“Never complain, never explain”: why robots may not have to be explicable after all

Gabriele Trovato, *Member, IEEE,* Yueh-Hsuan Weng, *Member, IEEE,* Yegang Du

***Abstract*— Explainability in AI and robotics is at the centre of interdisciplinary discussions, which involve machine learning, computer science, ethics, and more. Transparency and explainability are generally seen as a need for systems that encompass AI. We offer a different point of view, arguing for re-thinking the necessity of explainability, taking inspiration by how AI is conceived in videogames user-centred design and proposing existing applications in social robotics where explainability is not an advantage.**

I. Introduction

"Never complain, never explain" is a motto attributed to British politician and statesman, Benjamin Disraeli, and originating in 1903. It has been associated with the British royal family, particularly with Queen Elizabeth II and her family, the Windsors. Citing T. Borman: "According to Proverbs: 'The heart of Kings is unknowable' This is particularly true of Elizabeth II, who throughout her long reign has played her cards very close to her chest” [1]. The idea behind the motto is that members of the royal family should maintain a dignified and reserved public image, avoiding any complaints or explanations that may be perceived as unbecoming or undignified.

It is somehow understandable that “Never complain” could be a desirable trait for a robot. The etymology of the word robot in fact comes from the Czech *robota*, meaning hard labour, and worldwide research on robots believes that they can be tools that will support human society – doing so without complaining. While a “dignified” and “reserved” robot servant may be desirable, the “Never explain” part of the motto, on the other hand, may appear counterintuitive. In this paper we argue that this motto can be extended to social robots, under certain situations. The paper is structured as follows: in Section II we overview some known arguments for explainability; in Section III a different point of view is provided, and Section IV contains some examples of situations in social robotics. Section V concludes the paper.

II. The arguments for explainability

Transparency and explainability (compared to explicability, a more common word used in the AI/robotics community) are related terms that are sometimes used interchangeably, but they refer to different aspects of a system or process. Transparency refers to the degree to which information about a system or process is accessible and visible to stakeholders, such as users, regulators, or the general public. A transparent system or process is one in which information is available to all stakeholders, and there are no hidden or opaque aspects that could affect the outcomes of the system or process. For example, a transparent decision-making process would provide clear and complete information about how decisions are made, who is involved, and what factors are considered [2]. Even though the information may be shown, it may not always be understandable. “Interpretability” is what is missing in this case. Explainability, on the other hand, refers to the ability to explain how a system or process works, and why it produces the outcomes that it does. In the particular nuance referring to natural language, an explicable system or process is one that can provide clear and understandable explanations to stakeholders about why certain decisions were made or why certain outcomes were produced. An AI model in this case would be able to provide clear and detailed explanations about how it arrived at a particular decision or recommendation [3]. However, whether explainable black boxes are preferable to inherently interpretable models is a topic of debate [4]. While making information visible, making that information understandable, and being able to explain it are slightly different concepts, in this paper span among them.

Explainability in computer science has seen significant advancements in recent years. With the increasing complexity of machine learning models and their impact on decision-making processes, there has been a growing demand for models that can provide clear and understandable explanations for their outputs. One approach that has gained popularity is the use of model-agnostic methods to generate explanations. These methods can be applied to a wide range of machine learning models and provide insights into how these models arrive at their decisions. Examples of such methods include LIME (Local Interpretable Model-agnostic Explanations) [5] and SHAP (SHapley Additive exPlanations) [6]. Explainable AI models that are specifically designed to provide interpretable outputs also exist. These models typically use simplified, rule-based approaches that are easier to understand and interpret than complex neural networks. Examples of such models include decision trees, rule-based systems, and Bayesian networks. There has also been a growing emphasis on the evaluation and validation of explanations generated by machine learning models. Researchers have proposed several metrics for evaluating the quality and usefulness of explanations, such as the degree of fidelity to the underlying

*This work was supported by Ministry of Internal Affairs and Communications (MIC) of Japan (Grant no. JPJ000595).

G. Trovato is with the Innovative Global Program, Shibaura Institute of Technology, Tokyo 169-8050, Japan (+81 (0)3-3203-6449; e-mail: gabu@shibaura-it.ac.jp).

Y. H. Weng is with Frontier Research Institute for Interdisciplinary Sciences, Tohoku University, Sendai, Japan.

Y. Du is with Future Robotics Organization, Waseda University, Tokyo, Japan, and Visiting Researcher in Shibaura Institute of Technology, Tokyo, Japan.

model, the comprehensibility of the explanation, and the degree of relevance to the user's decision-making needs [7].

Within robotics, according to A. Winfield [8], not only explicability but also transparency is a crucial element in ensuring that users can comprehend how a robot might behave in varying situations. For instance, an elderly person may be uncertain about interacting with robots. Thus, it is essential that their assisted living robot is supportive, predictable, and does not engage in any activities that may confuse or alarm them. Moreover, the robot must be safe to use. The elderly user should find it effortless to learn about the robot's functions and its rationale behind specific actions in diverse circumstances, enabling them to develop a cognitive model of their robot. A possible solution could be the robot's ability to provide natural language explanations when asked questions such as "Robot, why did you just do that?" or "Robot, what would you do if I fell down?". Winfield went as far as proposing an "ethical black box" (intended as "flight data recorder") to explain robot behaviours in case of incident [9].

III. Parallelism with Videogames

In this section, we would like to introduce a parallelism that is relevant to the topic of explainability in social robots. For understanding user-centred robotics in these regards, we can broaden the view, seeing how automated behaviours are programmed in user-centred games. Videogames industry is, in fact, in a more mature stage compared to robotics industry, and lessons can be learnt from it.

Games typically divide in to two types: symmetric and asymmetric [10]. Belonging to the first group are competitive games like chess or draughts, where the same rules apply to both the human player and the computer player. Asymmetric games instead typically consist in an environment populated by agents, in which the human player digs into. In the former case, transparency and explainability are both out of question: it would mean for the computer player to play with its cards uncovered, where victory instead is the aim (excluding the case of tutorial). Complex algorithms and decision-making techniques are then supposed to be difficult for human players to understand or guess. This can make the game more challenging and interesting. The latter case is more interesting to examine in terms of how non-playing characters (NPCs) are created. Their scripted behaviours are typically based on some sort of FSM (Finite State Machine), and their role is to take part in the enjoyment of the human player. This role may go as far as "playing to lose" [10]. Players may find it more interesting and challenging to play against opponents that are unpredictable and behave in a seemingly "intelligent" manner, keeping the "suspension of disbelief". Even though machine learning in its strict sense is used sparingly in videogames for its lack of predictability [11], AI in general is a sensitive issue in game design. This is particularly true for single-player strategy games, which stand in the middle between symmetric and asymmetric, presenting the criticalities of having to provide a credible challenge to the human player while "entertaining". Here, completely hiding the reasoning and the mechanics is also used to create unfair advantages, or to hide flaws or biases in the game mechanics.

Figure 1. Diplomacy screen of Sid Meier's Civilization V, showing revealed modifiers

One example is the following. During the development of the best-seller strategy game Sid Meier's Civilization V, the lead designer Jon Shafer stated in interviews [12]: "Our goal was to make diplomacy feel more like interacting with other players or world leaders, rather than a system to be min-maxed. No longer are diplomatic modifiers shown since this used to give away pretty much everything your computer-controlled rival nations were thinking. […] Showing the numbers would either give everything away." and "We want there to be a sense of mystery to it, where the player doesn't know exactly what to expect from the other players." Eventually, some kind of modifiers were revealed again in later patches (Figure 1), however it is worth noticing how the design decision aimed for concealing the core of the calculations for the first release.

A more extreme case is the mod for Civilization called Rhye's and Fall of Civilization [14], which explored possibilities related to Procedural Content Generation (PCG). In [15] it is shown how keeping some mechanisms concealed to the human player is critical to perform some "invasive" interventions to the game environment, which go as far as changing the map board. These interventions go unnoticed to the human eye, and are made for the sole purpose of enhancing the user's experience.

IV. Never Explain?

A. *Human-Robot Interaction (HRI)*

Since robots, unlike game agents, exist in the real world, an upstream requirement will always be the safety for the user, and its related legislation, regardless of explainability. Taken this as granted, if we think of a robotic social agent as an NPC, many potential correspondences could be found. First however, it is necessary to make a distinction, if the intended transparency/explainability is to the developer or to the user. For a developer, having control of the system that is being created is certainly needed. There are some exceptions though, as the performance of AI systems has to come to a trade-off with explainability. This is the case of complex systems such as recognition algorithms for self-driving cars, but may as well be applied to robots. Security regarding hacking of the robot in another important concern against transparency first of all. Our discussion, though, is mainly focused on the user side.

As “opacity” can help managing biases and flaws in game design, it is also widely known how the Wizard-of-Oz technique [16] helps dealing with flaws in HRI. Children [17][18] and older adults [19] are common targets of social robotic applications, and it is where the Wizard-of-Oz can be more successful. However, for all kinds of users, other aspects of interaction may be critical. One is emotional dialogue, and all the aspects of affective computing. Here we find again the concept of suspension of disbelief, a term used to describe the act of temporarily setting aside one's doubts or scepticism about the implausibility of a fictional story or scenario in order to fully engage with and enjoy it. This concept is often used in literature, theatre, films, and other forms of storytelling to create a sense of immersion and believability for the audience or readers. The idea of "suspension of disbelief" in social robotics is based on the notion that a social robot is perceived as a genuine social entity only if it is assumed to have an internal purpose and intention behind its actions. Without this assumption, the robot would be seen as a mere collection of movements rather than a social being [16]. Suspension of disbelief is, in fact, also applicable in the context of video games. In fact, it is a crucial element in creating an immersive gaming experience. When players are able to suspend their disbelief, they can fully engage with the game world and characters, becoming emotionally invested in the story and gameplay. This can enhance the enjoyment and sense of accomplishment that players experience.

B. Theomorphic robots

The case of theomorphic robots is interesting because it brings the concept of inexplainability to the extreme. Theomorphic robots are defined are robots that are related, in the appearance and in the behaviour, to the sphere of the divine [20][21]. SanTO [22], a Catholic robot, has been shown to be very sensitive to the context, with its interaction experiments turning from relatively successful [23] (Figure 2) to poor [24] in terms of accomplished dialogues and perceived sacredness, depending on the context. Maintaining the “enchantment of technology” [25] and thus the suspension of disbelief seem to be critical for the robot’s credibility, and all efforts in the design should directed to conceal the machine-like element [26], which is not only visual, but also includes total obscurity of the algorithms that drive its responses.

C. Practical application in social care

Theomorphic robots may also cross with social care, this is the case of DarumaTO [27], employed in e-ViTA [28], a Horizon 2020 EU-Japan project which aims at creating a “virtual coach” to support healthy living of adults in the age range of 65–75. The framework of e-ViTA roughly contains a front-end device, typically a social robot, a network of sensors, a dialogue system, and a middleware. Virtual coaches are defined by Siewiorek et al. as personalised systems that continuously monitor the activities and environment of users [29]. They detect situations where an intervention would be desirable and propose such interventions. To accomplish this, coaches utilise activity sensors combined with a coaching application, located either on the internet, smartphones, sensors, or social assistive robots [30].

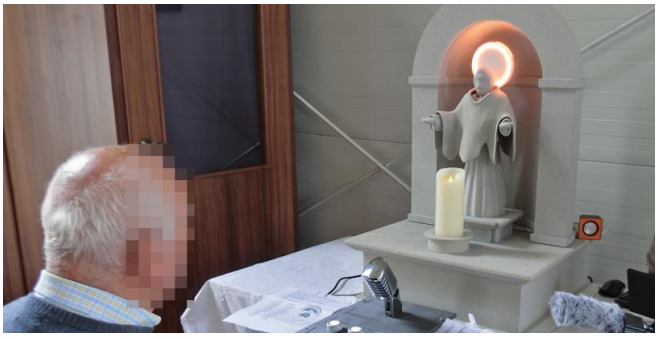

Figure 2. A Catholic user in front of SanTO-PL in a church

During the preliminary experiments in e-ViTA (Deliverable 4.4), it was seen that exposing sensors information, the framework, or the coaching application, may have a negative impact on the acceptance.Indeed, too much information can overwhelm elderly individuals when they interact with social robots, leading to confusion and frustration, and possibly hindering the acceptance of the whole system from the start, even if the whole robotic system was designed for them. Elderly individuals may have difficulties with memory and attention, making it harder for them to process and retain large amounts of information presented to them, especially if it is new.

One of the front-end devices used in these interactions was an android (Figure 3), whose results are reported in [31]. The extreme human-likeness brings an additional problem regarding explainability. All robot’s actions, in order to be appropriate to the external appearance, should be performed in a natural, human-like way. Explaining algorithms may be even more challenging to do in a natural language. The risk of an android with exposed AI is to cause a mismatch of cues, which is a prerequisite for being perceived uncanny.

D. “White Lie” Deception in HRI

Social care in robotics necessarily involves ethics. From the Kantian point of view, cheating shall be always avoided in human-robot interactions [32]. A question from a specific situation is whether healthcare robots shall be explicable when they were giving a “white lie” to patients. Due to the purpose of enhancing patients’ self-efficacy, adaptable AI-enabled healthcare robots from the Japanese Moonshot project can not disclose patients’ real health situations to them at the time they received support from robots [33].

E. Non-Spatial Proximity in HRI

Further aspects of HRI are worth mentioning. The concept of proximity, for example, which refers to “being close to something measured on certain dimension” [34]. In HRI, there were discussions on spatial proximity for robot navigation [35], verbal communication [36], nonverbal cues [37], etc. These spatial interactions often ask for high transparency on robots’ actions to allow human counterparts to take corresponding moves. However, it will be difficult and unnecessary to decide the criteria for robot transparency in matters of non-spatial proximity in HRI due to the incapability to standardise people’s virtual comfort zones with robots from different cultural and educational backgrounds, religious beliefs, social status [38][39].

Figure 3. Conversation with an android within the project e-ViTA

V. Conclusion

In these paper we argued against the explainability of social robots at least in some circumstances, and drew in some cases, such as theomorphic robots and socially assistive robots, from the point of view of the user. Taking inspiration by how AI agents are made in single-player videogames, we argue that a similar approach could be taken. As AI and robotic technologies develop, the concerns on AI may change; nevertheless, some design concepts will stay, as it is possible to witness from the evolution of videogaming industry.

References

[1] T. Borman, Crown & Sceptre: A New History of the British Monarchy. Atlantic Monthly Press, 2022

[2] Datta, Anupam, Shayak Sen, and Yair Zick. "Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems." 2016 IEEE symposium on security and privacy (SP). 2016

[3] Doshi-Velez, Finale, and Been Kim. "Towards a rigorous science of interpretable machine learning." preprint arXiv:1702.08608 (2017)

[4] Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 1, 206–215 (2019). https://doi.org/10.1038/s42256-019-0048-x

[5] Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "" Why should i trust you?" Explaining the predictions of any classifier." Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016

[6] Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." Advances in neural information processing systems 30 (2017).

[7] Doshi-Velez, Finale, and Been Kim. "Towards a rigorous science of interpretable machine learning." preprint arXiv:1702.08608 (2017)

[8] A. Winfield, Experiments in Artificial Theory of Mind: From Safety to Story-Telling. Front. Robot. AI 5:75, 2018. doi: 10.3389/frobt.2018.00075

[9] Winfield, A. F., van Maris, A., Salvini, P., & Jirotka, M. (2022). An Ethical Black Box for Social Robots: a draft Open Standard. arXiv preprint arXiv:2205.06564.

[10] S. Johnson, "Playing to Lose: AI and Civilization", GDC 2008 https://www.designer-notes.com/playing-to-lose-ai-and-civilization-gdc-2008/

[11] P. H. M. Spronck, Adaptive game AI. [Doctoral Thesis, Maastricht University], 2005. Datawyse / Universitaire Pers Maastricht. https://doi.org/10.26481/dis.20050520ps

[12] https://web.archive.org/web/20120119143839/http://e3.gamespot.com/story/6265330/civilization-v-qanda-first-e3-details?tag=topslot%3Bthumb%3B1&page=2 Last accessed 12 January 2019

[13] https://www.ign.com/articles/2010/06/15/all-about-civilization-v Last accessed 4 April 2023

[14] https://en.wikipedia.org/wiki/Rhye%27s_and_Fall_of_Civilization

[15] G. Trovato, S. Johnson, and P. Spronck: "Procedurally Generated History: building a game ecosystem through autoplay", Foundation of Digital Games, Pacific Grove, USA, June 2015

[16] P. Saulnier, E. Sharlin, and S. Greenberg, Exploring interruption in HRI using Wizard of Oz, in the 5th ACM/IEEE International Conference on Human-Robot Interaction (2010) 125–126

[17] Belpaeme, T., Baxter, P., De Greeff, J., Kennedy, J., Read, R., Looije, R.& Zelati, M. C. (2013). Child-robot interaction: Perspectives and challenges. In Social Robotics: 5th International Conference, ICSR 2013, Bristol, UK, October 27-29, 2013

[18] D. Tozadore et al.: "Wizard of Oz vs Autonomous: children's perception changes according to robot's operation conditionRO-MAN 2017, Lisbon, Portugal, August 2017.

[19] Getson, C., Nejat, G.: Socially assistive robots helping older adults through the pandemic and life after COVID-19. Robotics 10(3), 106 (2021)

[20] G. Trovato, F. Cuellar, and M. Nishimura: "Introducing 'theomorphic robots'", 2016 IEEE-RAS International Conference on Humanoid Robots, Cancún, Mexico, November 2016.

[21] G. Trovato et al: "Religion and robots: towards the synthesis of two extremes", International Journal of Social Robotics, p. 1-18, May 2019

[22] G. Trovato et al: "The creation of SanTO: a robot with "divine" features", 15th International Conference on Ubiquitous Robots, Honolulu, USA, 2018.

[23] G. Trovato et al: "Communicating with SanTO – the first Catholic robot", RO-MAN 2019, New Delhi, India, October 2019.

[24] G. Trovato and Y.-H. Weng: "Retrospective Insights on the Impacts of the Catholic Robot SanTO", Robophilosophy 2022, Helsinki, Finland, August 2022.

[25] Gell A (1994) The technology of enchantment and the enchantment of technology. In: Coote J (ed) Anthropology, art, and aesthetics. Clarendon Press, Oxford

[26] G. Trovato, C. Lucho, A. Huerta-Mercado and F. Cuellar: "Design strategies for representing the divine in robots", HRI 2018, Chicago, USA, March 2018.

[27] Z. Shen et al: "Participatory Design and Early Deployment of DarumaTO-3 Social Robot", International Conference on Social Robotics, Florence, Italy, December 2022.

[28] Jokinen, K., Homma, K., Matsumoto, Y., Fukuda, K.: Integration and interaction of trustworthy AI in a virtual coach–an overview of EU-Japan collaboration on eldercare. Advances in Intelligent Systems and Computing, vol. 1423, pp. 190–200. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-96451-1 17

[29] Siewiorek, Daniel, Asim Smailagic, and Anind Dey. "Architecture and applications of virtual coaches." Proceedings of the IEEE 100.8 (2012): 2472-2488

[30] op den Akker, Harm, et al. "Opportunities for smart & tailored activity coaching." 2013 IEEE 26th International Symposium on Computer-Based Medical Systems (CBMS). IEEE Computer Society, 2013

[31] F. Carros et al: "Not that uncanny after all? An Ethnographic Study on Android Robots Perception of Older Adults in Germany and Japan", International Conference on Social Robotics, Florence, Italy, Dec 2022.

[32] S. L. Anderson: "Asimov's "three laws of robotics" and machine Metaethics". AI Soc 22(4):477–493, 2008

[33] Y. Weng, Y. Hirata: "Design Cantered HRI Governance for Healthcare Robots", Journal of Healthcare Engineering

[34] J. Knoben and L.A.G. Oerlemans: "Proximity and inter-organizational collaboration: A literature review", International Journal of Management Reviews, Volume 8, Issue 9, pp. 71-89, 2006.

[35] T. Kruse, P. Basili, S. Glasauer, A. Kirsch: "Legible robot navigation in the proximity of moving humans", IEEE Workshop on Advanced Robotics and its Social Impacts, ARSO, 2012

[36] F. Babel et al: "Small Talk with a Robot? The Impact of Dialog Content, Talk Initiative, and Gaze Behavior of a Social Robot on Trust, Acceptance, and Proximity", Int J of Social Robotics, Vol. 13, pp. 1485–1498, 2021

[37] P. Saulnier, E. Sharlin, S.Greenberg: "Exploring minimal nonverbal interruption in HRI", International Conference on Robot and Human Interactive Communication (RO-MAN), 2011

[38] A. Winfield: "Ethical standards in Robotics and AI", Nature Electronics, 2, 46-48, 2019

[39] J. Bryson, A. Winfield: "Standardizing Ethical Design for Artificial Intelligence and Autonomous Systems", Computer, 50(5), 2017