vej 7, 8000 Århus C, Denmark

Introduction

The starting point of this paper is the relation between elderly citizens including those suffering from various forms of dementia, on the one hand, and robot pets such as the famous robotic seal called Paro, on the other. In particular, our focus will be on this human-robot interaction as it involves what seem to be genuine affective responses on behalf of elderly citizens to an entity that, on the face of it, does not seem to be deserving of these responses.

It may be argued that such relations are beneficial to the elderly and, hence, that there is reason for them to engage in such relations and perhaps reasons for others to promote or at least not hinder such relations. There is initial evidence to the effect that, for example, engaging with 'social' robots—systems that can be perceived as social entities that communicate with the user–increases positive mood, diminishes loneliness, alleviates stress, increases immune system response and even decreases existing dementia (Broekens et al. 2009: p. 98). Unfortunately, however, the value of this evidence is in many instances tarnished by the methodological weaknesses of the studies that evinced it (Broekens et al. 2009: p. 98–100). There are, however, at least some studies whose results are convincing (Kidd et al. 2006; Wada K, Shibata T. 2007). In light of the evidence they provide, we should at least consider the claim that the elderly have prudential reason to engage in these relations while others have moral reasons to promote or, at least, not hinder such relations.

As Sparrow (2002: 306) has correctly noted, while the moral issues raised by our relations to robot pets may not seem the most urgent, they do pose a problem that is structurally analogous to a potentially much more serious phenomenon of large scale 'ersatz companionship'. The advent of robot pets as potentially worthwhile companions

Published online: 06 August 2015

for the lonely aged is in other words a valuable opportunity for thinking about the larger scale advent of other companionship robot technologies, perhaps in the shape of human beings ('androids'). What should be the worth of such companions and of our relationships with them? How would their widespread advent affect our current values and modes of life? As we will see, in the end, a discussion of the sentimentality of relations with robot pets leads us precisely into such larger questions.

Coming back to robot pets, despite the potential for benefit, Sparrow himself finds moral reasons to oppose their use. According to him, the relation that we would have with robot pets is predicated on

> … mistaking, at a conscious or unconscious level, the robot for a real animal. For an individual to benefit significantly from ownership of a robot pet they must systematically delude themselves regarding the real nature of their relation with the animal. It requires sentimentality of a morally deplorable sort. Indulging in such sentimentality violates a (weak) duty that we have to ourselves to apprehend the world accurately. The design and manufacture of these robots is unethical in so far as it presupposes or encourages this. (Sparrow 2002: p. 306)

From this passage we extract a number of claims:

1. In order for individuals to derive significant benefit from interaction with a social robot, they ought to be deluded as to the real nature of the robot.
2. This delusion involves sentimentality.
3. Sentimentality is not always morally deplorable.
4. The sentimentality at issue here, however, is morally deplorable.
5. What makes sentimentality morally deplorable (at least here) is the fact that it involves the violation of a weak duty to ourselves to apprehend the world accurately.
6. Designing and producing these robots is immoral but its immorality derives from the violation of the duty in 5.

Sparrow's challenge has already received some critical attention (Blackford 2012). Much of the available discussion, however, revolves around points (4) and (5), i.e., whether the sentimentality at issue here is morally deplorable as opposed to a relatively benign kind of self-indulgence in forming beliefs about the world. In this article, however, I would like at least initially to shift the focus away from these moral issues, and to place it on the conceptual and empirical issues at play in points (1) and (2). Are the elderly being sentimental (and hence deluded) when engaging in genuine affective relations with robot pets? And do they have to be sentimental in order to benefit from interacting with robot pets?

Let us start by defining sentimentality. According to Midgley (1979: p. 385):

> Being sentimental is misrepresenting the world in order to indulge our feelings.

Hence, for example (Midgley 1979: 386):

> If somebody claimed that his car was frightened when he was in danger, that it was deeply devoted to him, jealous of his friends, and grateful for his attention to it, we should probably all agree that he was being sentimental.

The concept of sentimentality, then, involves a double claim. When being sentimental the subject (1) is misrepresenting the world and (2) is doing so with the aim of indulging some emotion or mental state. The misrepresentation involved in sentimentality, then, is not a simple form of misrepresentation to which we may inadvertently fall prey. It is rather something that we voluntarily entertain or at least fail to resist, in view of the desired feelings or mental states that come with misrepresenting.

The questions we should ask, then, are whether those who engage affectively with robot pets misrepresent the world, and whether they do so with the aim of indulging some feeling. Finally, we should also consider whether it is true that one must be deluded about the real nature of robot pets in order to derive significant benefit from interacting with them. These questions have not to this point received any critical attention. Yet the charges of immorality expressed by (5) and (6) rest precisely on the assumption that those who engage with robot pets are being sentimental. Showing that this assumption is false will disengage the moral objection that rests on it. While not ultimately offering a categorical dismissal of this assumption, in the next section, I argue that those who hold affective relations to robot pets are not necessarily and, in fact, not likely misrepresenting the world. In the same vein, there is no reason to believe that they have to misrepresent the world in order to derive significant benefit from their relation with the robot pet.

In the next section, then, I bracket discussion of moral issues while questioning whether sentimentality is at all involved. As we move to the section after that, however, I do the opposite. I grant the assumption of sentimentality while questioning its wrongness. I argue that what is ultimately wrong is not anything inherent to individual cases of sentimentality but, perhaps, the way in which large-scale ersatz companionship potentially endangers other values we hold dear.

Is senior being sentimental?

In order to have genuine emotional responses to a robot pet one must believe that it is not a mere simulation of a pet, but that it actually instantiates some properties that to our knowledge only real pets possess. When I lavish care and attention to my dog, say by caressing him, feeding him, and being attentive to his needs, I must believe that he genuinely has these needs and that he has an evaluative point of view from which it makes a difference whether I strike him with a stick or gently scratch behind his ears. I take the fact that he displays appreciation towards my attentions not as a simulation but as a genuine expression of appreciation or of a pro-attitude so intimately connected with this evaluative point of view.[1] Similarly my joy at his display of affection towards me must be anchored in the belief that the display is genuine and not simulated. If I were to believe that his "expressions" were mere simulations, I would stop feeling joy at the sight of the dog and if I continued to feel such joy, I would certainly consider it unwarranted.

Now, however, suppose that a growing number of individuals will not be put off by these simulative behaviours and will continue to lavish care and attention even when believing that what they are engaging with is not an actual dog with the relevant properties but a robot pet that can at most simulate having these properties.[2] Are these individuals' emotions unwarranted, irrational or inappropriate? Are these people misrepresenting the world when feeling such emotions? Remember that the latter question must be answered positively in order for the sentimentality charge to stick at all.

These questions, or, at any rate, questions very similar to these, have philosophical precedents. Rather similar issues have been the object of animated debate in the philosophy of art under the heading of the paradox of fiction (or emotional fiction). I therefore propose to tackle our questions by way of an examination of this paradox, its attempted solutions, and its similarities and dissimilarities to our original case.

In the article that was to start what became the debate on the paradox of fiction, Radford (1975: p. 68) considers the following example:

> Suppose that you have a drink with a man who proceeds to tell you a harrowing story about his sister and you are harrowed. After enjoying your reaction he then tells you that he doesn't have a sister, that he has invented the story… once you have been told this you can no longer feel harrowed. … But the possibility of your being harrowed again seems to require that you believe that someone suffered.

Radford is trying to illustrate a common reaction: when we realize that the situations to which our emotions are directed do not involve something that really happened to real individuals (i.e. concrete objects in the actual world), the emotions typically disappear and we think that they should disappear. According to Radford that is because being saddened by the circumstances affecting a person involves the belief that this person exists and suffers. In the absence of such belief, our sadness disappears and, if it didn't, we should at any rate consider it unwarranted or unfitting.

Now consider another common reaction, namely, the fact that individuals regularly exhibit apparently genuine emotional responses to fictional characters and situations that the individuals represent as being merely imaginary. When immersed in the novel *Anna Karenina*, for example, consider your sadness at reading that Anna suffers a major setback. This reaction is in an apparent tension with the kind of reaction you have in Radford's case. While upon learning that your friend's sister did not actually exist you probably cease to feel sad about her, in the fictional case, there is no accompanying or underlying belief that Anna Karenina exists as a concrete individual in the actual world, and yet you feel sad about her and perhaps think that this is just as it should be.

This tension has been presented as a paradox, the paradox of fiction (or of fictional emotions), which may schematically be put as arising when one endorses these three claims:

(a) *Response Condition*: S experiences genuine emotional responses towards *F*.
(b) *Belief Condition*: S believes that *F* is fictional.
(c) *Coordination Condition*: In order to have genuine emotional responses towards someone *F*, one must not believe that the character is fictional.[3]

[1] In the not so distant past many would have considered these claims as betraying sentimentality towards animals. See Midgley (1979).

[2] I assume that current robotic pets merely simulate emotional and intentional states—they display the salient criterial behaviour but the internal processes that produce the relevant body movements and sounds do not bear any functional similarity to those processes that produce the criterial behavior within an animal. I do not need to consider here the further question whether at some point in the future robot pets could do more than simulate behaviour, i.e., produce the relevant behaviour on the basis of an internal process architecture that is functionally equivalent to an animal's architecture. For discussion of this point see Seibt (2014). Note that in Seibt's (2014) classification, the robot pets at hand here merely "approximate" affection rather than imitate it.

[3] This is Gendler's (2013) way of putting the paradox. I am also indebted to Schneider (2006) and Neill (2005) for the following discussion of the paradox of fiction.

Now the analogy with the robot pet case is quite evident and can schematically be put as follows:

(d) *Response Condition*: S experiences genuine emotional responses towards *P*.
(e) *Belief Condition*: S believes that *P* is a robot pet that only simulates real pet properties.
(f) *Coordination Condition*: In order to have genuine emotional responses towards a robot pet, one must not believe that it merely simulates real pet properties.

In short, both sets of claims seem to involve an inconsistency at the level of the beliefs that are allegedly involved when we engage emotionally with, respectively, a fictional character or event, and a robot pet.

Philosophers attempting to solve the paradox of fiction have employed different strategies.[4] At the most general level, these can be understood as challenging in turn one of the three claims (a), (b), or (c), i.e., respectively the *Response Condition*, the *Belief Condition*, or the *Coordination Condition*. Walton (1978), for example, has famously argued against (a). We do not experience genuine emotions in the case of fiction but something else, namely, *quasi-emotions*, which, despite being both physically and psychologically very similar to genuine emotions differ from the latter and do so

> primarily in that they are generated *not* by existence beliefs (such as the belief that the monster I am watching on screen really exists), but by "second-order" beliefs about what is *fictionally* the case according to the work in question (such as the belief that the monster I am watching on screen *make-believedly* exists). … Thus, it is make-believedly the case that we respond emotionally to fictional characters and events due to the fact that our beliefs concerning the fictional properties of those characters and events generates in us the appropriate *quasi*-emotional states. (Schneider 2006)

Beside the difference in their cognitive basis, quasi-emotions differ from emotions with respect to their behavioural consequences: we tend not to act on fictional emotions in the same way as we act on our genuine emotions. Even if we are saddened by Anna Karenina's misfortunes, we do not attempt to console her.

What insights can we gain by applying this solution to the paradox of fiction to cases of alleged emotional engagement with robot pets? In short, the resulting claim would be that these would not count as genuine cases of emotional engagement but as at best instances of quasi-emotions. That would have to be explained by the second-order kind of beliefs according to which the robot pet one interacts with only *make-believedly* possesses the relevant properties. Yet, unlike the case of fiction, it would seem that the behavioural reactions linked to quasi-emotions would be quite similar to their counterpart genuine emotions. If the robot pet simulates sadness the individual who engages with it may well respond by attempting to console it.

Two quick comments are in order here, the first specific to the pet robot case, and the second directed to the idea of quasi-emotions more generally. Firstly, is it as a matter of fact the case that those who engage (quasi-)emotionally with pet robots display this type of second-order beliefs about their properties? This is not meant as a rhetorical question but as a genuine empirical question to which to this point we do not have an answer. Note, however, how a positive answer to this question implies that when we engage with robot pets we do not misrepresent the world at the level of second-order beliefs. If one could say that we correctly represent the properties of robots as existing relative to a make-believe scenario, the sentimentality charge would have to be dropped.

Secondly, as many have noted, this type of solution to the paradox of fiction seems to rest on a mistake. Walton's idea of a quasi-emotion simply assumes that the cognitive basis of genuine emotion necessarily involves existence beliefs about its object. There is, however, no evidence to that effect and in fact evidence to believe the contrary. At the level of anecdote, consider how we may all be able to generate, say, genuine disgust simply by *imagining* something really disgusting. There is no reason to think that the disgust at issue here is quasi-disgust. We shall pick up this line of thought again later, when discussing other attempted solutions to the paradox. It seems both artificial and unnecessary to postulate a new category, i.e., quasi-emotion, instead of accepting the much less demanding and phenomenologically more accurate option according to which beliefs, and in fact, existence beliefs, are just one kind of cognitive basis that emotions are susceptible to having.

Charlton (1984) provides another attempt to challenge (a), i.e., the claim that we experience genuine emotions towards fictional objects. Unlike Walton, instead of questioning the genuine nature of the emotions, Charlton questions the claim that the particular objects of these emotions are fictional. He argues that our emotional responses are directed not towards the characters or events in the fiction but rather towards appropriate real-world *surrogates* for or *counterparts* of those characters and events. So, for example, we don't feel sadness or pity for

[4] In what follows I will only mention strategies directed at solving rather than dissolving the paradox. See Tullmann and Buckwalter (2014) for a dissolving strategy. See Cova and Teroni (2015) for a convincing reply.

Anna Karenina, but rather for someone in the actual world who has led a relevantly similar life. According to Paskins (1977): 346) when we feel or realistically drawn characters like Anna

> our pity is or can without forcing be construed as, pity for those people, if any, who are in the same bind as the character.

The immediate objection raised against this view is that, in the case of fiction, we are supposing that our pity is directed at Anna Karenina and not at some potential actual person in the same bind. I will presently return to what this particular solution to the paradox of fiction has to offer to the case of affective engagement with robot pets.

First, however, let us consider a second type of solution to the paradox, one that challenges (b) i.e., the claim that subjects actually believe that the objects of their emotions are fictional. The idea is that those immersed in fiction experience a momentary lapse in their belief that the fictional characters are imaginary. They are somewhat confused, under the illusion that they are real, or in a state of 'suspension of disbelief'.[5] Contemporary philosophers typically dismiss this view. They note that when immersed in fiction we may well suspend some of our beliefs as to what is probable or even possible as, for example, when we suspend a general disbelief in the supernatural. Yet we do not suspend our belief that we are reading a novel or watching a film.

The last two solutions to the paradox offer interesting interpretations of what may actually happen to at least some of the individuals when emotionally engaged by robot pets. It may be as a matter of empirical fact the case that some of the subjects who relate to robot pets enter states of 'suspension of disbelief' or engage emotionally with real pet counterparts or surrogates. This may be less unlikely for those subjects who suffer from Alzheimer, are senile, or otherwise relevantly cognitively impaired. Note, however, that while these subjects would misrepresent the world, it seems unlikely that they would do so intentionally, in order to indulge some emotion. I would take the subjects at hand to be the objects of a case of misrepresentation *simpliciter* rather than of misrepresentation entertained and motivated by some further aim. As for relevantly healthy subjects, the attribution of misrepresentation *simpliciter* seems admittedly to offer an unlikely empirical interpretation of what goes on in the case of robot pets. This will become clearer as we study the third and most popular kind of solution to the paradox.

On this approach claim (c) comes under critical fire. We have already mentioned one strand of this strategy, sometimes referred to as the *Thought Theory*. On this view, although our emotional responses to *actual* characters and events may require beliefs in their existence, there is no good reason to hold up this particular *type* of emotional response as the model for understanding emotional response *in general*. We could actually postulate that the cognitive basis of emotions in fiction, rather than beliefs in the actual existence of fictional events and characters, consists in "mental representations" (Lamarque 1981), "entertaining in thought" (Carroll 1990), "imaginatively proposing" (Smith 1995) to ourselves or "*aliefs*" (Gendler 2008a, b), i.e., cognitive mechanisms that are indifferent between content that is represented as real or as merely imaginary. Let us refer to the cognitive modes described by these scholars as "non-serious" modes. A relevant way to capture the difference between beliefs, on the one hand, and non-serious cognitive modes, on the other, is the way in which they are regulated by the truth norm: a belief that *p* should be abandoned in the light of information that not-*p*. The same norm, however, does not apply to non-serious cognitive modes.

Another type of solution to the paradox that challenges (c) is proposed by Radford (1975) himself and goes under the heading of *irrationalism*. On this rather defeatist solution, we simply are irrational insofar as we do feel emotions about fictional characters that we should not rationally feel. In other words, *contra* (c), we do have genuine emotional responses about fictional characters because we are irrational about this.

The irrationalist solution connects this discussion with the question of the rationality or appropriateness of emotions (De Sousa 1987, 2002, 2007; Samela 2006; Teroni 2007; Prinz 2004). When assessing the rationality or appropriateness of emotions, we take account of at least two dimensions. Firstly, we take it that emotions have *evaluative* correctness conditions. On this view, each emotion type is partly to be understood in terms of some type of value or disvalue. Hence, for example, the relevant value for sadness is loss, for anger offence, for fear danger, etc. On this dimension, then, my sadness is appropriate or rational to the extent to which the event towards which it is directed amounts to a loss and similarly with anger and offence, fear and danger, etc. The second dimension has to do with what we called above the *cognitive* bases of emotions. To be sad about an event you need to represent the event through direct sensory perceptions, beliefs formed as a result of inferences, or perhaps through some other channel.[6]

[5] Coleridge (1907) is often mentioned (Charlton 1984; Gendler 2013) as possibly maintaining this view.

[6] Cova and Teroni (2015, my translation) combine these two dimensions in the following way: "…emotions inherit the correctness conditions of these cognitive bases and add an evaluative layer to them. In other words, if you are, say, visually aware of an event, then that experience is correct if and only if the event exemplifies the

Now *irrationalism* is predicated on the assumption that truth norm-regulated beliefs are the only kind of cognitive basis for emotions. If this assumption were correct, our emotions towards fictional characters and events would indeed be irrational or inappropriate insofar as these emotions would be based on a false belief, a belief that should rationally be abandoned in light of information available to us that denies their content. Theories of emotions, however, suggest that the *irrationalist's* fundamental assumption is mistaken: emotions may have non-serious cognitive bases that are not regulated by a truth-norm. Hence we can say that while my sadness for Anna Karenina, grounded on such non-serious cognitive bases, should rationally be *insensitive* to information about her existence in the actual world, my sadness towards the sister in Radford's example, grounded on a belief, should rationally be sensitive to such information.[7]

Let us now apply these points to the case of affective engagement with robot pets. The question here is not about whether emotions towards robot pets involve beliefs as opposed to non-serious cognitive bases *regarding the existence of the robot pet* that is the object of one's emotions. We are here assuming that the elderly, or whoever engages with robot pets, would have beliefs about the existence of the particular robot pet in question.[8] What is at issue here, however, are the properties relevant to the genuine emotional engagement that real pets instantiate and robot pets merely simulate. The question, then, is one about the type of cognitive mode at issue *with regard to these properties*: are our emotions towards a robot pet based on a belief that it instantiates the relevant real pet properties, or are our emotions based on non-serious cognitive modes that are insensitive to information about the instantiation of such properties?

If the former were the case, we would have to accept that our emotional engagement to robot pets involves a form of irrationality, for it would involve the belief that the pets instantiate properties, which we know they do not instantiate. I think this would be the closest we have so far reached to a case of misrepresentation of the world. At this point two remarks are in order. Firstly, we should once again ask whether this scenario correctly describes at least some cases of genuine affective engagement with robot pets, and, if so, how statistically significant this type of misrepresentation tends to be. Secondly, while noting that this scenario depicts a form of irrationalism similar to the one described by Radford, we should point out that there is a significant difference between the two. While Radford describes individuals as irrational insofar as they are unable to resist a blameless cognitive fault, in the scenario proposed here individuals are described as misrepresenting the world with the aim of basking in some feeling or other. Both descriptions involve misrepresentation of reality but the latter involves something over and above that. Attribution of sentimentality seems to involve a form of voluntariness in misrepresenting reality that the misrepresenting involved in irrationalism (*à la* Radford) does not necessarily involve. In conclusion, even if, as a matter of fact, some of the individuals that engage affectively with robot pets misrepresent reality, they may well do so without sentimentality.

If our emotions were based on non-serious cognitive modes, however, this form of irrationality would not be at issue and no misrepresentation of the world would be involved. Being emotionally engaged by robot pets would be akin to being emotionally engaged by a good novel or movie. The interactions with robot pets that lead to emotional engagement would be at least in part functionally equivalent to reading a book or watching a movie. Just as my sadness for Anna Karenina involves my imagining, accepting, mentally representing or entertaining the thought (without believing) that certain unfortunate events have occurred to her, my joy at the robot pet involves my imagining, accepting, mentally representing or entertaining the thought (without believing) that it is happy to see me. No need to misrepresent the world and hence no sentimentality. Whether this option is descriptively accurate, and whether it is more so than the irrationalist option is once again an open empirical question.

So far, then, I have argued that there is a relevant parallel between the paradox of fiction, on the one hand, and the case of affective engagement with robot pets, on the other, in that both cases seem initially to involve a case of misrepresentation. When applying the various solutions to the paradox to our case, however, we come to the following conclusion: either there is no reason to suppose a case of misrepresentation or, if there is misrepresentation, it is not necessarily due to sentimentality but can be due to the irrationality or cognitive deficits of blameless subjects.

Though we cannot categorically exclude that some subjects will actively misrepresent with that aim of basking in some feeling, I have provided a series of alternative ways of understanding what may be going on when we are affectively involved with a robot pet. In fact, it seems that

Footnote 6 continued
properties that you seem to be visually aware of. And if you react with sadness to what you see, your sadness is correct if and only if the event you see constitutes a loss." Despite the presence of these two dimensions, the following discussion will bear solely on the cognitive basis. As we shall see, settling rationality questions at this level will be sufficient to our purposes.

[7] See Cova & Teroni (2015, Sect. 5) for this as well as another solution to the paradox of fiction.

[8] To deny this much would amount to asserting that one suspends belief that this is just a robot pet or perhaps that one is in fact emotionally engaged to a real pet counterpart or surrogate of the robot pet. We would then fall back on something akin to the second kind of strategy discussed above.

in the case of healthy (i.e., non-cognitively impaired) subjects, it would be both interpretatively more charitable and empirically more likely to conceive of them as experiencing emotions based on non-serious cognitive modes as opposed to understanding them as involved in a systematic misrepresentation of the world. As for those who are cognitively impaired, perhaps some degree of misrepresentation is more likely to be going on, though it is not easy to imagine that it would be positively motivated by the desire to bask in some feeling.

If true, this protracted argument would take much of the steam out of the charge of sentimentality. Returning to the schematic reproduction of Sparrow's argument above, point (2) would be dismissed as would the subsequent part of the argument that hinges on that point. But what of the very first claim in that argument? Would it be true that in order for individuals to derive significant benefit from interaction with a social robot, they ought to be deluded as to the real nature of the robot? The discussion above provides reason to doubt also this claim. Just as in the case of fiction we may enjoy affective engagement with fictional characters based on non-serious (imaginings, "aliefs", etc.) cognitive states, we may derive benefit from affective engagement with robot pets based on non-serious cognitive modes about some of the robot pet's properties.

At this point, however, a moral objection may arise. Suppose for the moment that people who affectively engage with robot pets do indeed typically do so on the basis of non-serious cognitive modes. Suppose also that these were the only kind of affective engagement that these people were entertaining despite the availability of affective engagements based on (truth-sensitive) beliefs. The situation of these individuals would parallel that of those who spent much of their life immersed in their novels or watching their favourite TV-series. While literally speaking not misrepresenting reality, some would say that these individuals are in some sense living a life at some remove from reality. And while not misrepresenting reality, they may engage in these activities in order to indulge in some feelings (or perhaps in order to avoid other feelings that they would otherwise have to confront). While we will return on the moral status of such cases later, here we should note that, strictly speaking, this objection does not have to do with sentimentality.

The wrongness of sentimentality

If you found the above line of argument overall unconvincing, if, that is, you still believe that affective engagement with robot pets should be understood as involving sentimentality, you may want to consider the moral case for or against sentimentality. As shown above, Sparrow believes that sentimentality is morally wrong when it violates the weak duty to our selves to apprehend the world accurately. But what exactly would be wrong with failing to apprehend the world accurately, above all when by doing so no one else is being harmed and when there is a suspicion that a subject's acquiring or entertaining incorrect beliefs may actually be beneficial to her?[9]

In a very balanced contribution, Blackford (2012) challenges Sparrow's stance on the morality of sentimentality. While we should certainly admit that obtaining a correct apprehension of the world is usually valuable to ourselves as individuals, and to those with whom we interact, Blackford argues that some categories of misapprehension (usually or often) contribute to our well-being. Such is the case for example for the exaggerated ideas of our abilities and control which most of us tend to entertain (Taylor and Brown 1988; Snyder and Higgins 1988). Such self-misperceptions have been shown to contribute to our happiness and to have an adaptive advantage (Taylor and Brown 1988, pp. 194, 197–199). In light of this prudential advantage, and in the absence of harm to others, it is hard to see why we would have a moral duty to ourselves to cease to entertain these misapprehensions.

Blackford's argument has a similar structure with regard to cases of anthropomorphic sentimentalization e.g. of cars or computers, cases of exaggerated feelings that many people have for sporting teams, popular entertainers, fictional characters, and, finally, affective engagement with robot pets. In all such cases we have a propensity to experience strong emotions for individuals or entities, fictional or otherwise, who know nothing about our existence or who cannot otherwise reciprocate. Given this psychological propensity, however, and the fact that entertaining the misapprehension (allegedly) involved by having these emotions may actually do a certain amount of good at little cost to ourselves or others, Blackford rejects the idea of a moral duty to ourselves to cease to entertain these (alleged) misapprehensions.

In what follows, granting that these cases do involve misapprehension, I will not challenge Blackford's conclusions. I will, however, consider another argument for the wrongness of sentimentality, one that is not based on an alleged duty to apprehend the world correctly. I will then

[9] As David Pugmire colourfully puts it (2005: p.125): "The problem might instead be with the very animus against sentimentality. Mightn't that reflect a discomfort with emotion as such, especially with being agitated by emotion, a kind of psychic Calvinism? What can really be wrong with sunning oneself with the mellower sentiments and bathing in kindly feelings, in making opportunities for feeling good? Recessiveness might rather lie in the curmudgeonly allergy to this. Surely it is morbid to spurn the nurturing of benign feeling where that allows itself, which, after all, it so often doesn't."

apply this argument to the case of affective engagement with robot pets and show that even this argument fails.

Those interested in finding the wrongness of sentimentality must look somewhere other than in the direction of duties to apprehend the world correctly. A promising alley is to consider not merely the fact that in being sentimental one misrepresents reality but rather the fact that the misrepresentation is to some extent intentional or voluntary. Doing so allows us to place greater focus on the nature of the motivation that drives the misrepresenting. In order to make sense of this line of thought, we need to work with a characterization of sentimentality that is fuller than the one used so far. Pugmire (2005: p.127–128) offers precisely that:

> The notion that sentimentality is a kind of vice rests on an intuition… that somehow the emotion is being aroused dishonestly and is being used. We are struck by how the emotion has insinuated itself where it is out of keeping, and this not through confusion or mere ignorance but through an indulgent and even insistent disregard for its misplacement. When we bridle at emotion that runs to sentimentality, the dishonesty we register is [such] that the emotion is not sustained by a truthful picture of what it is about, of its ostensible focus of concern, which I will term its *theme*. … The construction a sentimentalist puts on the theme is not truthful, not the one an unneedy (or courageous) observer would put on it. …. [T]he sentimentalist will select, idealize (or vilify), sanitize, embroider, and even fantasize, as necessary. And he needs to be a person or persons who ought to know better, not someone, like a child, who is just fanciful or very naive, who may be under-equipped without being dishonest. The fond enthusiasms of the inadequately (or not yet) sophisticated may be callow, but quaintly, endearingly, or sadly, rather than corruptly, so. As for the sentimentalist we are concerned with, were he not needful of the emotion in question, he would not take the view he takes of its theme. He has powers the simple lacks but devotes them to sophistry.

A couple of points are worth highlighting here. First, the emphasis is on the "indulgent and even insistent disregard" that is actively pursued, and not as the result of confusion or ignorance. Pugmire excludes children, the naïve, the "simple" and those who are otherwise "underequipped" from the charge of sentimentality: they simply lack the relevant intention (and skills) to be viciously sentimental. It is not that these individuals are being sentimental without being viciously so. They are not being sentimental at all. Second, the intention of those who are being sentimental is motivated by a need to evoke some emotion. The emotion is actively sought because the person somehow needs to feel that emotion, irrespective of whether the emotion is called for by the theme. Sentimentality relies in portraying the theme in a way that is conducive to the desired emotional effect and subordinates the truth about it to this end. Sentimentality involves the attempt to use something to secure a desired feeling or emotional comfort.

Pugmire (2005: p. 126) does not think that all forms or instances of sentimentality are bad or vicious. While some of them are indeed degraded, others are innocent, and others salutary. It would not necessarily be bad to be sentimental for someone who has a helpless need for solace, for cherishing a benign memory of a deceased loved one. The moment one agrees, as both Sparrow and Pugmire seem to do, that sentimentality is not as such always wrong, bad or vicious, we must begin to look for features that do not essentially belong to it but which, when present, make instances of sentimentality wrong, bad, or vicious. Hence, according to Pugmire (2005: 136), the motivation behind sentimentality

> becomes corrupting where the concern of the type of emotion in question is with something outside oneself which should command our regard on its own account. It is less corrupting where the theme is not anything that could do this…

Now Pugmire continues (2005: 136), a thing is not regarded or respected for what it is

> when a feature of it that is more momentous than its power to stir me in a certain way is occluded in order that it may better serve to stir me. Then the emotion is profane. Where no such violation is involved, nothing is lost by so arranging the theme as to make one feel good.

Fault depends not just on one's relationship to feeling but on this together with the relation of the feeling to what it is about. To sum up, then, sentimentality is to be understood as (often) relying on the active occlusion of reality motivated by a need or desire to secure emotional comfort. This motivation, however, is not yet enough to make sentimentality bad, profane, vicious, or faulty. What is also required is that in being sentimental one does not respect a thing for what it is. If for example I betray to an audience the secrecy of my friend's private sorrows in order to bask in the approval, common commiseration and admiration of this crowd, my sentimentality is vicious. My friendship and my friend command greater respect than I show when I use them in order to stir some feelings in me.

Next we should assess whether engaging affectively with robot pets would be viciously sentimental in this sense. It is clear from this account of sentimentality that we may have to exclude at least some of the subjects we have

in mind from counting as being sentimental at all, for they may be underequipped to count as actively disregarding the truth in the way discussed above. But let's assume that at least some of our subjects do actively engage in this type of dishonesty in order to secure some emotional comfort. The question now is whether their doing so is bad or vicious, in other words, whether in doing so they are disrespecting something for what it is in order to achieve their end.

The task at hand, then, is to understand what these subjects would be disrespecting. Assuming that they would not be disrespecting the robot pets, it looks initially like we have two options. Either they would be disrespecting themselves or some aspects of themselves, or they would disrespect others or some aspects of others. As noted above, however, both options take us away from the idea that the wrongness of sentimentality resides somehow in itself. This would not as such be a problem, if we could find inherent or systematic connections between the sentimentality involved in affective engagement with robot pets, on the one hand, and certain types of disrespect, on the other. If, for example, the sentimentality at issue here systematically or even just typically involved, say, violations of duties towards one's family and friends, then, we could say that this type of sentimentality is wrong in this respect.

There is no reason to suspect, however, that sentimentalist robot pet owners would systematically or even typically commit more wrongs than any odd individual, nor that they would systematically or even typically commit wrongs of a *specific kind*. Of course, at times, we can expect some of these individuals' sentimentality to be vicious, insofar as it disrespects other persons, values, or duties that the individuals have. But this would not point to anything peculiarly wrong with being sentimental about robot pets, which would in this case be as wrong as any other self-interested pursuit that occasionally involves wrongdoing.

In conclusion, the aim of this section was to give the argument from sentimentality another shot by using a different account of sentimentality and its wrongness. On this account, the wrongness of sentimentality is, at least initially, understood as resting on the quality of one's will in misrepresenting, rather than on one's failed duty to represent the world accurately. What the section shows, however, is that even on such characterization our thinking about the issue is such as to move away from the idea that there is something inherently wrong with sentimentality. Whenever we find something wrong with it, sentimentality is wrong insofar as it involves other things that are themselves thought to be wrong. I think this is an important lesson, which should contribute to putting an end to appeals to the vice of sentimentality

Conclusion: the morality of ersatz companionship technology beyond sentimentality

The conclusions reached so far may be taken to make a positive moral case in favour of robot pet ownership and production and, from there, in favour of ersatz companionship technology more in general. The argument in favour of this case would go as follows: if sentimentality alone constituted the inherent wrong in the consumption and thereby the production of such technology, and we were correct in thinking that sentimentality is neither at issue, nor inherently wrong, then, production and consumption of such technology should be morally permissible. The question now, however, is whether sentimentality constitutes the only, or even the major, moral objection to the consumption and production of this technology. In this concluding section, I will argue that there is at least another important worry to consider.

In order to bring this worry to the fore, we have to shift the focus of the current discussion from the potential wrongness of instances of a type of action *considered in isolation*, to the potential wrongness of the same type of action *in the context of wide-spread diffusion of it.* To home in on what I have in mind, consider the moral implications of driving a standard car first in isolation, where, for all we know, this may be the only car being driven on the planet, and then in a context in which the number of cars in use surpasses one billion. This change in context or contextual knowledge makes a significant difference to our moral assessment of the action. Similarly, in light of the conclusions reached so far, I would argue that we stop considering the morality of individual acts of ownership or production of robot pets taken in isolation, and consider instead the wider discussion of ersatz companionship across the board and as a socially significant phenomenon.

When considered in this way, the fear is that what was originally to function as a *substitute* for human or pet companionship will progressively and irreversibly *replace* human or pet companionship as the standard form of companionship. Why bother with real pets when you can do with something that is as good in all respects but has no inconveniences? Think about an otherwise normal looking pet who does not die or get ill (unless programmed to do so), who does not need flee cures or sterilization, who does not shed its fur all over the house (despite having a nice and clean fur), who does not need to be left in a kennel during the summer holidays, with the financial costs and the bad conscience that ensue from it, and that does not otherwise restrict your freedom unless and to the extent to which you agree to it. Similarly, why have a real human friend or partner when you can have one that is as good

minus whatever inconveniences (e.g. lack of reliability, potential for betrayal, moodiness, disagreements, ugliness) one finds in friends or partners?

Many consider authentic and reciprocal human-to-pet relationships, and human-to-human relationships such as friendship and romantic relations not only to have intrinsic value but also to define who and what we are. Anything that threatens these relationships is therefore not only a threat to something that we take to be of value but also a threat to something that defines us as the kind of beings we are. To illustrate this point, I will consider three quick examples. First, we shall consider the case of autism. Many find those autistic individuals incapable of normal human-to-human social interaction as "alien" and "alienating".[10] We have trouble seeing the "human being" in another human being when the latter cannot relate to us according to the standard modes of social engagement, when the other's sociality is significantly different from ours. I do not cite this stance towards autistic individuals to defend it but to adduce it as evidence that certain forms of sociality are strictly connected to our self-understanding. If widespread diffusion of ersatz companionship technology altered or obliterated the way in which we relate to each other, our forms of sociality, it may bring in its train loss of valuable relations, forms of life, and self-understanding.

Second consider what, in his book about human happiness, Daniel Russell (2012: p. 5) had to say about the happiness of individuals that are not quite like the rest of us emotionally:

> Someone incapable of loving others [or emotionally childish] for instance, might have a life that is fulfilling for him as it can be, given that unique make-up of his, but we would not point to his life as a good example of happiness. (It's certainly no life one would wish on a friend). If that is so, then we can understand happiness only by keeping in view that it is *happiness for humans* that we are talking about.

Someone who cannot quite love like the rest of us do is considered emotionally childish, not quite a full or complete human being in that respect. According to Russell and like-minded people, this individual will not be able to enjoy happiness for humans. Once again, specific forms of (affective) relations are taken to be defining our self-understanding. Ersatz companionship may indeed make us emotionally childish in some sense, incapable or at any rate unable or unwilling of loving other human beings with all their "imperfections", of negotiating those difficulties that often make a bond between friends or lovers even stronger or deeper. One could indeed legitimately worry whether the happiness we would enjoy from relations with social robots is indeed *happiness for humans*, the kind of happiness that is proper to us as humans.

Finally, consider the nature of sexual relations. According to an established view (Ruddick 1975: pp. 83–84)[11]:

> The completeness of a sexual act depends upon the *relation* of the participants to their own and each other's *desire*. A sex act is complete if each partner allows himself to be "taken over" by an active desire, which is desire not merely for the other's body but also for his [the partner's] active desire. … "It is important that the partner be aroused, and not merely aroused, but aroused by the awareness of one's desire".

I assume that the day in which we can have complete sexual acts with robot partners is far away, if ever within reach. We may yet find robot sex convenient enough. And if we did so in a large enough scale, we may end up witnessing the gradual vanishing of our current understanding of complete sexual acts along with the valuable form of life that embodies it.

Whether ersatz companionship technology would have these effects, i.e., changing our forms of sociality, our capacity to love, and to have complete sexual relations, however, is an open question. For all we know, the new human-pet robot relationships could actually enable us to develop our humanity, namely by way of the emotions and attachments that they involve. In line with the argument above, just as large scale engagement with art and the emotions that it generates is presumably conducive to our humanity (Nussbaum 1997, 2001), large scale human-pet robot relationships may also become a training ground for emotions conducive to our humanity.[12]At this point in time, then, we should not take ourselves to have evidence that the large-scale advent of ersatz companionship technology is a danger to valuable relations, our self-understanding, and our form of life.

We should rather take the above examples to be enough of an argument to the effect that academics and society at large ought morally to assess the extent to which the fears I have just sketched are grounded. This would be a complex intellectual exercise involving, in its first phase, an assessment informed by disciplines such as anthropology, philosophy, robotics, and sociology, of what realistically we are to lose as the result of large scale ersatz companionship. In a second, even more speculative phase, we would have to figure out exactly what we are to gain from it, what new forms of life we can expect to arise and what

[10] See for example the experiences reported in Collins (2004).

[11] Ruddick refers to Jean-Paul Sartre, Maurice Merleau-Ponty, and Thomas Nagel as endorsing versions or elements of this view. The quote reported in the quote above is from Nagel (1969: p.13).

[12] I thank an anonymous referee for this suggestion.

would be good about them. It may turn out that the best, or perhaps only possible way of doing this is by introducing this ersatz companionship in society as an 'ongoing social experiment' that requires close monitoring and specific procedures (van de Poel 2011, in press).

Meanwhile, in the light of the discussion broached in the previous sections, it does not make much sense to focus on the immorality of each individual's affective engagement with robot pets, be it sentimental or non-sentimental. There is nothing inherently wrong with either kind of affective engagement *when considered in isolation* just as there is nothing *inherently* wrong with the production of the same technology. Of course, as argued above, we do know that this technology can be used in ways that are beyond the moral pale. It may, for example, be at least initially morally questionable for an individual to provide ersatz companionship to his ageing parent, when he could himself have attended to her to no great cost of his own (Sharkey and Sharkey 2012). At the same time, however, there may be circumstances in which providing the same type of technology to someone may be the best one can do and, in fact, something good to do. At the moment, that is, we cannot say that there is an inherent connection between technology of this kind and immoral behaviour or bad outcomes. Yet if enough individuals embraced enough ersatz companionship technology, we may end up living in a world quite unlike this one, where many of the valuable things we now enjoy would be lost.

It may help to envision our current moral situation with regard to the production and ownership of ersatz companionship technology as similar to our situation with regard to the production and ownership of automobiles as it would have looked like to someone at the beginning of the 20th century. Back then, no one would have believed that cars were to have the success that they have had, no one would have imagined the ways in which cars would change our mode of life (for better or worse), the wars, deaths, and exploitation that securing petrol supply would have contributed to cause, and the impact that cars and the mode of life they allowed would have on the environment. It is hard to find moral fault with individual ownership or production of cars at that point in time. Today, however, *in the context of mass ownership and production of cars*, we know or should know that individual usage of cars has negative consequences for the environment and for human beings now and for generations to come. The same goes for car producers. But this knowledge has come too late. Significant parts of our modes of life and societies are built around car ownership. Many find it extremely difficult if not impossible to switch to another, car-free, mode of life. The challenge should be understood as largely a societal one and not merely one of individual morality.

Just as things may have looked in the automobile case a century ago, at the moment, we are uncertain about the potential for development of such ersatz companionship technology (how good or realistic will artificial companions be?), we do not know how welcomed and therefore diffused they will be, and it is therefore hard to know whether they will have an impact on our modes of life, and if they will, what exactly the nature of that impact would be. Given our current context and our current knowledge about its likely development we cannot find an inherent connection between this technology and immoral behaviour or bad outcomes. Such connection may however appear if the context changed and this technology became sufficiently wide-spread. In such context, individual consumption and production of ersatz companionship technology may indeed mean loss of forms of life that are, both, intrinsically valuable and self-defining. What follows morally for us now, is the imperative to figure these issues out in a responsible manner.[13,14]

References

Blackford, R. (2012). Robots and reality: A reply to robert sparrow. *Ethics and Information Technology, 14*, 41–51.

Broekens, J., Heerink, M., & Rosendal, H. (2009). Assistive social robots in elderly care: A review. *Gerontechnology, 8*(2), 94–103.

Carroll, N. (1990). *The philosophy of horror; or paradoxes of the heart*. New York: Routledge.

Charlton, W. (1984). Feeling for the fictitious. *British Journal of Aesthetics, 24*(3), 206–216.

Coleridge, S. T. (1907). *Biographia Literaria*, J. Shawcross (Ed.), Oxford: Oxford University Press. http://www.gutenberg.org/files/6081/6081-h/6081-h.htm.

Collins, P. (2004). *Not even wrong: A father's journey into the lost history of autism*. New York: Bloomsbury.

Cova, F., & Teroni, F. (2015). Le paradoxe de la fiction: le retour. In Martin Rueff & Julien Zanetta (eds.) *L'expression des émotions: Mélanges en l'honneur de Patrizia Lombardo*, Genève, 2015, URL: http://www.unige.ch/lettres/framo/melangeslombardo.html.

De Sousa, R. (1987). *The Rationality of Emotion.* Xxxx MIT Press.

de Sousa, R. (2002). Emotional truth. *Proceedings of the Aristotelian Society*, *Supp. Vol.* 76.

de Sousa, R. (2007). Truth, authenticity, and rationailty. *Dialectica* XXxx, pp. 323–345.

Gendler, T. (2008a). Alief and belief. *Journal of Philosophy, 105*(10), 634–663.

Gendler, T. (2008b). Alief in action (and reaction). *Mind and Language, 23*(5), 552–585.

[13] Once again, Van de Poel's (2011; in press) idea of the introduction of new technologies in society as 'ongoing social experiment' that requires close monitoring and specific procedures may be very relevant here.

[14] I woudl like to thank Johanna Seibt and two anonymous referees for this journal for their extremely insightful comments, the audience at the Robophilosophy Confernce 2014 at Aarhus University, and the Velux Fonden for its generous financial support.

Gendler, T., (2013). "Imagination, Edward N. Zalta (ed.). *The Stanford Encyclopedia of Philosophy* (Fall 2013 Edition) http://plato.stanford.edu/archives/fall2013/entries/imagination/.

Kidd CD, Taggart W, Turkle S.(2006). A sociable robot to encourage social interaction among the elderly. In: Proceedings of ICRA 2006: IEEE International Conference on Robotics and Automation; pp 3972–3976; doi:10.1109/ROBOT.2006.1642311.

Lamarque, P. (1981). How can we fear and pity fictions? *British Journal of Aesthetics, 21*(4), 291–304.

Midgley, M. (1979). Brutality and sentimentality. *Philosophy, 54*(209), 385–389.

Nagel, T. (1969). Sexual perversion. *The Journal of Philosophy, 66*, 1.

Neill, A. (2005). Art and Emotion. In J. Levinson (Ed.), *The oxford handbook of aesthetics* (pp. 421–435). New York: Oxford University Press.

Nussbaum, M. (1997). *Cultivating humanity: A classical defense of reform in liberal education*. Cambridge: Harvard University Press.

Nussbaum, M. (2001). *Upheavals of thought: The intelligence of emotions*. Cambridge: Cambridge University Press.

Paskins, B. (1977). "On being moved by Anna Karenina and *Anna Karenina*". *Philosophy, 52*, 344–347.

Prinz, J. (2004). *Gut reactions: A perceptual theory of emotion*. Oxford: Oxford University Press.

Pugmire, D. (2005). *Sound sentiments*. Oxford: Oxford University Press.

Radford, C. (1975). How can we be moved by the fate of Anna Karenina? *Proceedings of the Aristotelian Society, Supplemental, 49*, 67–80.

Ruddick, S. (1975). Better sex. In R. B. Baker & F. A. Elliston (Eds.), *Philosophy and sex* (pp. 83–104). Buffalo: Prometheus.

Russell, D. (2012). *Human happiness*. Oxford: Oxford University Press.

Salmela, M. (2006). True emotion. *Philosophical Quarterly, 56*(224), 382–405.

Schneider, S., (2006). "The Paradox of Fiction," in *The Internet Encyclopedia of Philosophy*, http://www.iep.utm.edu/f/fict-par.htm.

Seibt, J. (2014). Varieties of the"As-if": Five ways to simiulate an action. In: J. Seibt, R. Hakli, M. Nørskov, *Sociable robots and the future of social relations*. Proceedings of Robo-Philosophy 2014, IOS-Press, Amsterdam. pp. 97–105.

Sharkey, A. J. C., & Sharkey, N. E. (2012). Granny and the robots: Ethical issues in robot care for the elderly. *Ethics and Information Technology, 14*, 27–40.

Smith, M. (1995). Film spectatorship and the institution of fiction. *Journal of Aesthetics and Art Criticism, 53*(2), 113–127.

Snyder, C. R., & Higgins, R. L. (1988). Excuses: Their effective role in the negotiation of reality. *Psychological Bulletin, 104*, 23–35.

Sparrow, R. J. (2002). The march of the robot dogs. *Ethics and Information Technology, 78*(3), 346–359.

Taylor, S. E., & Brown, J. D. (1988). Illusion and well-being: A social psychological perspective on mental health. *Psychological Bulletin, 103*, 193–210.

Teroni, F. (2007). Emotions and formal objects. *Dialectica, 61*(3), 395–415.

Tullmann, K., & Buckwalter, W. (2014). Does the paradox of fiction exist? *Erkenntnis, 79*, 779–796.

Van de Poel, I. (2011). Nuclear energy as a social experiment. *Ethics, Policy, and the Environment, 14*(3), 285–290.

Van de Poel, I. (in press). Society as a laboratory to experiment with new technologies. In E. Stokes, D. Bowman & A. Rip (Eds.), *Embedding and governing new technologies*. Singapore: Pan Stanford Publishing.

Wada, K., & Shibata, T. (2007). living with seal robots— its sociopsychological and physiological influences on the elderly at a care house. *IEEE Transactions on Robotics, 23*(5), 972–980.

Walton, K. (1978). Fearing fictions. *Journal of Philosophy, 75*(1), 5–27.