Voice-based virtual assistants design and European legislation: the interpretation of subliminality, manipulation and deception*

Vittoria Caponecchia[1]

[1]Ph.D. student in Artificial Intelligence for Society, XXXVIII Cycle, University of Pisa - Scuola Superiore Sant'Anna, Pisa

Abstract

This position paper, which is a work in progress of my PhD research, highlights the terminological unclarity of article 5(1)(a) of the AI Act regarding the terms subliminal, manipulative and deceptive techniques, that the article itself prohibits. Since it is unclear how these concepts should be interpreted, it is difficult to determine when AI systems should be considered subliminal, manipulative or deceptive. In order to prevent deployers and providers from developing, deploying, or commercializing such systems, it is necessary to specify the meaning of these expressions. To do so, the European regulatory framework on digital services (DSA), data (Data Act) and consumer protection will be analyzed, both at the European level (Unfair Commercial Practices Directive) and Italian level (Legislative Decree No. 145/2007 and Legislative Decree No. 146/2007). The context in which this analysis will be conducted is that of voice-based virtual assistants, in order to understand whether and how these increasingly used software can deceive or manipulate consumers.

Keywords

Voice-Based Virtual Assistants, Subliminal techniques, Manipulative techniques, Deceptive techniques, AI Act.

1. Introduction

In a world pervaded by artificial intelligence (AI) it is necessary for the law to maintain a predominant position, guaranteeing the protection and preservation of human rights and interests, especially in terms of legal certainty. This is because, while AI undoubtedly brings benefits in any field, it also entails risks for both individuals and society.

It is proving increasingly problematic, however, to ensure that the law keeps pace with the development of new technologies, which run much faster and therefore become difficult to regulate. Precisely for this reason, several regulations have been proposed and even adopted at EU level, the most recent of which, in final approval phase, is the Artificial Intelligence Act

Mobilizing Research and Regulatory Action on Dark Patterns and Deceptive Design Practices Workshop at CHI conference on Human Factors in Computing Systems, May 12, 2024, Honolulu, HI (Hybrid Workshop)

*This contribution is a reworking and elaboration of a policy recommendation written by the author, in Rossi, A., & Comandé, G. (2023). D.7.6 BRIEF Report on policy design and advice. Zenodo. https://doi.org/10.5281/zenodo.10869888, [1]

vittoria.caponecchia@santannapisa.it (V. Caponecchia)

(AI Act)[1], which fits perfectly within the European digital strategy[2], the aim of which is to create a single European data space (single market for data) while leaving a central position for humans[3].

The AI Act establishes harmonised rules for artificial intelligence, with the aim, among others, of meeting the requirements of a well-functioning internal market[4], ensuring a high level of data protection, digital rights and ethical standards[5], and addressing the opacity and complexity of AI systems, as well as a certain degree of unpredictability and partially autonomous behaviour of certain AI systems, to ensure their compatibility with fundamental rights and to facilitate the enforcement of legal rules[6].

Nonetheless, although the specific objectives of the proposal include ensuring legal certainty and improving the effective application of existing legislation, the proposal itself emphasises, in recital 28, how artificial intelligence today *"can also be misused and provide novel and powerful tools for manipulative, exploitative and social control practices"*.

These tools also include so-called voice-based virtual assistants (VAs), software that allows users to control smart devices via voice commands[2][3]. They consist of Natural Language Interfaces (NLI) powered by machine learning and the collection of large amounts of personal data and used to access certain services (online stores, search engines, social networks, etc.).

Voice-based virtual assistants often succeed in influencing individuals, since the user interface (UI) – consisting in this case of a voice and not a graphical interface – and the user experience (UX), which constitute the digital architecture of the system, can implement deceptive design techniques.

Nevertheless, the article 5(1)(a) of the AI Act prohibits *"the placing on the market, the putting into service or the use of an AI system that deploys subliminal techniques beyond a person's consciousness or purposefully manipulative or deceptive techniques, with the objective, or the effect of, materially distorting the behaviour of a person or a group of persons by appreciably impairing their ability to make an informed decision, thereby causing a person to take a decision that that person would not have otherwise taken in a manner that causes or is likely to cause that person, another person or group of persons significant harm"*.

Therefore, first of all, it is necessary to try to specify the meanings of the terms *subliminal, manipulative* and *deceptive techniques*, in order to understand what they are and whether voice-

[1]Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislative acts, COM/2021/206 final. The last vote on the Regulation text took place on March 13th by the European Parliament, but it still has to be formally approved by the Council. The reference text for this paper is that of 13 March 2024, https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.pdf

[2]Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, "A european strategy for data", COM(2020) 66 final; Commission's Communication on "Shaping Europe's digital future", 2020.

[3]Also the White Paper On Artificial Intelligence states that AI should be a tool for people and a positive factor for society, with the ultimate aim of improving the well-being of human beings, in White Paper On Artificial Intelligence - A European approach to excellence and trust, COM(2020) 65 final.

[4]Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislative acts, COM/2021/206 final, p. 11.

[5]European Council, European Council meeting (19 October 2017) – Conclusion EUCO 14/17, 2017, p. 8

[6]Council of the European Union, Presidency conclusions - The Charter of Fundamental Rights in the context of Artificial Intelligence and Digital Change, 11481/20, 2020, p. 5.

based virtual assistants can fall under them. By reducing terminological uncertainty, legal uncertainty will also automatically be reduced.

2. Subliminal, manipulative and deceptive techniques: what European legislation provides about them

This paper was written after examining the most recent regulations that are applicable within the scope, and for the purposes, of the European digital strategy and having selected the most appropriate ones to analyze the issue at hand (Digital Services Act – DSA[7] – and Data Act[8]).

In addition to these, the most important consumer protection legislation was studied (Unfair Commercial Practices Directive – UCPD[9]; and, at Italian level, Legislative Decree No. 145/2007[10] and Legislative Decree No. 146/2007[11]), insofar as the aforementioned techniques can be classified as unfair commercial practices and therefore subject to the relevant discipline.

Specifically, after also considering other important legislation at European level, such as the GDPR [12] and the Digital Markets Act (DMA) [13], only DSA, Data Act and UCPD were selected, because they were closer to the topic at hand. The two implementing decrees of the latter directive in Italy were then included, in order to understand whether and what repercussions the issue can have in an European Union Member State.

The above-mentioned terms (subliminal, manipulative and deceptive techniques) were searched for in their texts and then it was proceeded to their interpretative analysis, also taking into account the contexts to which they refer.

It was observed that none of these regulations contain direct references to the notions of subliminal, manipulative and deceptive techniques, but they may contain references in general to subliminality, manipulation and deception, terms that are united by the fact that they fall within (or, as the case may be, contain the) category of dark patterns. Since dark patterns consist of *"instances where designers use their knowledge of human behavior (e.g., psychology) and the desires of end users to implement deceptive functionality that is not in the user's best interes"*[4][5],

[7]Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 october 2022 on a single market for digital services and amending Directive 2000/31/EC (Digital Services Act).

[8]Regulation (EU) 2023/2854 of the European Parliament and of the Council of 13 December 2023 on harmonised rules on fair access to and use of data and amending Regulation (EU) 2017/2394 and Directive (EU) 2020/1828 (Data Act).

[9]Directive 2005/29/EC of the European Parliament and of the Council of 11 May 2005 concerning unfair business-to-consumer commercial practices in the internal market and amending Council Directive 84/450/EEC, Directives 97/7/EC, 98/27/EC and 2002/65/EC of the European Parliament and of the Council and Regulation (EC) No 2006/2004 of the European Parliament and of the Council (Unfair Commercial Practices Directive).

[10]Legislative Decree No. 145 of 2 August 2007 "Implementation of Article 14 of Directive 2005/29/EC amending Directive 84/450/EEC concerning misleading advertising", published in the Official Gazette No. 207 of 6 September 2007.

[11]Legislative Decree No. 146 of 2 August 2007 "Implementation of Directive 2005/29/EC concerning unfair business-to-consumer commercial practices in the internal market and amending Directives 84/450/EEC, 97/7/EC, 98/27/EC, 2002/65/EC, and Regulation (EC) No. 2006/2004", published in the Official Gazette No. 207 of 6 September 2007.

[12]Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016, on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).

[13]Regulation (eu) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on contestable and fair markets in the digital sector and amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act).

it can be said that they may use subliminality, manipulation or deception, while the reverse is not always true (not all of these techniques are used online and target a user).[14]

To find an unambiguous meaning of the expressions mentioned in Article 5(1)(a) of the AI Act, or at least to try to better understand what they refer to, the aforementioned regulations have been examined.

2.1. Subliminal techniques

Starting with the notion of *"subliminal technique"*, none of the above-mentioned regulations contain such an expression. Taking into account the Italian area, the Italian Legislative Decree No. 145/2007, concerning misleading advertising, affirm, in article 5, the need for transparency in advertising and expressly prohibits subliminal advertising.

The same decree, in article 1, states that *"advertising must be clear, truthful and correct"*[15], while article 2 defines misleading advertising as *"any advertising which in any way, including its presentation, is likely to mislead the natural or legal persons to whom it is addressed or whom it reaches and which, by reason of its misleading character, is likely to prejudice their economic behaviour, or which, for that reason, is likely to harm a competitor"*.

At this point, three questions spontaneously arise. The first concerns the fact that this decree covers advertising, so what is not advertising how should it be regulated? Can this discipline be extended?

The other two, instead, concern the interpretation of the term *"subliminal"*:

- Does it refer to advertisement that is not *"clear, truthful and correct"*[16] (since, if an advertisement must be transparent in order not to be considered subliminal, then it must also be clear)? ;
- Assuming that *"transparent"* is equivalent to *"clear"*[17], is an advertisement that is not transparent then misleading? If so, does *"subliminal"* then fall under the latter definition?

2.2. Manipulative techniques

As far as *"manipulative techniques"* are concerned, this term is found in both the DSA and the Data Act, but with different nuances.

In the DSA the most relevant references to manipulation are to be found in some recitals, which do not provide a precise definition of the term in question, but allow us to understand what is meant, indicating it as a technique that *"alters the integrity of information transmitted or to which access is provided"* (recital 21), or that *"may have a negative impact on entire groups and*

[14]The term "dark pattern" was coined in 2010 by Harry Brignull, U.S. researcher and user experience designer, who defined them as *"a user interface that has been carefully crafted to trick users into doing things, such as buying insurance with their purchase or signing up for recurring bill"* (H. Brignull [6][7]). Then, it was taken up by many other scholars in later years, who defined them, substantially, as *"user interface design choices that benefit an online service by coercing, steering, or deceiving users into making unintended and potentially harmful decisions"*, in A. Mathur et al.[8]

[15]Personal translation of art. 1, Legislative Decree 145/2007, which states: *"La pubblicità deve essere palese, veritiera e corretta"*.

[16]Ibid.

[17]Ibid.

amplify social harm" (recital 69). Furthermore, recital 83, which concerns the fourth category of systemic risks that undermine online security through certain *"design, functioning or use of very large online platforms and of very large online search engines"*, mentions manipulation as a means by which these risks could materialise. Recital 84, indeed, also calls for an assessment of manipulation.

The main reference to manipulative techniques made by the Data Act, on the other hand, is contained in recital 38, which prohibits a third party from using coercive, deceptive *"or"* manipulative means (thus, implicitly differentiating them from each other, but without specifying why they differ) against the user, subverting or impairing the user's autonomy, decision-making or choices, including through a digital interface or a part thereof. It states also that third parties or data holders should not even refer to dark patterns in their design, describing them as *"design techniques that push or deceive consumers into decisions that have negative consequences for them"*. According to this recital, dark patterns do not correspond exactly to *"coercive, deceptive or manipulative means"*, but they are a subcategory of them and, in particular, of deceptive means. Moreover, the term *"persuasion"*, used in this context, suggests that deception can be associated with persuasion itself. Nevertheless, according to recital 38, the concepts of persuasion and manipulation could also be associated (*"those manipulative techniques can be used to persuade users"*) – in effect, the former can be seen as a subcategory of the second (some understand persuasion as the impulse that rationally convinces people to do something, thus never pushing them to perform an unwanted behaviour)[9]. Actually, it's not quite like that, as will be explained later. So, is deception also a subcategory of manipulation? And in what terms? And what does manipulation consist of?

With regard to dark patterns, specifically, recital 67 of the DSA (later implemented by art. 25) provides a definition similar to that of the Data Act, indicating them as *practices that distort or impair, either on purpose or in effect, the ability of recipients of the service to make autonomous and informed choices* and which, for that reason, should be prohibited.

Similarly to the Data Act, the DSA mentions deception rather than manipulation and it additionally refers to *"non-neutrality"*, which could be linked to the expression *"subliminal technique"*. Indeed, non-neutrality may consist of a partial or biased attitude, which can be held through subliminal techniques, in order to steer recipients in a certain direction, without explicitly stating intentions. At the same time, the use of subliminal techniques may serve precisely to achieve a purpose, in a more subtle way. And, in connection with what has been said above, in the analysis of the Italian Legislative Decree No. 145/2007, if subliminal technique were to be equated with a lack of transparency, the fact that the concepts of non-neutrality and subliminality can coexist would also include the concept of non-transparency: the subliminal (or non-transparent) technique can be the means by which non-neutrality is exercised or the very result of the experiment of a non-neutral action, thus the lack of transparency allows (or leads) to a non-neutral result.

2.3. Deceptive techniques

Coming finally to the analysis of the term *"deceptive technique"*, it could be argued that it is the least problematic since it is much more widespread in the regulatory texts mentioned so far. However, there is no definition of this term, which it is found in the form of *"misleading*

commercial practice" in the Unfair Commercial Practices Directive (later incorporated by Italian Legislative Decree No. 146/2007).

In this context, it is assumed that the terms *"misleading"* and *"deceptive"* can be considered synonymous, since articles 6 and 7 expressly contain the statement *"a commercial practice shall be regarded as misleading if it contains false information and is therefore untruthful or in any way, including overall presentation, deceives or is likely to deceive the average consumer"*.

In particular, the Directive defines *"misleading commercial practices"* as those which contain false, untruthful or deceiving information or which, in any way, are likely to deceive the average consumer (art. 6.1), even on the basis of the circumstances in which they are given (art. 6.2) or which, on the contrary, omit information (art. 7.1) or make it unclear, unintelligible or ambiguous (art. 7.2) and which, in any case, cause them to take a decision they would not otherwise have taken.

It should be recalled, however, that the Directive covers *"commercial practices directly related to influencing consumers' transactional decisions in relation to products. It does not address commercial practices carried out primarily for other purposes"* (recital 7). This means that everything outside the commercial scope and unrelated to a product is excluded from such discipline. Article 5 of the AI Act, on the other hand, concerns AI systems in general, so they could have negative implications both in commercial terms and non-commercial terms (e.g. they could aim at obtaining consent and personal data, just think of online phishing).

What is evident is the link between deceptive techniques, as defined in the commercial sphere, and dark patterns, which the community now calls *"deceptive design patterns"*[3][7][18], and which also have purposes other than purely commercial ones (e.g. showing the user a pre-selected check box for accepting political communications without clear or accurate information on the content or origin of such communications, could influence the political opinion of users).

3. Voice-based virtual assistants

In this context, voice-based virtual assistants are inserted. They are computer programs designed to interact with users and assist in performing various tasks, such as asking questions, providing information, executing specific actions, and more, all through voice conversation.

This is possible through the use of two key components, which allows them to learn techniques and social abilities to offer adequate usability experiences for users[11]: Natural Language Processing (NLP) and Application Programming Interfaces (APIs).

NLP is the set of methods for making human language accessible to computers. Drawing on many other intellectual traditions, it combines computational linguistics with statistical and machine learning models to enable computers and digital devices to recognize, understand and generate text and speech[12]. Virtual assistants use NLP to interpret users' voice requests by analyzing the meaning of words, sentence structure and conversation context. This allows them to understand users' questions and provide relevant and useful responses.

APIs are the interfaces *"between an application and a library"*[13] that consist of two parts: declaring code (the code an application developer needs to use to call upon specific functionality in the library) and structure, sequence and organization[14] (the taxonomy under which the

[18]The term "dark patterns" has also been criticised, in C. Sinders [10].

declaring code is structured[15]). So, essentially, APIs are sets of tools and definitions that enable different software applications to communicate with each other. Virtual assistants use them to access a wide range of services and functionalities offered by other applications and platforms. For example, through APIs, a virtual assistant like Alexa can integrate with music streaming services like Spotify, retrieve updated news from online sources, book a taxi through ride-sharing services like Uber, and much more.

NLP and APIs are used in conjunction with recommendation systems, information filtering systems providing a personalized item recommendation to a user in a service environment that can hold or collect various data[16][17]. They may suggest results on the basis of past interactions between users and items to predict future interactions (*collaborative filtering*), similarities between user preferences and item characteristics (*content filtering*), or contextual information, like user location, device, and time, into the recommendation process (*context filtering*)[18].

Given the widespread use of such tools, there is an urgent need to understand what kind of impact they may have on consumers. The challenge becomes arduous since, unlike visual interfaces, in VAs the voice is the only element that allows interaction with the user and which, for this reason, can be exploited by designers to deceive or manipulate them, even if *"there are seldom services that offer voice-only interfaces"*[3].

Indeed, a study[19] has shown how the use of microphones and the lack of transparency about smart speaker data practices are central to people's concerns. This could constitute a dark pattern, as the user is unaware of how their data is being collected and processed, thus making choices based on incomplete and misleading information. Because of these characteristics, virtual assistants are able to affect consumers concerns and persuade them[2][3], especially because of the mode used (voice instead of text) and the adoption of a human name[20] (instead of a robot name or being nameless)[21]. From the same study it emerged that consumers are more inclined to purchase if the virtual assistant possesses these features. Moreover, as people have less experience with receiving persuasive messages via voice-controlled devices than via screens, they are likely to be less aware that a recommendation is, in reality, a persuasive attempt[19].

Specifically, with regard to persuasion, participants of the study[19] were more aware that the recommendation was an attempt at persuasion when it was presented on a smartphone screen, whereas, if it came from a virtual assistant, they evaluated the brand more positively. However, if the virtual assistant had a human name (such as Alexa or Siri), they followed the recommendation even if it was a persuasion attempt. And, if both social cues are present (i.e., voice and the adoption of a human name[22]), it is likely that the effects will be additive.

Social cues are of five types[20], but voice-based virtual assistants mainly exploit two of them: psychological and language. The first one can lead people to infer, often subconsciously, that the product has emotions, preferences, motivations and personality, as if the machine (in this case the virtual assistant) had a psychology and empathy [21]. The second one, instead, can use written or spoken language to convey social presence and to persuade. By this means, it

[19]Ibid.

[20]Physical, psychological, language, social dynamics, social roles, in B. Fogg [23].

[21]Indeed, even those who are experts in the field treat computer products, as if they had preferences and personalities, ibid.

is possible to providing recommendations to users, as well as establishing a bond with them that is able to engage and retain them. It should be noted, furthermore, that most voice-based assistants not only have names resembling female ones, but also have female voices [22].

Instead, with regard to deception, a recent study[3] revealed that voice interfaces have six unique properties, which *"may be used to (intentionally or accidentally) implement deceptive design patterns"*[3] and more than half concerns the use of voice.

The latter are *discoverability* (it is difficult to know which command to use exactly to achieve a certain result, since the VA often does not respond), *physical domain* (the activation of the VA is often not realized until it occurs, maybe due to an accidental pronunciation of the wake-up word or the arrival of a notification), *linearity* (since there is no visual interface, it is no possible to go back without starting the whole process over again, so it is necessary to listen to the entire VA's speech) and *volume* (the VA can influence the user's choices based on the volume, tone, speed, fluency, pronunciation, articulation and emphasis of the voice). While, the properties not directly related to the voice are *multiple interfaces* (the VA is often connected to other services and, in order to give an answer, it must connect to them, forcing the user to change interface) and *unclear context* (often the VA relies on several contexts, e.g. features and skills, which enable it to respond to the user's request by sharing its data with third entities, often without the user being aware of it[23][25][26][3].

Nevertheless, the result of this study shows that users often do not perceive the virtual assistant's behaviour as manipulative and, on the contrary, justify it by recognising its limitations as non-human. Users often recognised the pattern as problematic, but normal, expected, satisfying or even useful[24]. However, obviously, if designed using deceptive patterns, virtual assistants may lead the consumer to make choices he would not otherwise have made.

Having established that virtual assistants can influence user behavior and persuade them to achieve a specific goal, some also utilize techniques of gamification and digital nudging to make the user experience more engaging and increase the usage of such systems, leveraging persuasive design systems[27]. It is precisely in this way that attempts at manipulation may arise, which, instead of guiding the user to make choices in his or her best interest, could lead to unexpected and risky outcomes.

Some forms of gamification may require the collection of users' personal data to function effectively. This could raise concerns about user privacy and data security. Moreover, it could lead to dependence on the use of the virtual assistant, with potential negative consequences on their mental health and lifestyle habits, and it could distract users from important tasks or reduce their ability to focus on specific activities.

[22] From a conducted survey, it emerged that humans tend to attribute human traits to computers and that the female voice, in virtual assistants, is perceived as friendlier and more pleasant than the male voice (CASA paradigm: describes the human tendency to assign human traits to computers, in Claus-Peter H. Ernst et al.[24]). Obviously, this has practical implications for the development and design of devices utilizing virtual assistants, which leverage these factors to persuade consumers, in Claus-Peter H. Ernst et al. [24].

[23] Smart speakers can also run automation applications such as *IFTTT* (*If This Then That*) that connect to other services on user's phones and in the cloud, in J. Lau et al. [19].

[24] A study showed that, in contrast to the elderly, who feel deprived of their autonomy by using VAs, people with disabilities feel more autonomous, in J. Lau et al. [19]

4. Conclusions

In light of this, it becomes increasingly urgent to regulate the use of virtual assistants within a framework of regulations aimed at minimizing potential negative consequences.

In response to the questions posed in this paper, it is believed that:

- Subliminal techniques differ from misleading techniques: the former are stimulus too weak to be perceived and recognised, but not so weak that they don't influence a person's behaviour or psyche; while the latter easily lead to false beliefs and can be more easily perceived by humans. However, like misleading techniques, subliminal techniques correspond to practices that are not clear, truthful, and correct;
- Manipulative techniques are not to be equated with deceptive techniques (which do not even represent a subcategory of manipulation): the former compromise a person's reasoning, for example, influencing the surrounding environment, controlling and/or distorting information and exploiting emotions; while the latter make one believe in a falsehood through misleading evidence[28]. So, manipulation may involve deceptive tactics, while the opposite is not true. Despite some discussing *"deceptive persuasive practices"*, deception should be distinguished from persuasion, as the latter involves attempting to convince people to do something (it is the type of influence that arises from social situations[29]) but without circumventing them. In any case, virtual assistants may include both manipulative and deceptive techniques;
- It's true that the Unfair Commercial Practices Directive only covers commercial practices directly related to influencing consumers' transactional decisions in relation to products, but virtual assistants can be considered as such, since they can influence consumer choices, personalizing recommendations and influence their buying process[30].

It is important to recognize that voice-based virtual assistants can employ a variety of techniques, including subliminal, manipulative and deceptive, to influence user decisions and behaviors.

AI Act in this matter lacks in terms of definitions and the other mentioned regulations are not very clear, leaving the task of uncovering the meanings of the above-mentioned terms to those who must interpret or apply it. This causes, on one hand, uncertainty, as those designing certain AI systems will not be sure of which rules to follow, and, on the other hand, mistrust in justice, as the idea may spread that similar cases could be judged differently simply because they are evaluated by different judges. As existing rules on design are often guidelines and therefore not binding, it is up to the law to be the beacon that lights the way.

This paper highlights the questions arising from the most recent Regulation on Artificial Intelligence (AI Act), the only one to directly mention subliminality, manipulation and deception with reference to AI systems. It indicates a possible interpretation of these terms, however, this is only a proposal, that will need to be further explored and modified in light of any developments following the final approval of the AI Act, expected in April 2024.

In any case, only through a combination of effective regulation, transparency and a law-compliant design of AI systems can it be ensured that voice-based virtual assistants are used ethically and responsibly, for the benefit of all stakeholders involved.

Acknowledgments

The author is very grateful to her Ph.D. Professors for their support and guidance, especially Arianna Rossi for the continuous stimuli she offers to her research and for her time, encouragement and advice. Finally, she thanks the anonymous reviewers for their helpful feedback on the previous version of this paper.

References

[1] A. Rossi, G. Comandé, D.7.6 brief report on policy design and advice, 2024. URL: https://doi.org/10.5281/zenodo.10869888. doi:10.5281/zenodo.10869888.

[2] S. D. Conca, The present looks nothing like the jetsons: Deceptive design in virtual assistants and the protection of the rights of users, Computer Law Security Review 51 (2023) 105866. URL: https://www.sciencedirect.com/science/article/pii/S0267364923000766. doi:https://doi.org/10.1016/j.clsr.2023.105866.

[3] K. Owens, J. Gunawan, D. Choffnes, P. Emami-Naeini, T. Kohno, F. Roesner, Exploring deceptive design patterns in voice interfaces, in: Proceedings of the 2022 European Symposium on Usable Security, 2022, pp. 64–78.

[4] S. Greenberg, S. Boring, J. Vermeulen, J. Dostal, Dark patterns in proxemic interactions: a critical perspective, in: Proceedings of the 2014 conference on Designing interactive systems, 2014, pp. 523–532.

[5] C. M. Gray, Y. Kou, B. Battles, J. Hoggatt, A. L. Toombs, The dark (patterns) side of ux design, in: Proceedings of the 2018 CHI conference on human factors in computing systems, 2018, pp. 1–14.

[6] H. Brignull, M. Leiser, C. Santos, K. Doshi, Deceptive patterns – user interfaces designed to trick you, 2023. URL: https://www.deceptive.design/.

[7] H. Brignull, Deceptive Patterns: Exposing the Tricks Tech Companies Use to Control You, Harry Brignull, 2023. URL: https://books.google.it/books?id=PysW0AEACAAJ.

[8] A. Mathur, G. Acar, M. J. Friedman, E. Lucherini, J. Mayer, M. Chetty, A. Narayanan, Dark patterns at scale: Findings from a crawl of 11k shopping websites, Proceedings of the ACM on Human-Computer Interaction 3 (2019) 1–32.

[9] D. Susser, B. Roessler, H. Nissenbaum, Online manipulation: Hidden influences in a digital world, Geo. L. Tech. Rev. 4 (2019) 1.

[10] C. Sinders, What's in a name? unpacking dark patterns versus deceptive design., 2022. Https://medium.com/@carolinesinders/whats-in-a-name-unpacking-dark-patterns-versus-deceptive-design-e96068627ec4.

[11] V. Claessen, A. Schmidt, T. Heck, Virtual assistants, in: Everything Changes, Everything Stays the Same? Understanding Information Spaces. Proceedings of the 15th International Symposium of Information Science (ISI 2017), 2017, pp. 116–130. doi:http://dx.doi.org/10.18452/1445.

[12] J. Eisenstein, Introduction to natural language processing, MIT press, 2019.

[13] T. M. Gage, Whelan associates v. jaslow dental laboratories: Copyright protection for

computer software structure - what's the purpose note, Wisconsin Law Review 1987 (1987) 859.

[14] K. Taylor, Oracle am., inc. v. google inc. 750 f.3d 1339 (fed. cir. 2014) survey, Intellectual Property Law Bulletin 19 (2014-2015) 221.

[15] P. Sagdeo, Application programming interfaces and the standardization-value appropriation problem, Harv. JL & Tech. 32 (2018) 235.

[16] H. Ko, S. Lee, Y. Park, A. Choi, A survey of recommendation systems: Recommendation models, techniques, and application fields, Electronics 11 (2022). URL: https://www.mdpi.com/2079-9292/11/1/141. doi:10.3390/electronics11010141.

[17] R. Van Meteren, M. Van Someren, Using content-based filtering for recommendation, in: Proceedings of the machine learning in the new information age: MLnet/ECML2000 workshop, volume 30, Barcelona, 2000, pp. 47–56.

[18] F. Isinkaye, Y. Folajimi, B. Ojokoh, Recommendation systems: Principles, methods and evaluation, Egyptian Informatics Journal 16 (2015) 261–273. URL: https://www.sciencedirect.com/science/article/pii/S1110866515000341. doi:https://doi.org/10.1016/j.eij.2015.06.005.

[19] J. Lau, B. Zimmerman, F. Schaub, Alexa, are you listening? privacy perceptions, concerns and privacy-seeking behaviors with smart speakers, Proceedings of the ACM on human-computer interaction 2 (2018) 1–31.

[20] S. T. Völkel, D. Buschek, M. Eiband, B. R. Cowan, H. Hussmann, Eliciting and analysing users' envisioned dialogues with perfect voice assistants, in: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, CHI '21, Association for Computing Machinery, New York, NY, USA, 2021. URL: https://doi.org/10.1145/3411764.3445536. doi:10.1145/3411764.3445536.

[21] H. A. Voorveld, T. Araujo, How social cues in virtual assistants influence concerns and persuasion: the role of voice and a human name, Cyberpsychology, Behavior, and Social Networking 23 (2020) 689–696.

[22] S. S. Sundar, The MAIN model: A heuristic approach to understanding technology effects on credibility, MacArthur Foundation Digital Media and Learning Initiative Cambridge, MA, 2008.

[23] B. Fogg, Persuasive Technology: Using Computers to Change What We Think and Do, Interactive Technologies, Morgan Kaufmann Publishers, 2003. URL: https://books.google.it/books?id=9nZHbxULMwgC.

[24] C.-P. H. Ernst, N. Herm-Stapelberg, The impact of gender stereotyping on the perceived likability of virtual assistants., in: AMCIS, 2020.

[25] D. Major, D. Y. Huang, M. Chetty, N. Feamster, Alexa, who am i speaking to?: Understanding users' ability to identify third-party apps on amazon alexa, ACM Trans. Internet Technol. 22 (2021). URL: https://doi.org/10.1145/3446389. doi:10.1145/3446389.

[26] A. Sabir, E. Lafontaine, A. Das, Hey alexa, who am i talking to?: Analyzing users' perception and awareness regarding third-party alexa skills, in: Proceedings of the 2022 CHI conference on human factors in computing systems, 2022, pp. 1–15.

[27] D. Benner, S. Schöbel, A. Janson, Exploring the state-of-the-art of persuasive design for smart personal assistants, in: Innovation Through Information Systems: Volume II: A Collection of Latest Research on Technology Issues, Springer, 2021, pp. 316–332.

[28] V. Krstić, Manipulation, deception, the victim's reasoning and her evidence, Analysis (2024) anad064.
[29] P. G. Zimbardo, M. R. Leippe, The psychology of attitude change and social influence., Mcgraw-Hill Book Company, 1991.
[30] D. Kim, K. Park, Y. Park, J. Ju, J.-H. Ahn, Alexa, tell me more: The effect of advertisements on memory accuracy from smart speakers (2018).