ETHICAL RISK ASSESSMENT FOR SOCIAL ROBOTS: CASE STUDIES IN SMART ROBOT TOYS

Alan F. T. Winfield, Bristol Robotics Laboratory, University of the West of England, alan.winfield@brl.ac.uk

Anouk van Maris, Bristol Robotics Laboratory, University of the West of England, anouk.vanmaris@uwe.ac.uk

Katie Winkle, KTH Royal Institute of Technology, School of Electrical Engineering and Computer Science, winkle@kth.se

Marina Jirotka, Department of Computer Science, University of Oxford, marina.jirotka@cs.ox.ac.uk

Pericle Salvini, Department of Computer Science, University of Oxford, pericle.salvini@cs.ox.ac.uk

Helena Webb, Department of Computer Science, University of Oxford, helena.webb@cs.ox.ac.uk

Arianna Schuler Scott, Department of Computer Science, University of Oxford, arianna.schuler.scott@cs.ox.ac.uk

Jaimie Lee Freeman, Oxford Internet Institute, University of Oxford, jaimie.freeman@oii.ox.ac.uk

Lars Kunze, Oxford Robotics Institute, Department of Engineering Science, University of Oxford, lars@robots.ox.ac.uk

Petr Slovak, Department of Informatics, King's College London, petr.slovak@kcl.ac.uk

Nikki Theofanopoulou, Department of Informatics, King's College London, nikki.theofanopoulou@kcl.ac.uk

Abstract (max. 200 words) [110]

Risk Assessment is a well known and powerful method for discovering and mitigating risks, and hence improving safety. Ethical Risk Assessment uses the same approach, but extends the scope of risk to cover ethical risks in addition to safety risks. In this paper we outline Ethical Risk Assessment (ERA), and set ERA within the broader framework of Responsible Robotics. We then illustrate ERA, first with a hypothetical smart robot teddy bear (RoboTed), and later with an actual smart robot toy (Purrble). Through these two case studies this paper demonstrates the value of ERA and how consideration of ethical risks can prompt design changes, resulting in more ethical and sustainable robots.

Keywords: Ethical Risk Assessment, Responsible Robotics, Social Robots, BS8611, Smart Robot Toy

1. Introduction

Risk assessment is a well-known method for discovering and mitigating risks, and hence improving safety. Ethical Risk Assessment is not new; it is essentially what research ethics committees do [1]. But the idea of extending the scope of safety risk assessment of intelligent systems to encompass ethical risks is new. Given the growing awareness of the ethical risks of intelligent systems in recent years, ethical risk assessment offers a powerful method for systematically identifying and mitigating the ethical, societal and environmental risks associated with the use of robots and artificial intelligence (AI).

In Section 2 we first define ethical risk assessment (ERA) with reference to British Standard BS8611. Then, in section 3, we determine whether this standard can be used as a guideline for ERA by testing it on a hypothetical smart robot teddy bear we call RoboTed. A fictional robot is used here as it allows us to evaluate a broad range of technological features that may not all be available in existing robots. In order to determine whether ERA is applicable to real-world robots, section 4 goes on to evaluate an existing smart robot toy called Purrble[1]. We believe this to be the first work applying ethical risk assessment to a commercial smart robot toy. The paper concludes with a comparison of the assessment of the two robot toys and provides an appraisal of both the benefits and limitations of ERA. This paper is an extended version of a paper presented at ICRES 2020 [2].

2. Ethical Risk Assessment

Risk Assessment is a process that typically has three stages:

1. identify and analyse potential events (hazards) that may cause harm to individuals, property, and/or the environment;
2. make judgments on the acceptability and likely impact of the harm arising from exposure to the hazard (risks), then
3. determine what steps should be taken to mitigate those risks and hence minimise or eliminate possible harms.

Standards for risk assessment are well established in safety critical systems. ISO 14971:2007 Application of risk management to medical devices, for instance, provides requirements and guidance for risk assessment for medical devices. And ISO 12100:2010 *Safety of machinery - Risk assessment and risk reduction* sets out requirements for performing risk assessments, notably including risk analysis focused on hazard identification.

Almost certainly the world's first explicitly ethical standard in robotics is BS8611-2016 *Guide to the ethical design and application of robots and robotic systems.* "BS8611 is not a code of practice, but instead guidance on how designers can undertake an ethical risk assessment of their robot or system, and mitigate any ethical risks so identified. At its heart is a set of 20 distinct ethical hazards and risks, grouped under four categories: societal, application, commercial and financial, and environmental. Advice on measures to mitigate the impact of each risk is given, along with suggestions on how such

[1] https://purrble.com/

measures might be verified or validated" [3]. Societal hazards include, for example, anthropomorphisation, loss of trust, deception, infringements of privacy & confidentiality, addiction, and loss of employment.

BS8611 defines an ethical harm as "anything likely to compromise psychological and/or societal and environmental well-being", an ethical hazard as "a potential source of ethical harm"; and an ethical risk as the "probability of ethical harm occurring from the frequency and severity of exposure to a hazard" [4]. Ethical risk assessment thus extends the scope of risk assessment to include ethical harms, hazards and risks (in addition to physical harms, hazards and risks).

Psychological safety is a well-known topic in Human-Robot Interaction (HRI) studies, although often overlooked [5]. It is concerned with the reduction of stress and anxiety caused by a robot's appearance and behaviour (such as size, shape, adherence to social norms) and the potential impact of robots that can emulate human or animal characters (such as feelings and expression of emotions).

Among the most relevant guidelines on psychological safety for robotics and AI currently available, or in draft, are: British Standard BS8611, a guide to the ethical design and application of robots [4], the European Parliament report on recommendations to the Commission on Civil Law Rules on Robotics [6] and the IEEE standards project P7014™ on ethical considerations in emulated empathy in autonomous and intelligent systems [7].

Ethical Risk Assessment is a fundamental part of the practice of Responsible Robots, which we define as "the application of Responsible Innovation [8] in the design, manufacture, operation, repair and end-of-life recycling of robots, that seeks the most benefit to individuals and society and the least harm to the environment" [9]. We would expect ERA to be undertaken within a framework of responsible innovation – such as EPSRC's AREA framework[2] – and alongside ethically aligned [10] and values-based design [11].

[2] https://epsrc.ukri.org/research/framework/area/

In the following two sections, ERA will be applied to both fictional and existing smart robot toys, to determine whether ERA can be applied to a broad range of technological features in smart robot toys.

3. Case Study with a Fictional Smart Toy: RoboTed

Consider a hypothetical smart toy in the form of a teddy bear named RoboTed. RoboTed is a simplified version of the robot Teddy from the 2001 movie *A.I. Artificial Intelligence*, directed by Steven Spielberg[3], to better reflect today's capabilities (see Figure 1).

Figure 1: Teddy

This fictional RoboTed is designed to:

1. recognise its owner, learning their face and name, turning its face toward the child,
2. respond to physical play such as hugs and tickles,
3. tell stories, while allowing a child to interrupt the story to ask questions or ask for sections to be repeated,
4. sing songs, while encouraging the child to sing along and learn the song, and
5. act as a child minder, allowing parents to remotely listen, watch and speak via RoboTed.

[3] as a tribute to Stanley Kubrick.

For these functionalities, RoboTed is based upon the following technology:

1. It is an Internet (WiFi) connected device,
2. It has cloud-based speech recognition and conversational AI (chatbot) and local speech synthesis,
3. Its eyes are functional cameras allowing the robot to recognise faces,
4. It has motorised arms and legs to provide it with limited baby-like movement and locomotion – not walking but shuffling and crawling, and
5. It has touch sensors which allow it to respond to physical play.

As a worked example we now consider the ethical hazards and risks of RoboTed, under the four categories of: physical (safety) risks, psychological risks, privacy & security risks, and environmental risks.

3.1 Physical Risks

Tripping – as RoboTed crawls on the floor, it has the potential to become a trip hazard. A mitigation strategy might be to have RoboTed make an audible crawling sound when it is moving, to alerts users (particularly adults) to its presence.

Battery overheating – There is a risk that defective batteries or battery chargers can overheat or *in extremis* catch fire. In mitigation RoboTed should be designed to make use of low-risk consumer rechargeable batteries rather than high-risk Li-Ion batteries. In addition parents should be advised to supervise battery recharging.

3.2 Psychological Risks

Addiction – RoboTed might be so compelling that it leads to a child playing obsessively with RoboTed and neglecting his or her family [12]. This also increases the risk of emotional distress should RoboTed's behaviour change or fail in any way (e.g. if the facial recognition was to fail and no longer recognises the child). A mitigation strategy might be to explore the addition of a RoboTed 'needs to sleep' function, as a way of limiting length of play times.

Deception – There is a risk that the child comes to believe that RoboTed has feelings for her [13]. To mitigate this risk we could design the chatbot to avoid language that suggest feelings, so that RoboTed never says things like 'I like you' or 'Why are you sad?'

Over trusting by the child – Building on deception there is a risk that the child cannot tell whether RoboTed is operating autonomously or is in the child minder mode. This may result in her sharing sensitive information in the belief that no one else will hear it, when actually her parents are watching and listening. The reverse is also true, in that she may share something she wishes her parents to know but is too embarrassed to raise face to face, when actually the robot is operating autonomously and her parents are *not* listening. Mitigation strategies would be concerned with making the mode of operation as obvious to the child as possible, for example only using RoboTed's speech synthesis when in autonomous operation.

Over trusting by parents – The risk here is that parents become over reliant on RoboTed's child minder function [14]. The risk and its consequences are so great as to suggest the child minder function should be removed altogether.

The Uncanny Valley – The Uncanny Valley can lead to a fearful reaction when a robot is close to but not 100% lifelike [15]. The risk of this is probably low with RoboTed, both because RoboTed is not human-like at all, and children are already familiar with teddy bears. However, the risk should be explored by engaging children in early trials of RoboTed, and if the uncanny valley reaction is demonstrated it might be mitigated by, for instance, equipping the robot with a cartoon voice.

3.3 Privacy and Security Risks

Weak security – Weak security could lead to malicious hackers gaining access to RoboTed's sensors & control functions. This could be very frightening for a child and her parents. To reduce the risk we need to implement strong encryption of the communications between RoboTed and the cloud, alongside best practice password protection to make it very hard for hackers to guess the password.

Privacy – Here the risk is that personal data, including images and voice recordings of children (and the house they live in) are stolen. One way of reducing this risk would be to ensure that personal data sent to the cloud is deleted immediately after it has been used.

Lack of transparency – The risk is that if there were an accident in which RoboTed harmed a child, that could be either physical or psychological harm, it would be very difficult to investigate what happened to cause the accident unless the robot keeps a data log of its actions and responses. This is a serious risk and to mitigate the risk a secure data logger (*ethical black box*) needs to be built into RoboTed [16]. The data would be stored locally, and only the most recent few hours of data would need to be saved.

3.4 Environmental Risks

Unsustainability of materials – Here the risk is that the robot uses unsustainable or high carbon materials. To mitigate this risk we could use materials (e.g. RoboTed's fur) from sustainable sources. We could also avoid plastics by, for instance, using wood for RoboTed's skeleton.

Unrepairability – This leads to the risk that the robot's lifetime is limited because faults cannot be repaired or parts replaced. This risk can be minimised by designing RoboTed for ease of repair, using replaceable parts as much as possible (especially the battery). Additionally, RoboTed's manufacturers should provide a repair manual so that local workshops can fix most faults.

Unrecyclability – All products will eventually come to the end of their useful life, and if they cannot be repaired or recycled we risk them being dumped in landfill. To mitigate this risk, RoboTed should be designed to make it easy to recycle parts. Ideally after these parts have been recovered for recycling, the remaining materials are biodegradable.

Table 1: Ethical risk assessment of RoboTed

Hazard	**Risk**	**Level**[4]	**Mitigation**
Physical risks			
Tripping	User(s) trip over RoboTed when it is crawling on the floor	M	Audible crawling sound to alert users to its presence
Battery overheating	Defective batteries or battery chargers can overheat or *in extremis* catch fire	M	Design to make use of consumer rechargeable batteries rather than high-risk Li-Ion batteries
Psychological risks			
Addiction	Child plays with RoboTed obsessively and neglects family	M	Explore 'RoboTed needs to sleep now' function
Deception (of child)	Child believes that RoboTed has feelings (for her)	M	Design chatbot to avoid language that suggests feelings
Over trusting (by child)	Child cannot distinguish mode of operation	H	Notification when child minder mode activated, uses parents' voice rather than RoboTed's voice
Over trusting (by parents)	Parents come to rely on the childminder function	H	Remove the childminder function
The Uncanny Valley	Child becomes fearful of Robot	L	Use cartoon voice; engage children in early user trials
Privacy and security risks			

[4] *Note:* Risk level: (H)igh, (M)edium or (L)ow

Weak security	Malicious hackers gain access to RoboTed's sensors and control function	H	Implement strong encryption together with best practice password protection
Privacy	Personal data, including images and voice recordings of child are stolen	M	Put in place auditable measures to ensure personal data is deleted immediately
Lack of transparency	Lack of data logs makes it hard or impossible to investigate accidents	H	Build a secure local data logger into RoboTed
Environmental risks			
Unsustainability (of materials)	Robot uses unsustainable or high carbon cost materials	M	Use materials (e.g. RoboTed's fur) from sustainable sources, avoiding plastics
Unrepairability	Robot's lifetime is limited because faults cannot be repaired or parts replaced	M	Design for ease of repair with replaceable parts – especially battery
Unrecyclability	End of life robots are dumped in land fill	M	Design for ease of recycling parts and materials

3.5 Discussion of ERA for RoboTed

The evaluation of potential risks of RoboTed, as summarized in Table 1, has demonstrated the value of ethical risk assessment. It has shown that a focus on ethical risks can:

- suggest new functions, such as 'RoboTed needs to sleep now',
- draw attention to how designs can be modified to mitigate some risks,

- highlight the need for user engagement, and
- reject some product functionality as too risky.

Testing ERA with a fictional robot has value as it allows us to cover a broad range of functionalities and hazards. However, for a more complete understanding of the application and limitations of ERA, we next consider an existing smart robot toy.

4. Case Study with Existing Smart Toy: Purrble

Our second case study focusses on an existing commercial product emerging from HCI research, named Purrble (see Figure 2). Purrble is a small, interactive plush animal that has been specifically designed to provide an in-situ emotion regulation support to children, which is specifically 'child-led': this means that the design assumes that no training should be required for the child (or their adults) for the emotion regulation effects to occur. A full description of the design process and the intervention theory of change can be found in Theofanopoulou et al [17] and Slovák et al [18].

Figure 2: Purrble, taken from www.purrble.com {accessed 10-01-2020)

In summary, the robot is presented to children as a vulnerable creature, which is often anxious (indicated by a fast heartbeat-like vibration) that calms down when cuddled. The expectation is — and empirical data from previous studies suggests — that by ‘soothing’ the robot, the child down-regulates themselves; that children will start seeking the Purrble when they themselves are distressed, and that these repeated engagements can constructively shift emotion-regulation processes in the family.

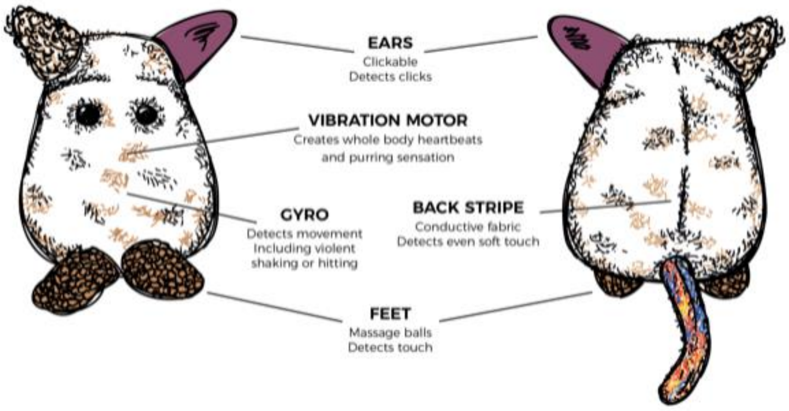

Figure 3: Technological features of Purrble (taken from [17])

Technologically, the robot is very simple, as presented in Figure 3:

1. The only modes of communication are vibration patterns (fast to slow heart-beats, and a purr when soothed), as well as several squeaks added for the commercial units (growls and chirps, no language)
2. Sensors consist of

   a) touch sensors on the back and sides, detecting (any) kind of touch, and
   b) a gyro, tuned so that rapid movements ‘scare’ the creature as does turning it ‘upside down’.
3. No network connection or data collection is present on the commercial units; the software is flashed onto the boards and thus immutable.
4. No movement functionalities are present, apart from the vibration.
5. The ‘AI’ of the robot is a very simple finite-state machine, representing a linear counter that changes vibration in response to previous state and outputs, by simple `points’ addition/subtraction (e.g., touching back == +1 point; toy upside down == -20 points). As a result, the empirically observed assumptions of ‘liveness’ or ‘emotions’ are brought in by emotion projection of those using the toy.

Before presenting our ERA for Purrble, it should be highlighted that some risks are equally applicable to many toys or artifacts (e.g. a choking hazard if eyes fall off). We contacted the company Sproutel, the developers of Purrble, to ask about their approach towards these hazards. They responded that Purrble meets international toy safety standards, and that they obtained certificates to confirm compliance with requirements on product safety. Therefore, such risks will not be addressed in this work and the ERA will solely focus on risks unique to *robot* toys.

4.1 Physical Risks

Injury – the battery pack of Purrble is hidden underneath a layer of fur. As this layer is relatively thin, Purrble’s underside is relatively hard. This can result in physical harm if the Purrble is dropped on a person or thrown by the child (if she is upset perhaps). This risk can be mitigated by either providing a denser layer of fur over the battery pack, or relocating the battery pack less close to the surface of the toy.

Hygiene – due to its functionality, it is likely that Purrble will be shared, for example between siblings or in a classroom. Currently, the fur of Purrble is not washable, presenting a hygiene risk[5]. To reduce this risk the Purrble should be updated such that the fur can be taken off the robot and washed.

4.2 Psychological Risks

Influence relationships between people – the use of Purrble may impact relationships between people. For example, it may disrupt home dynamics (e.g. in contexts such as trauma, fostered or adopted families), or exacerbate already existing tensions. As there are many possible scenarios for this hazard with different risk levels, a more thorough analysis on human relationships – and different circumstances in which Purrble can be used – is needed in order to develop mitigation approaches. We do not address this any further here as it goes beyond the scope of this study.

Dependency – if the child becomes too dependent on Purrble, there is a risk of increased anxiety and/or stress if Purrble is unavailable when needed. This could occur for example, if the batteries die, or another child is using Purrble, or it is lost or stolen. This can be mitigated by providing an indicator when the batteries are running low, and ensuring a child's time with Purrble is limited to prevent her becoming too dependent on it.

Overreliance – people might become too reliant on the supportive function of Purrble, using it as a 'quick fix' during situations for which it was not designed and not addressing deeper underlying issues. This risk is currently low as Purrble is still new to the market, but may increase with a larger user base. The risk can be mitigated by stressing the importance of carefully considering whether Purrble is the correct support system to use, case by case.

[5] Highlighted by the 2019-21 Covid-19 pandemic.

Nurturing Machines – children may feel compelled to look after Purrble [19]. The main function of Purrble is that it needs to be supported to calm down. The risk of feeling compelled to care has been mitigated by designing Purrble such that it is not too needy, resulting in relatively short interaction times.

Impact on interaction skills – intensive use of Purrble may lead to the child forgetting about, and interacting less with, other living creatures (e.g. classmates, siblings or pets). Replacement by Purrble may become the new norm (a phenomenon also known as environmental amnesia [20]), resulting in a decrease in development of interaction skills. This risk can be mitigated by limiting interaction time with Purrble.

Animal welfare – due to the autonomy and interactive abilities of Purrble it may be more likely to be considered a 'companion' for a child that otherwise may have been a pet. Currently, this risk is low as Purrble is still new. However, this risk highlights the advantage that Purrble can be used in situations in which pets are disallowed.

Negative symbolism – the goal of Purrble is to support emotion-regulation [17]. However, Purrble can become a symbol for mental health challenges which can lead to incorrect conclusions regarding the person's mental health, bullying or unwanted emotional conversations that people may prefer to keep private. The level of this risk will increase once people become more familiar with Purrble and can be mitigated by considering the way Purrble is introduced.

4.3 Privacy and Security Risks

Currently, Purrble does not incorporate any technologies that may raise privacy and/or security risks. It is not hackable as it is fully autonomous and does not require any connections; is it not open source nor does it store data. If any of these factors change potential privacy and security risks will need to be considered.

4.4 Environmental Risks

The environmental risks for Purrble are the same as for RoboTed (section 3.4), which means they include:

1. Unsustainability of materials,
2. Unrepairability, and
3. Unrecyclability.

Table 2: Ethical risk assessment of Purrble

Hazard	**Risk**	**Level**[6]	**Mitigation**
Physical risks			
Injury	The battery pack may injure a person if the robot is used forcefully	L	Increase the thickness of the layer of fur covering the battery pack
Hygiene	Spread virus/disease	M	Update device such that fur can be taken off and cleaned
Psychological risks			
Disturb relationships	Disrupt home dynamics, exacerbate existing tensions	L/M/H	See description
Dependency	Increased anxiety if Purrble is unavailable when needed	M	Decrease interaction time with Purrble, provide indicator for battery level
Overreliance	Purrble used as 'quick fix' instead of addressing underlying issues	L/M	Ensure Purrble is not the only option considered to address issues
Nurturing machines	The child feels responsible for Purrble	L	Ensure behaviour displayed by Purrble is not too 'needy'

[6] *Note*: Risk level: (H)igh, (M)edium or (L)ow

Impact interaction skills	Interaction skills are impacted due to less interactions with other living beings	M	Ensure interaction time with Purrble is limited and interaction with humans or animals encouraged
Animal welfare	Purrble is substituted for a pet	L	Similar to the hazard 'overreliance' Purrble should not be considered as the only option to address issues
Negative symbolism	Purrble becomes a symbol of mental health challenges, resulting in unwanted interactions	M	Consider how Purrble is advertised
Privacy and security risks: none determined for the current version of Purrble as it uses no technologies that can result in risk			
Environmental risks			
Unsustainability (of materials)	Robot uses unsustainable or high carbon cost materials	M	Replace unsustainable materials with materials from sustainable sources
Unrepairability	Robot's lifetime is limited because fault cannot be repaired or parts replaced	M	Adapt design for ease of repair with replaceable parts – exchange battery for rechargeable one
Unrecyclability	End of life robots are dumped in land fill	M	Adapt design for ease of recycling parts and materials

4.5 Discussion of ERA for Purrble

The ERA of Purrble provided several interesting insights, indicating that ERA is a useful tool for existing smart robot toys as well as the fictional RoboTed. Comparing the ERAs for Purrble and RoboTed, we see that the risks are less distributed over the four risk categories for Purrble than they are for RoboTed. This was expected, as RoboTed presents a broad range of functionalities that allowed for assessment of all four categories, where the functionality of Purrble is constrained to providing psychological support.

Purrble has been developed, and is advertised, as a research-based product. However, this ERA highlighted that there is a risk of the toy becoming a source of anxiety, flipping its original purpose. This reveals a difference between the ERAs for Purrble and RoboTed. Psychological risks for RoboTed were directly linked to the functionality of the smart robot toy, whereas some of the psychological risks for Purrble were indirect consequences of the use of a smart robot toy that provides psychological support (e.g. the concern regarding negative symbolism).

It should be noted that Purrble is still in the early stages of production. If sales take off, other aspects besides the ones presented in this ERA should be considered. For example, currently the electronics used are single use as they are cheaper to produce. If demand for the smart toy increases, the sustainability of the design should be considered, and whether the lifetime of the robots could be extended through providing options for repair.

Furthermore, even though Purrble has been developed to support children with emotion-regulation, research may indicate that the toy can be beneficial to other users. In this case, an additional ERA would be required to determine whether there are new risks that do not apply to children as a user group or have been overlooked. For example, Purrble seems to be opening up conversations that allow the parents and their child to talk about intimate personal emotional concerns. However, such conversations require different approaches depending on the user group. It is essential to carefully consider how Purrble is best introduced to specific user groups to ensure its functionality as well as the expectations from the user are understood.

5. Discussion and Conclusions

This paper has shown the value of ethical risk assessment through both a fictional and a real-robot case study. The assessments have indicated that attention to ethical risks can:

- suggest new functions,
- draw attention to potential design modifications to mitigate some risks,
- highlight the need for user engagement,
- reject product functionality as too risky, and/or
- indicate potential future issues, highlighting the need for periodic reassessments.

ERA is however, not guaranteed to expose all ethical risks. It is a subjective process which will only be successful if the risk assessment team are prepared to think both critically and creatively about the question: what could go wrong? As Vallor *et al.* [21] write, design teams must develop a "habit of exercising the skill of *moral imagination* to see how an ethical failure of the project might easily happen, and to understand the preventable causes so that they can be mitigated or avoided".

The ethical hazards and risks set out in BS8611 are an excellent starting point, but the standard does not provide an exhaustive taxonomy of ethical hazards, encompassing all domains of robotics. The RoboTed case study has identified several additional ethical hazards, some of which are specific to social robots, including the Uncanny Valley, weak security, lack of transparency (for instance the lack of data logs needed to investigate accidents), unrepairability and unrecyclability. Additionally the Purrble case study has indicated that there are psychological risks which do not follow directly from the use of the device, but indirectly, for example a disturbance of the relationship between siblings, which may lead to bullying. Such risks are not easily identified solely from the use of BS8611 and were exposed by co-authors with experience in many areas such as AI, responsible innovation, ethics, social robotics, human centred computing and psychology. Evaluating and quantifying psychological risks is especially difficult given that there are no agreed measures for hazards such as over trusting, the Uncanny Valley or when the child becomes too dependent on the smart robot toy. Assessment is further complicated by the likelihood that cultural and individual differences may lead to lower risks of psychological harm for some individuals than others. For these reasons design teams cannot rely on their own judgement and instead should engage with potential users from across the full range of age, gender and ethnic diversity, and seek guidance from psychologists and/or social scientists, both to ask the user group the right questions and interpret their responses.

Given also that ERA is not a one-time process but one that should iterate throughout a product life-cycle, good practice suggests that in-house assessments undertaken early in the design process would be shared with user groups during later iterations as the product undergoes user trials. During these iterations, the carbon footprint required for the production of the smart robot toy should be established.

Consider also what impact ERA might have on cost and user acceptance. If all mitigation strategies were applied, these might diminish the marketability of the product – which may not be a bad thing if the product is unethical (e.g. the doll 'My Friend Cayla' banned in Germany for privacy reasons [22]). However, whether companies will be able to make an adequate return on investment if they adhere to the suggested mitigations remains an open question. Note that there would be considerable value in a quality mark for responsibly designed and developed robot toys, similar to the FRR quality mark developed by the Foundation for Responsible Robotics that is currently in its pilot phase[7].

In summary, ethical risk assessment is a powerful and essential addition to the responsible roboticist's toolkit. ERA can also be thought of as the opposite face of robot accident investigation [8], seeking – at design time – to prevent risks becoming accidents.

Acknowledgments

[7] https://responsiblerobotics.org/quality-mark/

The work of this paper has been conducted within EPSRC project RoboTIPS, grant reference EP/S005099/1 *RoboTIPS: Developing Responsible Robots for the Digital Economy.* The authors are also grateful for the comments of the anonymous reviewers of ICRES 2020. The work of this extended paper was partially supported by the Assuring Autonomy International Programme, a partnership between Lloyd's Register Foundation and the University of York.

References

[1] Bernabe R, van Thiel G, Raaijmakers J, van Delden J (2012) The risk-benefit task of research ethics committees: An evaluation of current approaches and the need to incorporate decision studies methods. BMC Medical Ethics. doi: 10.1186/1472-6939-13-6

[2] Winfield AFT and Winkle K (2020) RoboTed: a case study in Ethical Risk Assessment, presented at the 5th International Conference on Robot Ethics and Standards (ICRES 2020), 28-29 September 2020. arXiv preprint: 2007.15864

[3] Winfield A (2019) Ethical standards in robotics and AI. Nature Electronics 2:46-48. doi: 10.1038/s41928-019-0213-6

[4] BSI, BS8611:2016 Robots and robotic devices, Guide to the ethical design and application of robots and robotic systems. British Standards Institute, 2016.

[5] Lasota P, Fong T, Shah J (2017) A Survey of Methods for Safe Human-Robot Interaction. Foundations and Trends in Robotics 5:261-349. doi: 10.1561/2300000052

[6] Delvaux M (2017) Report with recommendations to the commission on civil law rules on robotics. *European Parliament*, *27*.

[7] IEEE P7014, "Standard for Ethical Considerations in Emulated Empathy in Autonomous and Intelligent Systems". Retrieved from https://standards.ieee.org/project/7014.html (accessed 7-1-2021).

[8] Stilgoe J, Owen R, & Macnaghten P (2013) Developing a framework for responsible innovation. *Research policy*, *42*(9):1568-1580. doi: 10.1016/j.respol.2013.05.008

[9] Winfield A F, Winkle K, Webb H, Lyngs U, Jirotka M, & Macrae C (2020) Robot Accident Investigation: a case study in Responsible Robotics. arXiv preprint arXiv:2005.07474.

[10] IEEE, "The IEEE global initiative on ethics of autonomous and intelligent systems. Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems, first edition," tech. rep., IEEE Standards Association, 2019.

[11] Spiekermann S, Winkler T (2020) Value-based Engineering for Ethics by Design. arXiv preprint arXiv:2004.13676.

[12] Sharkey N, Sharkey A (2010) The crying shame of robot nannies. Interaction Studies 11:161-190. doi: 10.1075/is.11.2.01sha

[13] Sparrow R, Sparrow L (2006) In the hands of machines? The future of aged care. Minds and Machines 16:141-161. doi: 10.1007/s11023-006-9030-6

[14] Feil-Seifer D, Mataric M (2011) Socially Assistive Robotics. IEEE Robotics & Automation Magazine 18:24-31. doi: 10.1109/mra.2010.940150

[15] Moore R (2012) A Bayesian explanation of the 'Uncanny Valley' effect and related psychological phenomena. Nature Scientific Reports 2, 864. doi: 10.1038/srep00864

[16] Winfield AFT and Jirotka M (2017) The Case for an Ethical Black Box. In: Gao Y., Fallah S., Jin Y., Lekakou C. (eds) Towards Autonomous Robotic Systems. TAROS 2017. Lecture Notes in Computer Science, vol 10454. Springer, Cham. doi: 10.1007/978-3-319-64107-2_21

[17] Theofanopoulou N, Isbister K, Edbrooke-Childs J, Slovák P (2019) A Smart Toy Intervention to Promote Emotion Regulation in Middle Childhood: Feasibility Study. JMIR Mental Health 6:e14029. doi: 10.2196/14029

[18] Slovák P, Theofanopoulou N, Cecchet A et al. (2018) I just let him cry... Proceedings of the ACM on Human-Computer Interaction 2:1-34. doi: 10.1145/3274429

[19] Turkle S (2007) Authenticity in the age of digital companions. Interaction Studies 8:501-517. doi: 10.1075/is.8.3.11tur

[20] Kahn Jr, P H, Severson, R L, & Ruckert, J H (2009) The human relation with nature and technological nature. *Current directions in psychological science*, *18*(1):37-42. doi: 10.1111/j.1467-8721.2009.01602.x

[21] Vallor, S., Green, B., & Raicu, I. (2018). Ethics in Technology Practice. *The Markkula Center for Applied Ethics at Santa Clara University.*

[22] Madnick, S., Johnson, S., & Huang, K. (2019). What countries and companies can do when trade and cybersecurity overlap. *Harvard Business Review*, *4*.