An exploratory research about ethical issues on a smart toy: The Hello Barbie case study

Otávio de Paula Albuquerque[1], Marcelo Fantinato[1], Sarajane Marques Peres[1]
Patrick C. K. Hung[2]

[1]School of Arts, Sciences and Humanities – University of São Paulo
São Paulo – SP – Brazil

[2]Faculty of Business and IT – Ontario Tech University
Oshawa – Ontario – Canada

{otavioalbuquerque,m.fantinato,sarajane}@usp.br

patrick.hung@ontariotechu.ca

Abstract. *Smart toys are becoming increasingly present in children's lives, reinforcing the relevance of this market niche. Advances in user interfaces and artificial intelligence have been incorporated into smart toys to provide greater autonomy and inductive reasoning skills through machine learning. However, machine learning embedded in smart toys not only brings benefits but also potential problems of bias, possibly related to prejudice and discrimination. This work aims to explore Mattel's Hello Barbie smart toy in a case study, seeking to analyze its knowledge base and conversational functionality to identify possible ethical issues that could cause harm to children. The intention is to show unknown risks that can occur in the evolution's process of smart toys.*

1. Introduction

As part of the Internet of Things (IoT), toys have become part of the digital world with the rise of smart connected toys. Smart connected toys, or simply smart toys, are devices that comprise physical components of traditional toys connected to computer systems with online communication services [Hung et al. 2016b]. Smart toys are also expected to feature Artificial Intelligence (AI) functionalities and may also offer augmented reality experience to users [Tang and Hung 2017]. As a substantial part of human development, toys maintain a daily presence for billions of individuals of all ages. The context of smart toys is similar. The trend is for the smart toy market to grow considerably over the years, increasing its market share by almost 200% from 2018 to 2023 [Jupiter Research 2018].

Smart toys come in many shapes, including wearable gadgets and anthropomorphic ones, the latter representing one of the most popular shapes for regular toys. Examples of anthropomorphic toys are stuffed animals, dolls, and robots. In addition, smart toys have networking and reasoning capabilities, and their technology solutions can range from augmented reality to Artificial Intelligence-based conversations [Mahmoud et al. 2017, Rafferty et al. 2017]. One of the most famous smart toys is Hello Barbie, a fashion doll released in earlier 2015[1] by Mattel Inc. and ToyTalk PullString, whose main functionalities are a two-way conversation through speech recognition and a

[1]This smart toy has been discontinued by the manufacturer in 2017.

progressive learning feature which should improve the conversation experience, added to its Wireless connection's capabilities [Mattel's Hello Barbie 2017].

Smart toys have provoked both relevant debates in the press media [Mathews 2017, Harris 2015] and scientific research, mainly related to children's data privacy such as the perceptions regarding smart toys and privacy in different countries [Fantinato et al. 2017, Fantinato et al. 2018], privacy risks that the context brought to light [Hung et al. 2016a, Albuquerque et al. 2019, Fantinato et al. 2020], proposed solutions to preserving privacy [Albuquerque et al. 2020, Albuquerque et al. 2022], the issues and consequences which involve related to their complexity, and the prospect of market growth. However, an equally relevant issue, but probably more difficult to define, understand and solve, has been overlooked or even neglected: the risks of the harmful behavior that smart toys built with inductive reasoning implemented in machine learning strategies could perform. Like real-world examples of machine learning discriminatory behavior that happened recently [Angwin et al. 2016, Hern 2020], the smart toy is susceptible to making potentially discriminatory behaviors if not designed carefully. In this context, a discriminatory bias built into your dataset or knowledge base will be reflected in learned behaviors and will have an effect on human-toy interaction.

The goal of our study is to explore ethical issues regarding prejudice and discrimination in the knowledge base and conversation functionality of the smart toy through a case study of Mattel's Hello Barbie. The specific goals in this study are: (1) verify the smart toy design; (2) analyze the knowledge base used by the conversation function, looking for sensitive subjects that may represent any issue of discrimination or prejudice; and (3) compile the results of an analysis about the conversation functionality, considering its effectiveness in maintaining a proper dialogue, the possible occurrence of misunderstandings on the subject in question, and determine if this can lead to discriminatory or prejudiced responses[2].

2. Background

Children are at a sensitive stage, forming their conceptions of morality, ethics, and society; thus, for them, interactions that generate misbehavior can be extremely harmful. Whether such misbehavior is considered just misconduct or incivility in social behaviors such as insult and scurrility, or crimes such as discrimination and violence, depends on to location and culture in which it occurs.

An example of such harm is "bullying", which represents a discriminatory behavior commonly present in relationships among children or adolescents, especially at school [de Oliveira-Menegotto et al. 2013]. Bullying refers to a type of physical or psychological aggression that begins with discriminatory behavior, affecting the well-being and social functioning of children and adolescents [Nansel et al. 2001]. As the literature in the education and psychology fields shows, bullying can pose serious risks to children's physical and mental health. This behavior extends to online interactions with the practice of cyberbullying.

Discrimination concerns acting based on prejudice, resulting in unfair treatment of people for belonging to a category, regardless of individual merit [Pedreshi et al. 2008].

[2]The researchers carried out the analysis of the dialogue functionality through the interaction with the doll in the period that the service was still available.

Two types of discrimination can be considered: direct discrimination, which comprises rules that explicitly mention minority or disadvantaged groups based on sensitive discriminatory attributes related to group membership; and indirect discrimination, which includes rules that, although not explicitly mentioning discriminatory attributes, whether or not intentionally, can generate discriminatory decisions [Hajian and Domingo-Ferrer 2013].

The smart toy environment, called also as toy computing, comprises three elements: a physical toy, a mobile device, and mobile applications [Rafferty et al. 2015]. The first element has a conventional appearance but includes sensors, electronic components, and software with wireless communication capability. The second enhances the smart toy's functionalities through mobile services. The third represents applications that interact with the physical toy to process and store data [Albuquerque et al. 2022]. For the scope of our study, we are analyzing the interaction of the physical toy with a user, which is usually a child, through the mobile application represented by the speech recognition functionality (see Figure 1).

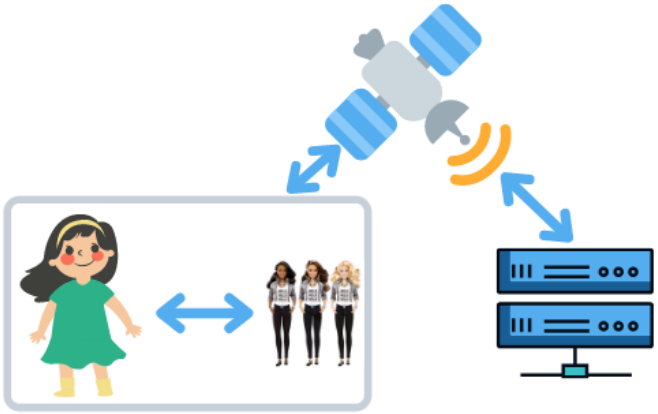

Figure 1. Smart toy environment

3. Related work

Concerns and discussions on prejudice and discrimination in machine learning have been in evidence [Bird et al. 2019, Holstein et al. 2019]. Studies have been conducted mainly regarding handling biased training datasets [Hajian and Domingo-Ferrer 2013, Hajian et al. 2016, Pedreshi et al. 2008]. Angwin et al. [Angwin et al. 2016] studied the outcomes of a biased algorithm to classify people for crime recurrence based on a biased dataset; the model classified criminals by giving a risk level to each committing a future crime. However, black criminals received a significantly higher likelihood of recurrence than white criminals, even if the crime committed by the white person was more serious.

The Natural Language Processing (NLP) filed has faced issues of stereotyping amplification, such as in gender and race, in biased machine learning models which use word embeddings as features. Bolukbasi and collaborators [Bolukbasi et al. 2016] provide a method that tries to change a biased embedding, removing the gender stereotypes and disassociating words that perpetuate the problem such as "receptionist" with "female". Another study presents a method based on the previous, but now for unbi-

ased word embeddings at multiclass level, treating stereotypes such as race and religion [Manzini et al. 2019].

Friedler and collaborators [Friedler et al. 2019] applied support vector machines, decision trees, Gaussian Naïve Bayes classifier, and logistic regression to study the bias present in the datasets. Bias discrimination has been explored under two aspects based on data mining strategies: *(i) the discrimination discovery* which comprises unveiling contexts of discriminatory practices in a dataset of historical decision records and *(ii) the discrimination prevention* that comprises inducing patterns that do not lead to discriminatory decisions even if the original training datasets are biased [Hajian and Domingo-Ferrer 2013, Ruggieri et al. 2014, Hajian et al. 2016]. Data anonymization and generalization techniques can be used also to prevent prejudice and discrimination, besides protecting privacy [Ruggieri et al. 2014, Wang et al. 2016].

No studies have yet been found associating smart toys with issues related to prejudice or discrimination, whether or not the toys are based on machine learning.

4. Method

In order to familiarize ourselves with smart toy analysis and the corresponding knowledge base, we first carried out an exploratory research. Exploratory research aims to provide greater familiarity with the problem to make it more explicit or build hypotheses [Gil 2002]. Based on the results of this initial exploration, we chose Mattel's Hello Barbie as the case study's object. The case study method seeks to know in-depth how and why a given situation occurs, which is supposed to be unique in many aspects, seeking to discover what is most essential and characteristic of it. The researchers do not intend to intervene in the object to be studied, but to reveal it as they perceive it [Gil 2002].

The exploratory research was made into three empirical analyses: the first one refers to the smart toy design, considering the physical component; the second analysis refers to the knowledge base; and the third analysis refers to the outcomes of a previous analysis in the conversational interaction capabilities. Regarding to the second analysis, we studied possible sensitive subjects, considering the common discriminatory behaviors in childhood raised by a literature review, which are related to gender, race/ethnicity and body type [Wang et al. 2010, Liu and Graves 2011, Powlishta et al. 1994], as well as words, phrases, and context used in questions and answers by the doll, aiming to discover sensitive subjects that can become discriminatory interaction. For analysis about conversational interaction capabilities, the speech recognition feature was evaluated in relation to the ease of dialog understanding, and the NLP-framework was evaluated in relation to the capability of the smart toy to process the speech and to answer accordingly, with the aim of capturing possible gaps in the dialogue functionality that could lead to misunderstandings and incoherent answers, hence, the risk of causing discriminatory behaviors.

The choice of the Hello Barbie doll as the case study's object was due to different factors: (1) its popularity, both in the toy market and scientific literature; (2) the previous studies and experiences of the researchers with the doll; (3) the availability and robustness of the knowledge base, or as Mattel calls "dialog lines" [3] (with a fixed set of 8,000 phrases

[3]Mattel's Hello Barbie dialog lines base is available at: http://hellobarbiefaq.mattel.com/wp-content/uploads/2015/11/hellobarbie-lines-v2.pdf

from many subjects); and (4) the availability of its feature documentation, presenting how the feature works, including technical issues.

Regarding the documentation specific for the two-way conversation functionality logic, the smart toy used *question-answering* datasets to build its machine learning models, besides their own knowledge base of questions and answers. Pullstring's ToyTalk service, used by Hello Barbie on its dialog feature, applied machine learning sequence-to-sequence (*seq2seq*) models [Sutskever et al. 2014], commonly used for speech recognition tasks. This type of toy used sample phrases from *question-answering* datasets to establish a dialogue with the child. This type of dataset can be seen as mapping a sequence of words representing the question to a sequence of words representing the answer.

5. Results and discussion

Mattel's Hello Barbie is a smart toy that was intended for children and teenagers (6 and 15 years old people), entitled to itself as the first fashion doll that could provide a two-way conversation with the user. Embedded with a microphone, speaker, and Wi-Fi capabilities, the speech recognition feature was activated by a push-and-hold on the belt buckle. The doll was available in three different skin tones, trying to bring an ethnically diverse doll. However, the dolls presented the same stereotypical appearance, such as body shape and hairstyle, which actually showed a shallow diversity, contrary to what Mattel itself already did with other lines of Barbie dolls in inclusion and diversity projects.

The content of the knowledge base was made internally by hand, specifically with children's usage in mind, trying to limit the subject's variety, and not including content coming from open web search [Mattel's Hello Barbie 2017]. Hello Barbie could talk about different topics of interest including but not limited to fashion, school, family, friends, holidays, and animals. These topics had the aim of creating an integrated conversation with the user.

Some subjects could be considered sensitive, given the indirect consideration of "race" when talking about haircut style embedded in the fashion topic, "sexual orientation" when talking about family, and "religious beliefs" when talking about holidays topic. Thus, this type of issue needs to be treated carefully in technology that has the child as the main user, such as smart toys. To mitigate issues, the Hello Barbie doll's knowledge base document presented different answers to the same subject questions, including the sensitive ones, in an attempt to diversify and adequate the dialog to different scenarios (See in the Table 1).

In the conversation interaction analysis, we verified Hello Barbie's two-way conversation feature gave the user a good range of freedom to start and conduct the dialog, but the doll could conduct the conversation if the child asked for it. However, the doll's conversation followed a question-answer tree when necessary, which was designed to redirect inappropriate conversations [Mattel's Hello Barbie 2017]. It was programmed not to repeat the most popular curse words (called also as swear words) or give answers to questions that involve words such as blood, death, and violence. The mechanism was designed to respond by asking another question and starting a new subject.

Mattel's Hello Barbie progressive learning feature meant the doll could tailor responses based upon the user's play history with the doll, creating the ability to talk about

Table 1. Questions and respective answers options about hairstyle

Questions	Answers
[1] Hi! I was just playing around with different hairstyles. Can you help me pick out the perfect one for today?	[1] A ponytail! Classic and chic!
	[2] Adorable! I love pigtails. They're just so playful!
	[3] Oooh updos are so elegant!
	[4] Braids are so beautiful! And there are so many different kinds it's fun learning. how to do them!
[2] I always love your look. How do you think I should wear my hair today?	[5] I love retro hairstyles! It's so fun to imagine what it would be like to live in a different era.
	[6] What a great idea! I think curly hair is so beautiful.
	[7] The higher the hair, the happier I feel!
	[8] Dreadlocks are beautiful!
	[9] Awesome! Short hair is super stylish.

some of the user's favorite subjects and timeline, remembering what was talked about in the past [Mattel's Hello Barbie 2017]. This feature had a relevant role in the appropriate dialog build, forcing the right scenario on a question-answer dialogue, and avoiding misleading answers and possible discriminatory responses. An identified problem involves the speech technology behind the dialog functionality. The doll had difficulty recognizing the accents of non-native English language speakers, which could cause incompatible answers.

The doll was targeted for children and teenagers (6 and 15 years old people), covering more ages than the common age group of smart toys, which is 5 to 11 years old. This could explain the freedom of usage, intended to create a dialogue closer to the human. The question-answer tree proved to be a good feature to help to prevent some prejudiced attacks, in view as discriminatory content such as racist words being treated as bad language. This mechanism was important to avoid the doll from learning how to misbehave through the progressive learning feature.

6. Final remarks

This research is an ongoing study, in which the analysis of the doll design has already been completed, while the knowledge base analysis has been partially carried out, covering only the issues related to ethnicity. In the next steps, the knowledge base analysis will be refined and expanded for covering matters related to religion, sexual orientation, and other sensitive topics. As future works, we aim to explore other smart toys and companion robots which have children as main users and use a children-specific knowledge base. The idea will be to continue analyzing knowledge bases construction and NLP features to explore the risk of prejudiced and discriminatory behaviors in the interaction with kids.

Acknowledgments

The authors of this work would like to thank the Center for Artificial Intelligence (C4AI-USP) and the support from the São Paulo Research Foundation (FAPESP grant #2019/07665-4) and from the IBM Corporation. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Fi-

nance Code 001 and the Conselho Nacional de Desenvolvimento Científico e Tecnológico - Brasil (CNPq grant #312630/2021-2).

References

Albuquerque, O. d. P., Fantinato, M., Eler, M. M., Peres, S. M., and Hung, P. C. K. (2020). A study of parental control requirements for smart toys. In *IEEE International Conference on Systems, Man, and Cybernetics*. *submitted*.

Albuquerque, O. d. P., Fantinato, M., Hung, P. C., Peres, S. M., Iqbal, F., Rehman, U., and Shah, M. U. (2022). Recommendations for a smart toy parental control tool. *The Journal of Supercomputing*, pages 1–39.

Albuquerque, O. d. P., Fantinato, M., Kelner, J., and Albuquerque, A. P. (2019). Privacy in smart toys: Risks and proposed solutions. *Electronic Commerce Research and Applications*, 39:1–15.

Angwin, J., Larson, J., M. S., and Kirchner, L. (2016). Machine bias. http://tiny.cc/sazdaz. ProPublica.

Bird, S., Kenthapadi, K., Kiciman, E., and Mitchell, M. (2019). Fairness-aware machine learning: Practical challenges and lessons learned. In *12th International Conference on Web Search and Data Mining*, pages 834–835.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

de Oliveira-Menegotto, L. M., Pasini, A. I., and Levandowski, G. (2013). O bullying escolar no brasil: uma revisão de artigos científicos. *Psicologia: Teoria e Prática*, 15(2):203–215. *(in Portuguese)*.

Fantinato, M., Albuquerque, O. D. P., De Albuquerque, A. P., Kelner, J., and Yankson, B. (2020). A literature survey on smart toy-related children's privacy risks. In *53rd Hawaii International Conference on System Sciences*.

Fantinato, M., Hung, P. C., Jiang, Y., Roa, J., Villarreal, P., Melaisi, M., and Amancio, F. (2017). A survey on purchase intention of hello barbie in brazil and argentina. In *Computing in Smart Toys*, pages 21–34. Springer.

Fantinato, M., Hung, P. C. K., Jiang, Y., Roa, J., Villarreal, P., Melaisi, M., and Amancio, F. (2018). A preliminary study of Hello Barbie in Brazil and Argentina. *Sustainable Cities and Society*, 40:83–90.

Friedler, S. A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E. P., and Roth, D. (2019). A comparative study of fairness-enhancing interventions in machine learning. In *Conference on Fairness, Account., and Transp.*, pages 329–338.

Gil, A. C. (2002). Como elaborar projetos de pesquisa. *São Paulo*, 5(61):16–17. *(in Portuguese)*.

Hajian, S., Bonchi, F., and Castillo, C. (2016). Algorithmic bias: From discrimination discovery to fairness-aware data mining. In *International Conference on Knowl. Disc. and Data Mining*, volume 13-17-Aug, pages 2125–2126.

Hajian, S. and Domingo-Ferrer, J. (2013). A methodology for direct and indirect discrimination prevention in data mining. *IEEE Transactions on Knowledge and Data Engineering*, 25(7):1445–1459.

Harris, S. (2015). 'Hell No Barbie' campaign targets Hello Barbie over privacy concerns. http://tiny.cc/pbzdaz. CBC.

Hern, A. (2020). Twitter apologises for "racist" image-cropping algorithm. https://www.theguardian.com/technology/2020/sep/21/twitter-apologises-for-racistimage-cropping-algorithm.

Holstein, K., Wortman Vaughan, J., Daumé III, H., Dudik, M., and Wallach, H. (2019). Improving fairness in machine learning systems: What do industry practitioners need? In *Conference on Hum. Fact. in Comp. Sys.*, page 600.

Hung, P. C. K., Fantinato, M., and Rafferty, L. (2016a). A study of privacy requirements for smart toys. In *20th Pacific Asia Conference on Information Systems*, pages 1–7.

Hung, P. C. K., Iqbal, F., Huang, S.-C., Melaisi, M., and Pang, K. (2016b). A glance of child's play privacy in smart toys. In *2nd International Conference on Cloud Computing and Security*, pages 217–231.

Jupiter Research (2018). Smart toy revenues to grow by almost 200% from 2018 to $18 billion by 2023. https://www.juniperresearch.com/press/smart-toy-revenues-grow-almost-200pc-by-2023.

Liu, J. and Graves, N. (2011). Childhood bullying: A review of constructs, concepts, and nursing implications. *Public Health Nursing*, 28(6):556–568.

Mahmoud, M., Hossen, M. Z., Barakat, H., Mannan, M., and Youssef, A. (2017). Towards a comprehensive analytical framework for smart toy privacy practices. In *7th Workshop on Socio-Technical Aspects in Security and Trust*, pages 64–75.

Manzini, T., Lim, Y. C., Tsvetkov, Y., and Black, A. W. (2019). Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. *arXiv preprint arXiv:1904.04047*.

Mathews, L. (2017). The latest privacy nightmare for parents: Data leaks from smart toys. http://bit.do/e3cCY. Forbes.

Mattel's Hello Barbie (2017). Frequently asked questions (FAQ). http://hellobarbiefaq.mattel.com/.

Nansel, T. R., Overpeck, M., Pilla, R. S., Ruan, W. J., Simons-Morton, B., and Scheidt, P. (2001). Bullying behaviors among us youth: Prevalence and association with psychosocial adjustment. *JAMA*, 285(16):2094–2100.

Pedreshi, D., Ruggieri, S., and Turini, F. (2008). Discrimination-aware data mining. In *International Conference on Knowledge Discovery and Data Mining*, pages 560–568.

Powlishta, K. K., Serbin, L. A., Doyle, A.-B., and White, D. R. (1994). Gender, ethnic, and body type biases: The generality of prejudice in childhood. *Developmental Psychology*, 30(4):526.

Rafferty, L., Fantinato, M., and Hung, P. C. K. (2015). Privacy requirements in toy computing. In Hung, P., editor, *Mobile Services for Toy Computing*, pages 141–173. Springer.

Rafferty, L., Hung, P., Fantinato, M., Peres, S. M., Iqbal, F., Kuo, S., and Huang, S. (2017). Towards a privacy rule conceptual model for smart toys. In *50th Hawaii International Conference on System Sciences*, pages 1–10.

Ruggieri, S., Hajian, S., Kamiran, F., and Zhang, X. (2014). Anti-discrimination analysis using privacy attack strategies. In *Joint Eur. Conference on Mach. Learn. and Knowl. Disc. in Datab.*, pages 694–710.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc.

Tang, J. K. and Hung, P. C. (2017). *Computing in Smart Toys*. Springer.

Wang, J., Iannotti, R. J., Luk, J. W., and Nansel, T. R. (2010). Co-occurrence of victimization from five subtypes of bullying: Physical, verbal, social exclusion, spreading rumors, and cyber. *Journal of Pediatric Psychology*, 35(10):1103–1112.

Wang, K., Wang, P., Fu, A. W., and Wong, R. C.-W. (2016). Generalized bucketization scheme for flexible privacy settings. *Inform. Sci.*, 348:377 – 393.