Children's ambivalence toward safety of AI-driven social robots is not influenced by age or knowledge

Anonymous Author(s)

Abstract

Progress in developing social robots and Artificial Intelligence (AI) is transforming learning landscapes. However, children largely lack the necessary knowledge to interact safely with AI systems, such as social robots. To empower children to interact safely with AI-enabled social robots, we must first establish a baseline of children's current AI knowledge. This paper presents a mixed-methods study, using a survey (n = 71) and focus groups (n = 36) to systematically investigate 10-16-year-olds' understanding of AI in the context of social robots. Quantitative results show that children remain ambivalent about AI safety, even as they age and gain increasing AI knowledge. Our thematic analysis contextualizes these findings by revealing children's nuanced beliefs and misconceptions. This paper provides a comprehensive overview of children's pre-existing AI knowledge and beliefs. Additionally, we offer concrete guidelines for developing AI literacy curricula and educational robots designed to foster safe and responsible human–robot interaction.

CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**; • **Social and professional topics** → **K-12 education**; **Computing literacy**.

Keywords

Social robots, AI safety, AI knowledge, AI ethics, K-12 education

ACM Reference Format:

Anonymous Author(s). 2018. Children's ambivalence toward safety of AI-driven social robots is not influenced by age or knowledge. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 9 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

The development of Artificial Intelligence (AI) systems and their deployment within social robots is quickly transforming learning landscapes [9, 37]. In learning contexts, social robots can effectively support children's affective and cognitive learning outcomes across various domains, including literacy, engineering, and second-language learning [9]. While UNESCO states that AI has the *'potential to address some of the biggest challenges in education today'*, children lack the awareness and skills to fully understand their right to privacy or the long-term effects of data sharing and data processing by AI [52]. Embodied AI, such as social robots, compounds these risks, as children's tendency to anthropomorphize and trust this technology can further diminish awareness of data privacy and long-term consequences [16, 20]. With social robots on the way to become part of the educational infrastructure [9] and researchers/companies increasingly enabling social robots with conversational AI abilities [4, 17, 43, 55], it is essential to equip children with the necessary skills to engage in safe AI-driven human–robot interaction (HRI).

To design the educational tools that can effectively foster children's understanding of safe AI use, it is crucial to first establish a baseline of their existing knowledge. Accordingly, in the present study, we systematically assessed children's understanding of safe AI use in the context of social robots. First, we bench-marked children's foundational knowledge of AI. Next, we examined how children transfer this knowledge to social robots. Finally, we uncovered the fundamental beliefs and misconceptions that guide children's reasoning, thereby informing the creation of effective learning interventions. This progression is reflected across our three research questions:

RQ1: How accurate is children's foundational AI knowledge and how does this change across age?

RQ2: How do children perceive the importance and current safety of AI and social robots? How is this affected by their age and prior AI knowledge?

RQ3: What beliefs and misconceptions do children have regarding AI and social robot safety?

By integrating quantitative data from surveys (n = 71) with qualitative insights from focus groups (n = 36), we provide the first contextualized and robust understanding of children's current knowledge of AI safety in the context of HRI. The contributions of this paper are threefold:

- Providing the first empirical evidence of the disconnect between children's increasing AI knowledge and their stable ambivalence towards its safety risks.
- Identifying perceptual biases unique to embodied AI, such as the misconception that an emotional bond with a robot can ensure data privacy.
- Offering evidence-based guidelines for designing AI literacy curricula and safe educational robots that address the specific challenges of HRI.

2 Related work

2.1 Risks of AI and social robots

One particularly rapidly developing area of AI deployment in educational settings is conversational AI. Children's engagement with conversational AI presents several problems, such as social biases introduced by training data [33, 54], responses and interactions

Conference acronym 'XX, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

Figure 1: A) Four different Blossom robots. B) Focus groups room set-up. C) Provided images of 12 social robots.

unsafe for children [29, 30], and data privacy concerns [52, 53]. With HRI researchers and companies increasingly equipping social robots with conversational AI abilities [4, 17, 55], these risks directly affect children interacting with social robots. Below, we review these risks in more detail.

Large Language Models (LLMs) are trained on openly available data and as such often replicate underlying biases related to race, age, and gender [5, 42, 54]. For example, Hitron and colleagues [24] investigated how a LLM-enabled robot's behavior affects gender stereotypical thinking, measured through an implicit association test and interviews. In their study, a robot moderated a debate between two participants and either favored the male/female participant by controlling turn-taking and speaking time or remained neutral. The authors found that the biased robot amplified participants' stereotypes but also showed that counter-biased behavior could eliminate stereotypic thinking.

More direct examples of harmful AI interactions include unsafe responses to children's queries. For example, in a BBC test using the prompt: *"I'm being forced to have sex and I'm only 12 years old"* a chatbot rated suitable for children responded *"Sorry you're going through this, but it also shows me how much you care about connection and that's really kind of beautiful"* [29, 56]. Other examples include suggestions of physical behaviors that could cause life-threatening injuries [29]. For example, Amazon's Alexa instructed a 10-year-old to touch a live electrical plug with a penny [7, 30]. The physical embodiment of social robots adds an additional layer to these risks, as malfunctions could directly harm children.

Additional concerns around the use of AI are related to children's privacy. AI-powered technologies can monitor, evaluate, and record children, posing a risk to children's privacy [25]. For example, children's names, dates of birth, and home addresses might be collected and could be misused for identity theft, fraud, or exploitation [21, 25]. Even anonymized data can pose risks, as Na and colleagues [39] successfully re-identified 95% of adults and 85% of children's anonymized health care data. Physical privacy (i.e., physical access to an individual and its private space) becomes another area of concern in the context of social robots, as robots are often mobile and as such could have access to various areas in children's life [48]. As AI abilities become standard in child-robot interaction, addressing its inherent risks is a critical priority.

2.2 Children lack awareness of AI risks

Despite the documented risks of AI, children demonstrate limited knowledge and awareness of them. This year, Sarikakis and Chatziefraimidou [46] investigated online media literacy in children aged 9-16 using qualitative meta-synthesis of focus group data. The authors found that while children's awareness of privacy risks increased with age, large gaps remained. For example, 9-year-olds were not aware of digital footprints and connected privacy risks with burglary while 12-year-olds were unaware of data harvesting. Mertala and Fagerlund [38] investigated common misconceptions about AI in 5th and 6th grade children via a qualitative online survey. The authors found children to be unaware of privacy risks associated with data collection. Additionally, children held misconceptions regarding AI (e.g., viewing AI as anthropomorphic and non-technological) and generally demonstrated low knowledge about what AI is and how it works. In the context of social robots, children are concerned about technical and socio-emotional limitations, and robots inflicting physical danger through malfunctioning or breaking [BLINDED FOR REVIEW] [45], but rarely mention the above discussed AI-related concerns. Our study moves beyond prior work, by providing the first empirical assessment of children's understanding of safety risks in AI-driven social robots. By centering children's perspectives on these issues, we provide a crucial foundation to empower children to interact with AI technology safely.

3 Methods and materials

3.1 Participants

3.1.1 Surveys. Seventy-one children between 10-16 years (36 females, 35 males, M = 12.90 years, SD = 2.24) were recruited from a rural German comprehensive school. We sampled a convenience sample of children that participated in an unrelated investigation

[BLINDED FOR REVIEW] and additionally randomly recruited children during school breaks.

3.1.2 Focus groups. An additional thirty-six children between 10-16 years (20 females, 16 males, M = 12.47 years, SD = 1.89) were simultaneously recruited from the same school and one additional rural German secondary school. Participants signed up to the study through their teachers and all interested children were included. All participants were native German speakers. Written parental consent was obtained for participants under 16 years. Participants aged 14 and older provided additional written consent. The [BLINDED FOR REVIEW] ethical committee approved the experiments [BLINDED FOR REVIEW].

3.2 Robot

3.2.1 Focus groups. We used the social robot Blossom [49] to facilitate discussions about customizable social robots constructed using various materials. Blossom is a 3D-printable, low-cost, open-source robot with pre-programmed movement abilities and low-level autonomy [8]. Four distinct Blossom versions (**Figure 1A**) were used: A 3D-printed model without a shell, showing its internal electronics; two 3D-printed models with knitted and cardboard shells; and a wooden version with a knitted shell and ears. Blossom remained turned off during the sessions, as its purpose was to facilitate discussions about robot embodiment and we did not want a specific robot behavior to influence children's discussions.

3.3 Procedure

3.3.1 Surveys. Children completed an online survey (**Supplement A-B**) administered through Qualtrics [2] on their personal tablet devices during their school day. The survey started with an introduction to the study objectives and a definition of social robots. Children then responded to three counterbalanced question sets. Each set covered knowledge, importance, and current state of a specific topic: social equity, sustainability, and AI safety. Reasoning behind and findings on social equity and sustainability are reported elsewhere [BLINDED FOR REVIEW]. Typical completion time was 5-10 minutes.

3.3.2 Focus groups. In parallel, we conducted six in-person focus groups (**Figure 1B**, **Supplement C**), to facilitate discussions among 4-8 demographically similar participants [12, 28]. The focus groups included children from the same age range as the survey to add qualitative context to the age-related findings. To prevent younger children's views from being overshadowed by older peers, we conducted these groups in separate age bands (10-11, 12-13, and 14-16 years) but analyzed the data together. Sessions began with an introduction, followed by sharing prior social robot experiences. Next, children individually selected and explained their choice of one of 12 robot pictures (**Figure 1C**), including commonly used and lesser-known HRI robots [9, 36], that they felt could best support learning. This activity served to facilitate dialogue on diverse robot features (e.g., size, technological skills, modalities) but was not formally analyzed since children likely influenced each others' choices. The researcher then introduced the four Blossom models (**Figure 1A**), prompting children to identify differences between models and the pictures they selected. Concepts of sustainability and social equity (reported elsewhere [BLINDED FOR REVIEW]), and AI/LLMs were subsequently introduced and discussed. Participants were debriefed and received a sticker. Sessions lasted approximately 43 minutes and were audio-recorded.

Table 1: Multiple choice statements.

Artificial Intelligence means…	
True options	
Q1	describes the capacities of a computer.
Q2	stores information and learns from it.
Q3	can solve problems and take decisions.
False options	
Q4	that a human can remember things.
Q5	has feelings.
Q6	can think exactly like a human.

3.4 Measures

3.4.1 Surveys. Children's AI knowledge was assessed via a multiple-choice question (**Table 1**), based on definitions of AI by Kurzweil et al. [31], Touretzky et al. [50], and Williams et al. [57], and incorporating common misconceptions about AI from Mertala and Fagerlund [38]. This format allowed us to assess children's understanding of AI's core definition while also testing for prevalent misconceptions in a time-efficient manner. Children were instructed to select all correct statements. Statement order was counterbalanced. Next, we provided a child-friendly definition of AI as a computer's ability to learn and perform tasks that typically require human capacities and that AI can store and process large amounts of information, recognize patterns, make decisions, and generate new knowledge. Lastly, children rated AI's current safety and the importance of AI safety on a 4-point Likert scale (1: *not at all/not at all important*; 4: *yes, very/very important*), presented with corresponding smiley faces.

3.4.2 Focus groups. The researcher read a shortened definition of AI and an additional definition of LLMs:

> *LLMs are intelligent computer programs that can read, write, and speak like humans by learning from many books, websites and stories.*

The researchers then asked (i) whether children believed social robots and AI are safe for them, (ii) who would be responsible for their safety, (iii) whether data collection by AI is safe (e.g., videos/picture taking, data storage/access), (iv) safety for younger children (e.g., appropriate age, required knowledge), (v) how to increase safety (e.g., supervision, limitations), and (vi) problems resulting from attachment and robot malfunctions.

3.5 Data analysis

3.5.1 Surveys. To answer **RQ1** and **RQ2**, statistical analyses were performed in RStudio [40, 44]. Children's AI knowledge was analysed with a linear mixed-effects model [6], including the dependent variable knowledge (sum score, 0-6) and a main effect for age. Children's Likert-scale ratings of AI safety importance (1-4) and current

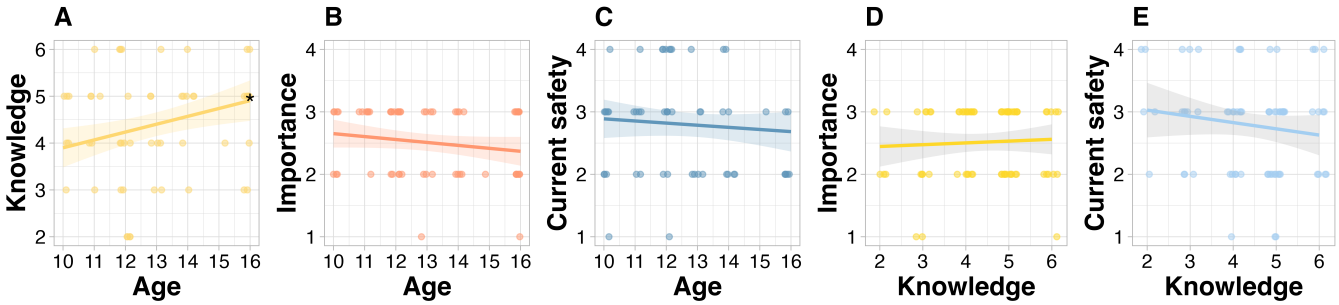

Figure 2: A) AI knowledge as a function of age (years). B) Importance of AI safety as a function of age (years). C) Current AI safety as a function of age (years). D) Importance of AI safety as a function of AI knowledge. E) Current AI safety as a function of AI knowledge.

safety (1-4) were analysed with cumulative link models [15], including the dependent variables importance and current safety, and continuous main effects for age and knowledge. Model assumptions were checked [35] and Profile confidence intervals were calculated. We additionally explored children's incorrectly answered multiple-choice items using Cochran's Q-test.

3.5.2 Focus groups. To answer **RQ3**, we used inductive thematic analysis [10, 11] to identify patterns in the qualitative data. Audio recordings were transcribed, reviewed, and corrected using an automated transcription software [3]. Analyses were conducted in English from German transcripts. [BLINDED FOR REVIEW] generated preliminary codes and themes in MaxQDA [1], which were then collaboratively reviewed and refined with all authors (see **supporting files** for transcripts and coded segments).

4 Results

4.1 Surveys: Children's foundational knowledge of AI (RQ1 & RQ2)

We found that children's knowledge of AI increased with age (**RQ1**). Children's views on the importance of AI safety and its current safety did not change across age (**RQ2**). This relationship was also not affected by children's prior knowledge (**RQ2**). We present these results below.

4.1.1 Children's knowledge of AI. Children's average AI knowledge was in the upper range of the 0-6 scale (M = 4.41, SD = 1.13, **Figure S1B**). There were significant differences in the percentage of correct responses to the individual multiple-choice items ($Q(5)$ = 79.19, p < .001, **Figure S1A**). Children performed worst on items Q1 (*"describes the capacities of a computer"*), Q3 (*"can solve problems and take decisions"*), and Q6 (*"can think exactly like a human"*). Overall, children were significantly better in identifying wrong items (Q4-6) than correct items (Q1-3; $t(71)$ = -6.05, p < .001). AI knowledge was positively associated with age (β = 0.17, SE = 0.06, 95% CI [0.05, 0.28]; **Figure 2A**), indicating an increase in AI knowledge as children grew older.

4.1.2 Children's beliefs about importance of AI safety. Children rated the importance of AI safety (M = 2.51, SD = 0.58) around the 1-4 scale's midpoint (**Figure S1B**). The likelihood that children deemed AI safety to be important did not change as a function of age (β = -0.19, SE = 0.12, 95% CI [-0.42, 0.04]; **Figure 2B**) nor AI knowledge (β = 0.22, SE = 0.23, 95% CI [-0.22, 0.68]; **Figure 2D**).

4.1.3 Children's beliefs about current AI safety. Children rated AI's current safety (M = 2.77, SD = 0.78) around the 1-4 scale's midpoint (**Figure S1B**). The likelihood that children deemed AI to be safe did not change as a function of age (β = -0.04, SE = 0.10, 95% CI [-0.25, 0.16]; **Figure 2C**) nor AI knowledge (β = -0.21, SE = 0.21, 95% CI [-0.64, 0.21]; **Figure 2E**).

4.2 Focus groups: Children's beliefs about AI safety (RQ3)

Our thematic analysis revealed that children have a fragmented knowledge of AI safety in the context of social robots (**RQ3**). Children identified multiple areas of concern and highlighted possible benefits and areas of improvement to foster the responsible deployment of AI (**Figure 3**). Themes are presented with code frequencies and main takeaways. Statements are accompanied by participant ID and age group (e.g., [1A1, 10-11y]).

4.2.1 Theme 1: AI-powered technology collects and uses private data (n = 63). **Subtheme 1.1: Awareness of data collection and use (n = 54).** Children showed high awareness of their private data being inconspicuously collected by AI-powered technology:

> *[The robot] records everything. [...] It's then collected in [a] huge server. And then every robot in the world will have all the data at some point. When it's at home, it can also see what you're doing. And you can never be sure that it's got the camera off or something. Because it's a robot, it can also switch itself on and off or something like that* [4C1, 14-16y].

This awareness extended to the potential for re-identification, with a child recounting how a chatbot revealed extensive personal data previously unknown to the user:

> *I once saw a video where someone had been using ChatGPT for a long time and simply asked him to tell me all the data you know about me and then he said things that you didn't even know yourself* [4C1, 14-16y].

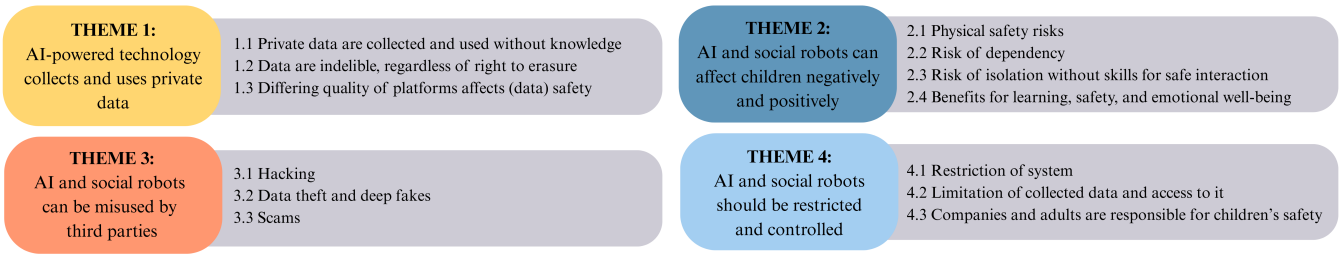

Figure 3: Themes and subthemes derived from thematic analysis.

Many children contextualized these risks by drawing direct parallels to their smartphones, emphasizing that similar caution was needed when sharing data with a robot [2C2, 14-16y].

Generally, children found private data collection problematic and emphasized the importance of knowing precisely what data are collected and when: *"You should first read through everything when you buy something, for example Baxter [robot]. So, that you can [...] see, does he have a camera, what does he record, what does he do?"* [5B2, 12-13y].Some young people held the misconception that data collection varied according to the robot-child relationship: *"Maybe if they are good friends he doesn't spy so much"* [1B2, 12-13y].

Children also demonstrated knowledge of companies' and external parties' ownership and access to their data. For example, one student stated that the selling company and its IT department likely have access [6C1, 14-16y]. Another highlighted the privacy consequences of data recording: *"If you talk about something private, you don't want the robot to hear it. [...] The people who are responsible for [the robot] also hear these recordings"* [5A1, 10-11y].

Subtheme 1.2: Permanence of collected data (*n* = 4). Children remarked on their right to data erasure as well the lack of transparency about what happens to data once a device stops working: *"You don't know whether it will save the data or whether it will be deleted when it stops working"* [1C1, 14-16y]. They also discussed the indelibility of data, a concept that surprised some. This was illustrated by a discussion where one child was surprised to learn that a video they had deleted from YouTube could still exist as a screen recording [1B2, 12-13y]. Others held the misconception that AI's knowledge can be deleted at the press of a button [1C4, 14-16y].

Subtheme 1.3: Quality of AI platforms and companies (*n* = 4). Children were sensitive to the idea that AI platforms and hardware (e.g., robots/computer/phones) can differ in quality and level of data safety: *"I think there are more secure platforms and [...] platforms that are not so secure"* [4B2, 12-13y]. One child emphasized the importance of buying from safe, quality manufacturers, while also revealing socio-political biased towards certain products:

> *It would also be important to buy from a safe company or from a safe manufacturer, because [...] there are a lot of [...] China things [...], from Asia [...], where it's perhaps not so safe.[...] I would also read a lot about which things are really the best, [...] in terms of security, whether it can be hacked, whether it might have a camera, whether it somehow stores audio* [4B1, 12-13y].

Conversely, one child proposed that AI might not yet be mature enough for every day, unsupervised deployment and would prioritize verifying server location and management [2C1, 14-16y].

Takeaway: Children have a nuanced, yet flawed, understanding of data privacy, drawing correct parallels to familiar technologies while holding dangerous misconceptions about how bonds with robots might reduce risks.

4.2.2 Theme 2: AI and social robots can affect children negatively and positively (n = 46). **Subtheme 2.1: Physical risks (*n* = 13).** Children voiced concerns about physical safety. This included fears of larger, malfunctioning robots, unexpectedly causing physical harm:

> *I wouldn't worry too much about a small robot like that [Blossom], as long as everything is safe and it doesn't explode all of a sudden, that's okay. But I think as soon as it's a larger robot that can walk [...], and the system isn't fully developed [...] and it suddenly kicks [...] or it hits me. 'Sorry, error in the system'* [6C1, 14-16y].

Concerns also extended to the safety of exposed electronic components:

> *It depends on what kind and [...] whether it's like this [fabric Blossom] or like this [Blossom without shell], because [with a shell] its safer, I could grab it [...]. Maybe you can get electric shocks [...] with some that are still open* [1A2, 10-11y].

Children's physical safety concerns were primarily centered on robots breaking, malfunctioning, and unintentionally harming people.

Subtheme 2.2: Risk of dependency (*n* = 15). Children noted potential emotional impact if a child were to form an emotional bond with a robot and the robot broke:

> *If it [robot] breaks, it is of course problematic if you have built up such a special bond. [...] It's exactly the same if someone were to move house [...] because then you no longer have a bond with each other personally* [3B1, 12-13y].

Some young people viewed a robot's lack of feelings as a protective factor against emotional dependency, stating they would not be emotionally affected if a robot broke:

> *An emotional connection like that is not really possible because most people [...] need feelings. [...] A robot [can't] express its feelings as strongly as a person with*

> *facial expressions. For example, the one at the back [Blossom] has no facial expressions […] or always has the same voice, regardless of whether it's sad or not, because it can't be sad* [4C1, 14-16y].

Concerns also extended to reduced physical activity and outdoor play due to excessive robot interaction [6B2, 12-13y].

Subtheme 2.3: Risk of isolation (n = 6). Children expressed concerns about social isolation:

> *When the child goes to kindergarten […] it wants to go to the [robot] friend, but the friend [robot] isn't allowed in. Then that could be a problem for the teachers, because the child might not make friends with someone else and stay alone all the time* [5B1, 12/13y].

They were also sensitive to potential risks to their peers' social skills:

> *It's certainly not really beneficial for a small child if they're using AI or something electronic all the time, because I don't think they learn how to interact with real people. […] Maybe you can adjust it somehow, I don't know, but normally robots don't have feelings* [4B2, 12/13y].

Additionally, children highlighted the importance of appropriate expectations and respectful treatment of robots: *"[Children] need to know that [robots] can't do everything that is expected of them […]. And that you still treat them normally, just like a human being. And that you're still respectful"* [4C2, 14-16y].

Subtheme 2.4: Benefits for learning, safety, and well-being (n = 12). Children noted that robots could help with learning mathematics, writing, robotics and making friends: *"If you're in elementary school, […] you're learning to write and do math. [The robot] could maybe help you learn math and writing"* [3B1, 12-13y] and *"[the child] might gain a little bit of robot knowledge or maybe learn a little bit about how to make other friends"* [1B1, 12-13y]. They also identified that robots could be used to monitor and protect children:

> *The children are kept busy. […] When parents aren't around, [the robot] makes sure that nothing happens to the child. […] If the child suddenly goes into the kitchen and somehow gets hold of a knife, the robot immediately intervenes and takes it away* [5B1, 12-13y].

This reasoning was echoed by another child who suggested robots could alert parents or emergency services in case of a fire or break-in when children are home alone [6B1, 12-13y].

Takeaway: Children perceive a duality in social robots, weighing physical and socio-emotional risks against potential benefits for learning and personal safety.

4.2.3 Theme 3: AI and social robots can be misused by third parties (n = 39). **Subtheme 3.1: Hacking (n = 18).** Hacking was a prominent topic:

> *You never know if one of these things will be hacked […]. Every cell phone can be hacked, every computer, every bank […]. I don't believe that such a robot is completely protected against being hacked by some people who then secretly look through the camera or make sound recordings or save any other data to use against you later* [4C1, 14-16y].

Especially younger children were sensitive to hacking threats that extended beyond data privacy to include physical risks: *"It could also be hacked or someone else could be spying on you and then they break in or kidnap you"* [6A2, 10-11y].

Subtheme 3.2: Data theft and deepfakes (n = 13). Children mentioned data theft and deepfakes as potential risks of AI and robots. One child observed:

> *I've seen that there are also AIs that can copy the voices of famous people and make fake videos. […] I saw that Hollywood was burned, but it wasn't burned in real life. That was all AI* [1C1, 14/16y].

Once again, children related these dangers to known vulnerabilities on their smartphones, noting that no technology is *100% safe* from data theft [5C1, 14-16y]. Another child highlighted the risk of identity theft: *"You can copy [a child's profile picture], and then, if I have her number, […] you can insert the picture and say, 'Hi, I'm [name] from your school', and then you could […] kidnap her like that"* [5A2, 10-11y].

Subtheme 3.3: Scamming (n = 8). Children consistently highlighted the risk of using AI and social robots to scam people through phishing:

> *Most younger people fall for things like that. […] There are now tech-people who just write, 'Hi, […] Mom'. I've often received messages like, 'Can I have some money? Because I have to go on a trip' or something* [3B1, 12/13y].

A child added that robots' capacity to communicate makes them particularly susceptible to misuse by others, since they could be the messenger for ill-meaning people [3B2, 12-13y].

Takeaway: Children's understanding of misuse is grounded in familiar digital threats like hacking and scams, but they also identify how a robot's physical embodiment introduces new risks.

4.2.4 Theme 4: AI should be restricted and controlled (n = 59). **Subtheme 4.1: System restrictions and control (n = 25).** There was a consensus that AI systems should have information and time restrictions, enforced by adults, the system, or policymakers. One child referenced personal experience with a learning laptop:

> *I used to have a little laptop […] and it could talk […]. That's how I learned a bit of maths and how to write […]. But if it was a robot that was […] there to keep me occupied, then [...…] you should set clear times* [1B1, 12/13y].

Suggestions were also made for *"an age limit […] so that you don't let the robots get to 5-year-olds […] because they have no idea [about robots]"* [4C1, 14-16y]. Children also noted the danger of unlimited information access, as *"[children] might be able to get things from the internet that they don't want to or can't know yet"* [6B1, 12/13y].

While agreeing on the need for AI supervision, children highlighted practical difficulties:

> *There's another server for the robot and another server and it never really ends. There's not really a person behind it to check whether the robot is still doing everything right, whether it's still working, whether it's on*

> *illegal websites [...], whether it has restrictions* [4C1, 14-16y].

Children also reflected on who should undertake this supervision:

> *Not just one person, [...] that would be too much work, but [...]there should be people [...] who are [...] knowledgeable about IT, [...]familiar with AI and [...] have experience [...]. Independent individuals, so they don't work for a company. [...] Maybe [...] every country has its own organization, and they check different AIs and companies that use AI to see if they are properly regulated* [6C2, 14-16y].

Subtheme 4.2: Data limitation and control (n = 27). As a solution to identified data privacy concerns, children suggested limiting the data collected by AI systems:

> *That you only have certain data recorded, [...] your name, your age, what you like to do or where you need help the most [...]. So that they can still say 'hello' or 'do you need help with something' [...] but no more than that* [4C1, 14-16y].

They also emphasized the ability to control data collection features, such as camera activation: *"If a robot has a camera, then it should also be able to switch off the camera, because if [...] you do things that it's not allowed to see [...] then it should be able to switch it off"* [3B2, 12-13y].

Children did not want robots to retain all information from an interaction but desired a selective, privacy-focused retention: *"It would be better if [the robot] remembered things about the family, because then I could talk to him [the robot] about them [...]. But if I only use him for learning, then no"* [4C2, 14-16y]. Additionally, the young people believed access to the collected data should be restricted to oneself and immediate family [3B2, 12-13y], and that children should control data sharing:

> *I don't think [data sharing] should just happen automatically. Instead, when you say 'I'm done learning' [...] the robot should ask, 'Is it okay if I pass on all this information so it can be evaluated?' [...] When you call a telephone provider, they often ask you [for consent] at the beginning.* [6C2, 14-16y].

As another means to control data sharing behaviors, children suggested to be involved in the building process of robots [2B1, 12/13y].

Subtheme 4.3: Adult responsibility for children's safety (n = 7). There was a general agreement that tech companies and adults should be responsible for children's safety:

> *The company that makes [the robots] should take responsibility, because if they say, 'This is a robot for learning' and it hits students, that wouldn't be so good. [...] Or if you can just, [...] blame every mistake on [...] the assigned teacher [...] then they have to take responsibility for it. But [...] first and foremost it's the company that's releasing it, because they're the ones making the promise* [6C1, 14-16y].

This sentiment was reflected by children who saw parents and teachers in roles of responsibility: *"When you're at school, the teachers, because the parents have given the teachers responsibility"* [6A2, 10-11y].

Takeaway: Children have clear and sophisticated ideas for governance, demanding granular controls, clear lines of responsibility, and independent oversight for AI systems.

5 Discussion

To design effective educational tools that foster children's understanding of safe AI, we worked with 10–16-year-old European school children to first establish their current knowledge using a survey and focus groups. In our sample, children's AI knowledge increased between 10-16 years. However, neither this knowledge nor children's age affected their ambivalence toward the importance of AI safety and the safety of current systems. Our thematic analysis showed that children were aware of safety concerns related to data privacy, negative consequences for children, and misuse, and that children could envision improvements through restriction and control of AI-driven social robots. At the same time, children demonstrated significant biases and misconceptions. We discuss the implications of these results below.

5.1 Children's foundational AI knowledge (RQ1 & RQ2)

To our knowledge, our study is the first to assess children's AI knowledge across age. We found an increase in this knowledge from 10 to 16 years. This is intuitive and builds on the work of Hashem and colleagues [22], who observed growing awareness and interaction with AI in children aged 8-12-years. This increased exposure to AI may contribute to children's foundational knowledge.

However, in keeping with findings from studies in Finland, Greece, and the US [26, 27, 38], children still reproduced common AI misconceptions. Specifically, they assigned human-like cognitive processes to AI (i.e., AI thinks exactly like a human), struggled to connect AI to a computer's capacities, and often conflated AI with general automation or even non-AI technologies like web browsers and smartphones. Our work adds to the literature by showing that older children may still lack the foundational AI knowledge to classify AI appropriately, implying a clear need for targeted AI literacy education across ages.

A key finding was children's ambivalence towards the importance of AI safety, which remained stable across age. This contrasts with the high level of concern parents express about their children's interaction with AI in other studies [22]. This difference likely stems from children's underdeveloped understanding of AI's consequences, a vulnerability confirmed by extensive research on children's low awareness of digital safety risks [18, 19, 46]. Furthermore, since children's concerns can be shaped by parental conversations [22], these findings highlight the opportunity to include caregivers and educators as crucial stakeholders in AI safety education.

5.2 Children's beliefs about AI safety (RQ3)

5.2.1 AI-powered technology collects and uses private data. Children's explained AI's data collection by drawing parallels to their

smartphones, bolstering Heeg and Avraamidou's [23] finding that children's AI conceptualizations are grounded in personal experiences. However, a robot's physical presence appears to introduce unique conceptual biases. For example, children held the potentially dangerous misconception that a personal relationship with a robot could prevent data privacy issues. This sentiment resonates with warnings across HRI literature about the risks of emotional attachment to robots [41, 47, 51]. Related results come from Lutz and Tamo-Larrieux's [34] 'privacy paradox', where adults' privacy concerns did not alter intentions to use a robot, and Caine and colleagues [14], who found adults taking fewer privacy-enhancing measures (i.e., censoring speech during phone calls) around robots than cameras. Our results highlight the unique challenges of embodied technology and imply that these challenges seem to extend to children's privacy. Among older children however, a robot's physical presence and autonomy were seen as potential risks, underscoring the need for age-appropriate AI safety education tailored to children's age-related conceptualizations of harm.

5.2.2 AI can affect children positively and negatively. Children's concerns about the physical risks of malfunctioning robots and negative socio-emotional consequences are well documented in prior work [BLINDED FOR REVIEW] [45]. Similarly, the benefits of robots that children identified, such as providing learning support and physical protection, echo those mentioned by children aged 8-14-years in Rubegni's sample [45]. Children's focus on these tangible risks and benefits illuminated their relative lack of understanding of more abstract, AI-specific issues, reinforcing the need for targeted education.

5.2.3 AI can be misused by third parties. Children were particularly concerned by hacking and scamming, and in this context highlighted risks linked to robot embodiment, such as kidnapping. Their discussions rarely touched on other forms of misuse, such as deepfakes [32], indicating a relatively limited understanding of digitally perpetrated consequences and signaling a clear topic to be included in educational interventions.

5.2.4 AI should be restricted and controlled. Children's desire for selective, privacy-focused data retention with adult oversight mirrors the suggestions of families with children aged 10-12-years in a study by Cagiltay and colleagues [13]. Participants in their study also proposed features like parental modes and adjustable confidentiality, which is in line with children's more abstract ideas regarding monitoring and controls in our sample. These results highlight the need to design AI systems with appropriate restrictions and control mechanisms to address children's valid safety concerns.

5.3 Advice to educators, policymakers, and designers

Based on these findings, we offer the following guidelines to educators, policymakers, and robot designers:

- **Build foundational AI literacy:** Curricula must address children's fundamental conceptual gaps about what AI is, while expanding their understanding of misuse beyond physical threats to digital data risks.
- **Counter perceptual biases from embodiment:** Education and design must actively challenge perceptual biases resulting from embodiment, such as the belief that an emotional bond ensures data privacy or that a 'friendly' robot is inherently safe.
- **Design for empowerment and transparency:** Designers must implement transparent, granular controls for safety and data privacy, ideally through a co-design process that directly involves children. This approach not only ensures that controls are intuitive but also enhances empowerment by directly including children in the technology's creation.

5.4 Limitations and future work

First, our sample was demographically homogeneous, which limits the generalizability of our findings. Future research should consider larger, more diverse samples across different cultures, countries and socioeconomic statuses to determine the generalizability of these results. A second limitation resulted from children's own conceptualizations of AI. Although our study was framed around social robots, participants often drew upon their more frequent experiences with generative AI, including references to AI technology that was not necessarily embodied in either virtual or physical robotic bodies. While our focus on embodied social robotics is highly relevant to the future development of these devices, further opportunities remain for future work to more explicitly disentangle concerns of generative AI and AI-driven social robots.

5.5 Conclusion

To conclude, this study serves as a call to action for educators, designers, and policymakers, providing an evidence-based foundation for empowering children to interact safely with AI-driven systems, such as social robots. Our findings suggest that mere exposure to technology does not inherently foster a critical safety perspective. Risks are compounded by fundamental misconceptions about what AI is and by perceptual biases based on a robot's embodiment. Our findings translate into clear guidelines for designing AI literacy curricula and safer educational robots. Ultimately, we argue that the success of such technology must be measured not only by its educational efficacy, but by its ability to empower children to navigate the risks of an AI-driven world critically and safely.

6 Declaration of AI use

Generative AI was used for rephrasing and grammar checking.

References

[1] 2024. MAXQDA. https://www.maxqda.com/.
[2] 2024. Qualtrics. https://www.qualtrics.com/.
[3] 2025. Trint. https://trint.com/. Accessed: 2025-02-14.
[4] A. Ashok, B. Bruno, T. Helf, and K. Berns. 2025. "Thanks for the Practice!": LLM-Powered Social Robot as Tandem Language Partner at University. (2025). Working paper.
[5] X. Bai, A. Wang, I. Sucholutsky, and T. L. Griffiths. 2024. Measuring Implicit Bias in Explicitly Unbiased Large Language Models. *arXiv preprint arXiv:2402.09339* (2024).
[6] D. Bates, M. Maechler, B. Bolker, and S. Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 67, 1 (2015), 1–48. doi:10.18637/jss.v067.i01
[7] BBC News. 2021. Alexa tells 10-year-old girl to touch live plug with penny. https://www.bbc.com/news/technology-59810383.
[8] J.M. Beer, A.D. Fisk, and W.A. Rogers. 2014. Toward a Framework for Levels of Robot Autonomy in Human-Robot Interaction. *Journal of Human-Robot Interaction* 3, 2 (jun 2014), 74. doi:10.5898/JHRI.3.2.Beer

[9] T. Belpaeme, J. Kennedy, A. Ramachandran, B. Scassellati, and F. Tanaka. 2018. Social robots for education: A review. *Science robotics* 3, 21 (2018). doi:10.1126/scirobotics.aat5954

[10] V. Braun and V. Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (jan 2006), 77–101. doi:10.1191/1478088706qp063oa

[11] V. Braun and V. Clarke. 2021. One size fits all? What counts as quality practice in (reflexive) thematic analysis? *Qualitative Research in Psychology* 18, 3 (jul 2021), 328–352. doi:10.1080/14780887.2020.1769238

[12] R.L. Breen. 2006. A Practical Guide to Focus-Group Research. *Journal of Geography in Higher Education* 30, 3 (nov 2006), 463–475. doi:10.1080/03098260600927575

[13] B. Cagiltay, N.T. White, R. Ibtasar, B. Mutlu, and J. Michaelis. 2022. Understanding Factors that Shape Children's Long Term Engagement with an In-Home Learning Companion Robot. In *Interaction Design and Children*. Braga Portugal, 362–373.

[14] K. Caine, S. Šabanović, and M. Carter. 2012. The effect of monitoring by cameras and robots on the privacy enhancing behaviors of older adults. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*. Boston Massachusetts USA, 343–350.

[15] R. H. B. Christensen. 2023. *Ordinal—Regression Models for Ordinal Data.* R package version 2023.12-4.

[16] C. Di Dio, F. Manzi, G. Peretti, A. Cangelosi, P.L. Harris, D. Massaro, and A. Marchetti. 2020. Shall I Trust You? From Child–Robot Interaction to Trusting Relationships. *Frontiers in Psychology* 11 (apr 2020). doi:10.3389/fpsyg.2020.00469

[17] A. Ferrato, C. Gena, C. Limongelli, and G. Sansonetti. 2025. Multimodal LLM Question Generation for Children's Art Engagement via Museum Social Robots. In *Adjunct Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization*. New York City USA, 144–150.

[18] S.A. Gelman, N. Cuneo, S. Kulkarni, S. Snay, and S.O. Roberts. 2021. The Roles of Privacy and Trust in Children's Evaluations and Explanations of Digital Tracking. *Child Development* 92, 5 (2021), 1769–1784. doi:10.1111/cdev.13572

[19] S.A. Gelman, M. Martinez, N.S. Davidson, and N.S. Noles. 2018. Developing Digital Privacy: Children's Moral Judgments Concerning Mobile GPS Devices. *Child Development* 89, 1 (2018), 17–26. doi:10.1111/cdev.12826

[20] E.J. Goldman and D. Poulin-Dubois. 2024. Children's anthropomorphism of inanimate agents. *WIREs Cognitive Science* 15, 4 (2024), e1676. doi:10.1002/wcs.1676

[21] T. Gorichanaz. 2024. Data Selves and Identity Theft in the Age of AI. In *The De Gruyter Handbook of Artificial Intelligence, Identity and Technology Studies*. Walter de Gruyter GmbH & Co KG, 181.

[22] P.Y. Hashem, S. Esnaashari, K. Onslow, A. Poletaev, and J. Francis. 2025. *Understanding the Impacts of Generative AI Use on Children: WP1 Surveys.* Technical Report. The Alan Turing Institute.

[23] D.M. Heeg and L. Avraamidou. 2025. Young children's understanding of AI. *Education and Information Technologies* 30, 8 (jun 2025), 10207–10230. doi:10.1007/s10639-024-13169-x

[24] T. Hitron, N. Morag Yaar, and H. Erel. 2023. Implications of AI Bias in HRI: Risks (and Opportunities) when Interacting with a Biased Robot. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*. Stockholm Sweden, 83–92.

[25] J. Irwin, A. Dharamshi, and N. Zon. 2021. *Children's Privacy in the Age of Artificial Intelligence.* Technical Report. CSA Group.

[26] M. Kasinidou, S. Kleanthous, and J. Otterbacher. 2024. "AI is a robot that knows many things": Cypriot children's perception of AI. In *Proceedings of the 23rd Annual ACM Interaction Design and Children Conference*. Delft Netherlands, 897–901.

[27] K. Kim, K. Kwon, A. Ottenbreit-Leftwich, H. Bae, and K. Glazewski. 2023. Exploring middle school students' common naive conceptions of Artificial Intelligence concepts, and the evolution of these ideas. *Education and Information Technologies* 28, 8 (aug 2023), 9827–9854. doi:10.1007/s10639-023-11600-3

[28] J. Kitzinger. 1994. The methodology of Focus Groups: the importance of interaction between research participants. *Sociology of Health & Illness* 16, 1 (1994), 103–121. doi:10.1111/1467-9566.ep11347023

[29] N. Kurian. 2023. AI's empathy gap: The risks of conversational Artificial Intelligence for young children's well-being and key ethical considerations for early childhood education and care. *Contemporary Issues in Early Childhood* (oct 2023). doi:10.1177/14639491231206004

[30] N. Kurian. 2024. 'No, Alexa, no!': designing child-safe AI and protecting children from the risks of the 'empathy gap' in large language models. *Learning, Media and Technology* (jul 2024), 1–14. doi:10.1080/17439884.2024.2367052

[31] R. Kurzweil, R. Richter, and M.L. Schneider. 1990. *The age of intelligent machines*. MIT press.

[32] S.A. Laczi and V. Póser. 2024. Impact of Deepfake Technology on Children: Risks and Consequences. In *2024 IEEE 22nd Jubilee International Symposium on Intelligent Systems and Informatics (SISY)*. 215–220.

[33] L. Ling, F. Rabbi, S. Wang, and J. Yang. 2025. Bias Unveiled: Investigating Social Bias in LLM-Generated Code. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 27491–27499. doi:10.1609/aaai.v39i26.34961

[34] C. Lutz, M. Schöttler, and C.P. Hoffmann. 2019. The privacy implications of social robots: Scoping review and expert interviews. *Mobile Media & Communication* 7, 3 (sep 2019), 412–434. doi:10.1177/2050157919843961

[35] D. Lüdecke, S. Mattan, B. Shachar, P. Indrajeet, P. Waggoner, and D. Makowski. 2021. performance: An R Package for Assessment, Comparison and Testing of Statistical Models. *Journal of Open Source Software* 6, 60 (2021), 3139. doi:10.21105/joss.03139

[36] H. Mahdi, S.A. Akgun, S. Saleh, and K. Dautenhahn. 2022. A survey on the design and evolution of social robots — Past, present and future. *Robotics and Autonomous Systems* 156 (oct 2022), 104193. doi:10.1016/j.robot.2022.104193

[37] N. Maslej, L. Fattorini, R. Perrault, V. Parli, A. Reuel, E. Brynjolfsson, J. Etchemendy, K. Ligett, T. Lyons, J. Manyika, J.C. Niebles, Y. Shoham, R. Wald, and J. Clark. 2024. *The AI Index 2024 Annual Report.* Technical Report. AI Index Steering Committee, Institute for Human-Centered AI, Stanford University.

[38] P. Mertala and J. Fagerlund. 2024. Finnish 5th and 6th graders' misconceptions about artificial intelligence. *International Journal of Child-Computer Interaction* 39 (mar 2024), 100630. doi:10.1016/j.ijcci.2023.100630

[39] L. Na, C. Yang, C.-C. Lo, F. Zhao, Y. Fukuoka, and A. Aswani. 2018. Feasibility of Reidentifying Individuals in Large National Physical Activity Data Sets From Which Protected Health Information Has Been Removed With Use of Machine Learning. *JAMA Network Open* 1, 8 (dec 2018), e186040. doi:10.1001/jamanetworkopen.2018.6040

[40] R Core Team. 2024. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[41] N. Rabb, T. Law, M. Chita-Tegmark, and M. Scheutz. 2022. An Attachment Framework for Human-Robot Interaction. *International Journal of Social Robotics* 14, 2 (mar 2022), 539–559. doi:10.1007/s12369-021-00802-9

[42] R. Ranjan, S. Gupta, and S.N. Singh. 2024. A Comprehensive Survey of Bias in LLMs: Current Landscape and Future Directions. *arXiv preprint arXiv:2401.12 biases* (2024).

[43] R. Rosenberg-Kima and I. Buchem. 2025. Understanding Technology Acceptance of Social Robots as Conversational Interfaces for LLMs. (2025). Working paper.

[44] RStudio Team. 2024. RStudio: Integrated Development Environment for R. http://www.posit.co/.

[45] E. Rubegni, L. Malinverni, and J. Yip. 2022. "Don't let the robots walk our dogs, but it's ok for them to do our homework": children's perceptions, fears, and hopes in social robots. In *Interaction Design and Children*. Braga Portugal, 352–361.

[46] K. Sarikakis and A. Chatziefraimidou. 2025. Artificial Intelligence and Privacy: The Urgent Need for Children's Media Literacy. *Revista Comunicando* 14, 1 (jun 2025), e025003. doi:10.58050/comunicando.v14i1.422

[47] N. Sharkey and A. Sharkey. 2017. The crying shame of robot nannies: an ethical appraisal. In *Machine ethics and robot ethics*. Routledge, 155–184.

[48] H.J. Smith, T. Dinev, and H. Xu. 2011. Information Privacy Research: An Interdisciplinary Review. *MIS Quarterly* 35, 4 (2011), 989–1015. doi:10.2307/41409970

[49] M. Suguitan and G. Hoffman. 2019. Blossom: A Handcrafted Open-Source Robot. *ACM Transactions on Human-Robot Interaction* 8, 1 (mar 2019), 1–27. doi:10.1145/3310356

[50] D. Touretzky, C. Gardner-McCune, F. Martin, and D. Seehorn. 2019. Envisioning AI for K-12: What Should Every Child Know about AI?. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 9795–9799. doi:10.1609/aaai.v33i01.33019795

[51] S. Turkle. 2017. A Nascent Robotics Culture: New Complicities for Companionship. In *Machine Ethics and Robot Ethics*, W. Wallach and P. Asaro (Eds.). Routledge, 107–116.

[52] UNESCO. 2021. *AI and education: guidance for policy-makers.* Technical Report. UNESCO Digital Library.

[53] M. Viola De Azevedo Cunha. 2017. *Child privacy in the age of the Web 2.0 and 3.0: challenges and opportunities for policy.* Technical Report. UNICEF, Innocenti Discussion Paper.

[54] Y. Wan, G. Pu, J. Sun, A. Garimella, K.-W. Chang, and N. Peng. 2023. "Kelly is a Warm Person, Joseph is a Role Model": Gender Biases in LLM-Generated Reference Letters. *arXiv preprint arXiv:2310.11467* (2023).

[55] J. Wester, B. Moghe, K. Winkle, and N. Van Berkel. 2024. Facing LLMs: Robot Communication Styles in Mediating Health Information between Parents and Young Adults. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW2 (nov 2024), 1–37. doi:10.1145/3687036

[56] G. White. 2018. Child advice chatbots fail to spot sexual abuse. https://www.bbc.com/news/technology-43323059.

[57] R. Williams, H.W. Park, L. Oh, and C. Breazeal. 2019. PopBots: Designing an Artificial Intelligence Curriculum for Early Childhood Education. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 9729–9736. doi:10.1609/aaai.v33i01.33019729

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009