10-9-2023

Uncovering Drivers for the Integration of Dark Patterns in Conversational Agents

Tim Kollmer
University of Innsbruck, Innsbruck, Austria, Tim.Kollmer@uibk.ac.at

Alessa Hauser
University of Innsbruck, Innsbruck, Austria, alessa.hauser@student.uibk.ac.at

Viviana Oberhofer
University of Innsbruck, Innsbruck, Austria, viviana.oberhofer@uibk.ac.at

Gregor Blossey
University of Innsbruck, Innsbruck, Austria; European University Viadrina, Frankfurt (Oder), Germany, gregor.blossey@uibk.ac.at

Andreas Eckhardt
University of Innsbruck, Innsbruck, Austria, andreas.eckhardt@uibk.ac.at

Uncovering Drivers for the Integration of Dark Patterns in Conversational Agents

Research Paper

Tim Kollmer[1], Alessa Hauser[1], Viviana Oberhofer[1], Gregor Blossey[1,2], and Andreas Eckhardt[1]

[1] University of Innsbruck, Innsbruck, Austria
{tim.kollmer, viviana.oberhofer, gregor.blossey, andreas.eckhardt}@uibk.ac.at
alessa hauser@student.uibk.ac.at
[2] European University Viadrina, Frankfurt (Oder), Germany

Abstract. Today, organizations increasingly utilize conversational agents (CAs), which are smart technologies that converse in a human-to-human interaction style. CAs are very effective in guiding users through digital environments. However, this makes them natural targets for dark patterns, which are user interface design elements that infringe on user autonomy by fostering uninformed decisions. Integrating dark patterns in CAs has tremendous impacts on supposedly free user choices in the digital space. Thus, we conducted a qualitative study consisting of semi-structured interviews with developers to investigate drivers of dark patterns in CAs. Our findings reveal that six drivers for the implementation of dark patterns exist. The technical drivers include heavy guidance of CAs during the conversation and the CAs' data collection potential. Additionally, organizational drivers are assertive stakeholder dominance and time pressure during the development process. Team drivers incorporate a deficient user understanding and an inexperienced team.

Keywords: Dark patterns, conversational agents, dark pattern integration

1 Introduction

In the age of digitization, organizations increasingly use conversational agents (CAs) to improve their processes (Schuetzler et al., 2021). By definition, CAs are artificial technologies that interact with individuals through text or speech in natural language (Luger and Sellen, 2016, Kundinger et al., 2019). For users, the interaction with CAs is very intuitive as they mimic human-to-human dialogues. Moreover, CAs enable direct responses to user inquiries without long waiting times (Maedche et al., 2019, Diederich et al., 2022).

Because of their anthropomorphic nature, CAs have wider spectrum of persuasion strategies at their disposal (Lehto and Oinas-Kukkonen, 2015). As a result, CAs are very effective in streamlining and steering user decisions toward a specific outcome.

18th International Conference on Wirtschaftsinformatik,
September 2023, Paderborn, Germany

However, these persuasion strategies are prone to be exploited by organizations to advance their own goals at the user's expense (Lehto and Oinas-Kukkonen, 2015). As CAs are limited to a pre-selected number of options, the decision space for the users is very restricted because a complete overview of possible options is not provided (Rheu et al., 2021).

Likewise, the application of persuasion strategies through user interface design elements on users' is not a new phenomenon. Studies have shown that guiding the user by highlighting a specific option increases the likelihood of this option being chosen by 39% on average (Hummel and Maedche, 2019). The potential to covertly influence users' decision-making process is increasingly exploited by organizations and often manifests in so-called dark patterns. These patterns refer to digital design compositions that coerce individuals into making uninformed decisions that benefit an organization (Kollmer and Eckhardt, 2023).

Recognizing the tremendous potential, organizations have begun to integrate dark patterns into CAs in an effort to maximize their capacity to manipulate behavior (Harjunen et al., 2018). For instance, if users want to turn off Apple's CA, Siri, it is not possible to do so via conversation. Instead, users are forced to manually navigate through the settings menu in order to achieve their objectives (Johnson, 2022). This dark pattern deliberately increases the effort for the given task and thus discourages individuals from deactivating Siri on their devices. Clearly, the capability to systematically influence human behavior on a grand scale gives reason for concern as this development comes with many ethical implications regarding the user's free will. Even though the integration of dark patterns in CAs is a concerning phenomenon, research in this field is very scarce. Extant literature mainly focuses on the consequences and perceptions of CAs (Diederich et al., 2022), which extends to ethical implications (Banks, 2019) and design guidelines for CAs to protect against dark patterns (Yang and Aurisicchio, 2021). Up to this date, the most effective protective measure is arguable to prevent their integration in the first place. As a first step, our objective is to deepen our understanding of the origin of dark patterns and the driving forces for their integration during the development process. Therefore, our work is guided by the following research question (RQ):

What drives the integration of dark patterns in CAs?

To address this issue, we conducted a qualitative study consisting of semi-structured interviews with experts in CA development. We identify technical, organizational and team drivers that foster the integration of dark patterns in CAs. The technical drivers include heavy guidance of CAs during the conversation and the CAs' data collection potential. Additionally, organizational drivers are assertive stakeholder dominance and time pressure during the development process. Team drivers incorporate a deficient user understanding and an inexperienced team.

The remainder of the paper is structured as follows. Section 2 provides the theoretical background on CAs and dark patterns by discussing relevant related literature. In section 3, the applied methodology is outlined. The results are reported in section 4, followed by implications for research and practice in section 5.

2 Theoretical Background

2.1 CA Design and Application

CAs are characterized as dialog systems to converse with humans using natural language (Diederich et al., 2022, Feine et al., 2019a). Overall, CAs exist in multiple different manifestations. For instance, CAs partially imply disembodied interactions through text (e.g., chatbots, digital agents), embodied interactions via speech (e.g., social robots), or disembodied interactions through speech (e.g., smart personal assistants, voice assistants, digital agents) (Diederich et al., 2022). Up to this date, existing research predominantly highlights the human-like characteristics of CAs during the interaction. This includes, for instance, the personality of the CA (Westhoven and Tegtmeier, 2021, Złotowski et al., 2016). Commercially available CAs simulate human-like social cues and human-to-human conversations to create a natural interaction between the agent and the user. These designs indicate to improve the interaction quality (Araujo, 2018, Chen et al., 2021), increase the willingness to reuse the technology, the enjoyment (De Cicco et al., 2021, Song and Kim, 2022), trust (You and Robert, 2018, Song and Kim, 2022), and the perceived human-likeness (Araujo, 2018).

CAs have proven beneficial in collaboration and decision support contexts. In general, CAs are adopted by organizations across multiple industries (e.g., healthcare, education, and business), filling different roles (e.g., tutor, consumer service agent, and therapist) (Ciechanowski et al., 2019, Xu and Lombard, 2017, Chin and Yi, 2021). From a technical perspective, CAs process the input provided from the user to assess the intent of the user based on artificial intelligence and past training data. As a result, CAs respond with the most suitable output in case a suitable intent for the user's inquiry is identified (Zierau et al., 2020). Ideally, CAs provide multiple advantages for both the organization and the user compared to human agents. On the one hand, CAs lead to profit increases due to their more efficient and less labour-intensive nature of CAs (Chung et al., 2020). On the other hand, users benefit from CAs due to faster response times and 24/7 availability (Chen et al., 2021). However, CAs also provide negative implications for users. Researchers have already called for more ethical design choices considering the user's privacy, network security, and information power redistribution to provide the user's dialogue intent (Murtarelli et al., 2021).

In the same vein, persuasive conversational designs may be adopted by CAs to manipulate user behavior in favour of the organization. For instance, CAs use statements that are unlikely to prompt users' disagreement (Schulman and Bickmore, 2009). Especially if CAs apply emotional speech, users are prone to be persuaded (Saunderson and Nejat, 2020). Moreover, research has shown that anthropomorphic characteristics in CAs increase the potential for persuasion (Pietrantoni et al., 2022, Diederich et al., 2020).

Nevertheless, the literature on the negative implications of CAs on users is still in its infancy (Diederich et al., 2022). Only one study investigated the ethical perspective of CA development by defining a measure for the perceived moral agency (Banks, 2019, Diederich et al., 2022). Thus, a closer investigation of ethical decision-making in the development process of CAs is of great importance as it has been established that CAs

can act as manipulators (Diederich et al., 2022) due to their capability of influencing user's behaviors, decisions, and affective reactions (Adler et al., 2016). This effect is widely known from human-computer interaction research, where the manipulative user interface design is also referred to as dark patterns (Gray et al., 2018).

2.2 Dark Patterns

The term dark patterns is characterized as "*user interface design elements that compromise user autonomy by preventing informed choices*" (Kollmer and Eckhardt, 2023, p. 202). Generally speaking, dark patterns manipulate and deceive users through fabrication, omission, complication, and composition of choices and information (Kollmer and Eckhardt, 2023). Thereby, dark patterns exploit psychological biases of the human brain (Hoch and Schkade, 1996). More precisely, to deal with the massive amount of choices in information systems, the human brain has an automatic system that performs immediate and instinctive choices (Osman, 2004, Barrett et al., 2004). This allows individuals to attribute more time to the reflective system, which implies self-conscious choices. Dark patterns subconsciously target the automatic system to alter individuals' choices (Thaler and Sunstein, 2008). To date, dark patterns are solely investigated in the context of user interface design characteristics of information systems (e.g., Mathur et al., 2021, Narayanan et al., 2020).

For instance, the leading social networking service Instagram utilizes a dark pattern concerning the activation of notifications. When individuals initially open the application, they are confronted with a modal dialogue to activate notifications. The user is only provided with the options "OK" and "Not Now". An option to permanently decline the notifications is absent. Thereby, individuals are steered into activating the notifications and consequently interact more frequently with Instagram (Gray et al., 2018). Generally, research has identified more than 100 different manifestations of dark patterns are already identified, and the number is constantly growing (Mathur et al., 2019). All these manifestations have in common that users are manipulated towards decisions or behaviors that favor organizational benefits at the expense of user benefits. Specifically, the organizational objectives are to generate profits, collect data, and draw users' attention (Narayanan et al., 2020).

From an organizational perspective, the creation of dark patterns occurs intentionally or inadvertently. Moreover, some dark patterns result from poor design choices, while others result from explicit design intentions (Gray et al., 2018). For instance, this involves increased complexity within the process of unsubscribing from newsletter subscriptions (Kollmer and Eckhardt, 2023)

The decisive attribute of a dark pattern is that the user's original intent is altered, and this causes users to select options they otherwise might not choose (Westin and Chiasson, 2021). The manipulative potency of dark patterns makes them attractive to many organizations, for it is no coincidence that about 11% of e-commerce websites already incorporate them (Mathur et al., 2019). However, technological progress is going to foster dark pattern applications in other systems like home robots, virtual reality, and, most importantly for the study at hand, in CAs (Lacey and Caudwell, 2019). In conjunction with the growing prevalence of CAs in our everyday lives, it is clear that we need to further our understanding of the usage of dark patterns within this

technology. Therefore, this work takes the first step by investigating the drivers for the integration of dark patterns in CAs.

3 Research Methodology

3.1 Data Collection

We recruited CA developers through Social Networks and professional communities. Overall, we conducted 11 interviews over a three-month period between April and June 2022. Each interview lasted between 45 and 90 minutes. Regarding our number of interviews, we are within the recommended range to ensure theoretical saturation (Thomson, 2010, Glaser and Strauss, 1967). In order to validate theoretical saturation, we utilized the last interview to test whether the identified categories and themes were sufficient (Marshall et al., 2013, Glaser and Strauss, 1967). Because no additional concepts were uncovered in the additional interview, we can confirm the theoretical validity of our data. Table 1 provides an overview of our interviewees.

Table 1. Interviewee Summary

Interviewee	Job-Level	CA Projects	Main Platform Experience
I1	Junior Developer	2 projects	Watson assistant
I2	Senior Developer	1 project	Amazon Alexa
I3	Scientist	2 projects	Multiple
I4	Senior Developer	6 projects	Multiple
I5	Senior Developer	3 projects	Self-developed
I6	Junior Developer	1 project	Watson assistant
I7	Senior Developer	7 projects	Multiple
I8	Junior Developer	6 projects	Google dialogue flow
I9	Consultant	2 projects	Self-developed
I10	Junior Developer	1 project	Open-Source
I11	Senior Developer	8 projects	Multiple

Concerning the sampling strategy, our general selection criteria for participation in our study was that the interviewees had experience in at least one software development project that involved the integration of CAs. In sum, on average, our interviewees were involved in 3,5 CA development projects. Likewise, our interviewees involve junior developers, senior developers, scientists, and consultants (see Table 1). All our interviewees have gained at least one year of professional experience after their last academic degree.

Additionally, we wanted to identify interviewees' expertise in multiple different technological platforms to develop a generalizable understanding of drivers for the integration of dark patterns in CAs. Our interviewees provide expertise in all major platforms, such as Watson Assistant, Amazon Alexa, Samsung Bixby voice assistant, and Google dialogue flow. Supplementary, two of our interviewees develop CAs by themselves without utilizing a pre-existing commercial platform. From a demographic standpoint, our sample involved three female and eight male interviewees aged between

22 and 41 years. Lastly, our objective was to generate unbiased insights about the causes of adopting dark patterns in CAs. We did not collect further identifiable information and anonymized identifiable information involving organizations, education, and professional experience outside of CA development in the subsequent transcription process.

3.2 Data Collection

For our data analysis, we followed the structured approach based on Strauss and Corbin (1997) and because of the explorative nature of our study. Table 2 depicts instances of coded segments found within the interviews.

Table 2. Coding Examples

Interviewee statement	Open Code	Sub Category	Core Category
"I mean, you have some data, but maybe not enough, but you still know how the user is talking at this specific domain. You actually also can collect data"	Data trade-off	Data collection potential	Technical driver
"So I guess a little bit of [time] pressure from upper management makes these choices that are implemented in the final step."	Stakeholder pressure	Time pressure	Organizational driver
"Yes, okay, we take the first one from the database. But then it's always just the first one and then maybe the others feel a little bit disadvantaged. Then maybe we have to give out a random one."	Random response	Deficient user understanding	Team driver

Open coding is utilized to discover concepts and categories within the interviews by naming and comparing similarities and differences within the collected interview data (Locke, 2000). Our open coding led to 341 codes in 16 subcategories and 10 categories. Subsequently, we continued with axial coding, linking the identified categories and subcategories (Matavire and Brown, 2013). Lastly, we selectively coded our data to reduce our categories, subcategories, and codes to the central concepts to answer our proposed research question regarding drivers of dark patterns in CAs (Strauss and Corbin, 1997). We excluded all codes that did not sufficiently relate to the three core categories (Glaser, 1992). For the data analysis of our interviews, we relied on two independent coders. After the selective coding, we calculated the resulting intercoder reliability using Cohen's kappa. The resulting value of 0,748 exceeds the recommended threshold by Landis and Koch (1977), which indicates substantial agreement between the coders.

4 Findings

4.1 Technical drivers

We found technological drivers of CAs that led to the integration of dark patterns. During our interviews, it became a predominant theme that the general role of the CA is likely to prompt dark patterns during the development process [I1-I4; I6-I11].

CAs are often applied in very specific areas, such as the initial error handling in customer support services. As a result, CAs are often only trained in the most common use cases. Consequently, CAs try to guide users into these use cases by asking suggestive questions within their knowledge base range, which can be perceived as dark patterns depending on the user's actual intention. Consequently, users are tied to predefined use cases, which can undermine their original intention. As interviewee 7 put it:

"[…] and if you now set up the CA in such a way that it always asks you a question and you are then forced to answer, - either with yes or no or something else - in principle already asking a bit of suggestive questions, then you can guide the customer exactly where you want."

In addition, the CAs are very fast and direct in the conversation, which may not always be in the organization's interest. CAs could directly respond and fulfil requests such as cancelling a membership or subscription. However, speeding up this process through CAs is not desirable from an organizational perspective. Because organizations try to prevent and avoid membership cancellations, such easy responses are intentionally not implemented in CAs. Instead, users are getting forwarded to an employee of the organization that is instructed to persuade the user to stay and to assess the reason behind the decision for the undesired action. Based on the exemplary cancellation process, interviewee 4 explained the advantages and reasons not to include direct responses to a cancellation request within the CAs:

"[…] I think it's much better to connect to a human so they can check like, "Oh, are you using the credit card correctly because it's for free? Maybe you are using it the wrong way". And if the agent is like, "Oh, [..] it's dangerous for them to have a credit card because they really do not understand it. They're going to cancel it. But I think this retention is good, but chatbots are not good at it. So, if you try to cancel anything with a chatbot they almost always reconnect you to a human."

Moreover, CAs' efficiency in fulfilling tasks seems especially advantageous if organization have the desires for the user to be active, such as, signing up for a membership. However, in cases where user inaction provides benefits (e.g., membership withdrawals), efficiency and speed are not what the organization is looking for. Instead, such procedures are often prolonged and overcomplicated to give users time to change their minds. Currently, CAs' capabilities are limited in differentiating between desirable and undesirable actions and acting accordingly. The usage of human agents is still warranted in some situations. In the same vein, CAs require training data to precisely understand users' intentions and to provide adequate responses. Therefore, collecting, storing, and utilizing vast amounts of user data is necessary. Although developers may improve the resulting user experience with the CA based on the collected data, our results suggest that data is also utilized for additional purposes, such

as, targeted advertisements and selling personal information. Generally, our interviewees highlighted the necessity of transparency regarding data collection. However, they also admitted that CAs incorporate dark patterns, such as, defaults within the data collection consent to utilize user data actively. Because, in the end, CAs, and their data collections are subject to organizational profit maximization.

4.2 Organizational drivers

A second theme we identified concerns the role of stakeholders and the organizational circumstances of the CA development project [I1-I4; I7-I10]. Generally, our interviewees highlight dominant stakeholders based on pre-existing conflicts of interest during the development process. While users and user experience designers aim for satisfying experiences, a marketing professional's objective is to provide tailored cross- and up-selling offers during the interaction. These practices may represent intentionally crafted dark patterns as they persuade users to purchase more expensive products or services. Moreover, within a diverse project team, the integration of dark patterns appears to be driven by the assertiveness of individual stakeholders. Similarly, stakeholders also pressure the project teams so that the CAs development follows their personal agenda. As interviewee 2 puts it:

"Well, there are 20 people sitting around a table and the one who shouts the loudest and is the most energetic, asserts his interests."

Currently, most CAs augment or substitute existing processes within organizations. Therefore, the organizational objective is to improve efficiency rather than provide an excellent user experience. As a result, the development and integration of CAs is only initiated when there is an underlying positive business case. Consequently, the development team is incentivized to adopt dark patterns within the CAs since it fosters profit increases for the organization. More precisely, our interviewees outlined that marketing departments foster dark patterns since they feel responsible for incorporating business objectives into CAs. Therefore, interactions may feature more marketing cues. As an example, interviewee 8 outlines the integration of further promotions of newsletter subscriptions at the end of regular conversations:

"I guess, marketing techniques where at the end of the message, you know, they'll say something like, "Hey, have you visited our new blog?" or, you know, "Do you want to subscribe to our newsletter?" or something like that."

As with many other application development projects, CA projects include strict deadlines for specific functionalities. Unlike in classical application development projects, the actual functionality of the CA is not easy to grasp because not all inputs and respective outputs can be easily showcased. Likewise, our interviewees mention that this often leads to requirement changes in the very late stages of the project. As a result, CAs are often first released with limited functionalities. This, in turn, fosters dark patterns in CAs, such as a wrong understanding of the user's intentions. Interviewee 10 illustrated this in the following statement:

"Currently, it's all just randomly selected and ranked. There's no particular logic to it. Unfortunately. Customers actually still want the functionality that those are ranked by preference."

4.3 Team drivers

Lastly, the team and its characteristics and experiences play an important role in applying dark patterns in CAs. Specifically, our interviewees highlighted two drivers on a team level [I1; I2; I5; I8, I11]. The first driver for the incorporation of dark patterns is that information about the user's needs is missing. Our interviewees reported that most CA projects are very organization-centric, so that, functionalities are prioritized and valued by clients, departments, and team members. Consequently, dark patterns are created inadvertently because of missing information about the user's needs within the team. Specifically, because CA developers primarily focus on the technical excellence of the CAs, such as accurate and precise responses. However, providing an accurate and precise response based on false assumptions about the user can also lead to incorporating dark patterns in CAs. Moreover, our results suggest that there is often no direct feedback loop from the end user to the development team, which makes it difficult for the CA development team to identify dark patterns. For instance, Interviewee 11 was working on a CA for a large hotel chain and described the following:

"We don't have to deal directly with the end user, i.e., the guest. We only get feedback from the hotels."

Secondly, we identified that deficient experience regarding the planning, process, and aftercare of CA development projects also fosters the integration of dark patterns. Specifically, the CA developers reported that little attention is paid to conversational design. This involves dealing with inquiries from a small fraction of users. Additionally, this also leads to poor error handling. For instance, dark patterns are created inadvertently when users ask a specific request that the CA cannot answer. Developers specifically decide if those requests are forwarded to a human agent, who can deal with the inquiry or are rather left unanswered by the CA. For instance, Interviewee 6 highlighted that the CA always answers with "sorry, I'm not sure how to answer this question", which represents a dark pattern as relevant information about how the user can proceed is omitted. In the same vein, the deficient experience leads to the integration of texts that are inappropriate for conversations. For instance, interviewee 2 outlines that he used to work on a project that utilized structured pre-existent website texts as input for the CA. However, the texts were very long, and therefore the answers of the CA took multiple minutes, which can be considered a dark pattern:

"And the […] approach was simply to take the texts from a website one-to-one from structured website data. […] nobody listens [to these texts] for ten minutes. And the texts themselves have to be designed editorially quite differently for CAs, as for, now for example, a website, yes? A book, a novel, is also written differently than a scientific treatise. So there also has to be something like audio texts for CAs […]"

5 Discussion

5.1 Drivers of Dark Patterns in CAs

The synthesis of our findings depicts an overview of dark pattern drivers in CAs. The conducted interviews allowed us to develop an overarching understanding of drivers for the integration of dark patterns in CAs. Our results indicate that the integration of dark patterns is fostered by technical drivers, organizational drivers, and team drivers. The nested visualization of the drivers for dark patterns in CAs results from the impact of these drivers.

In the highest order, we identified the technical drivers for the integration dark patterns in CAs. On the one hand, CAs are inclined to steer users through conversations since dialogue limitations hinder an open conversation simultaneously. Consequently, it is best practice to utilize suggestive questions within the CA design, which can be - if not sufficiently justified – manipulate user choices and therefore form dark patterns. On the other hand, CAs demand vast amounts of training data to function fully.

Our interviews revealed that CAs are more likely to collect sensitive user data than human agents; therefore, dark patterns that foster extensive data collection are adopted by the organization. In addition, our results suggest two organizational drivers. It became apparent that dominant stakeholders that enforce their interests, such as marketing professionals, impact the CA development process and drive the integration of dark patterns. In the same vein, CAs represent complex development projects that have the potential to miss important project deadlines due to late requirement changes, changes within the team, and time delays.

Moreover, our interviewees indicated that time pressure during CA development leads to the integration of dark patterns because the taken shortcuts during the development. Lastly, our findings depict that team characteristics play an essential role in considering the integration of dark patterns. Specifically, a deficient user understanding can lead to dark patterns. This involves missing feedback loops and non-existent user research. Furthermore, our interviewees outline that lack of experience regarding CA development fosters the incorporation of dark patterns because of the unawareness of technical challenges and the user interaction with the CAs (see Figure 1).

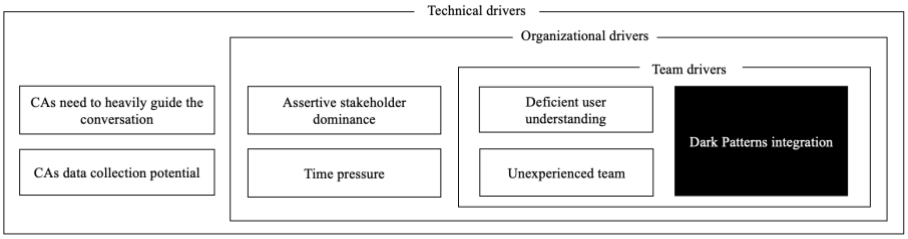

Figure 1. Drivers of Dark Patterns in CAs

5.2 Implications for Research

Our study provides several implications for research. Thereby, we contribute to the existing literature on CAs. As outlined by Diederich et al. (2022), the existing research on CAs lacks an ethical perspective. Our results further amplify the need for ethical discourses regarding the design and development of CAs. Furthermore, we identified multiple drivers for unethical CA designs, also called dark patterns (Mathur et al., 2021). Moreover, we contribute to existing research regarding deceptiveness as a conversational skill of CAs (Schuetzler et al., 2019).

We identified the technical drivers of CAs as a source of the following dilemma. Users require guidance to interact with CAs appropriately. Therefore, psychological principles within user interface designs have been adopted and proven successful over 25 years (Hix and Hartson, 1993). In contrast, our study indicates that, especially in the context of CAs, heavy guidance can lead to adverse consequences for the users. For instance, suggestive questions can steer the users' decision-making in a for them subsequently potentially harmful and undesired direction. Furthermore, we found that stakeholder involvement drives the incorporation of dark patterns in CAs. While classical information systems development attributes the primary responsibility for dark patterns to the interface designer within the team, we showcase that stakeholder involvement mainly drives their integration in CAs (Gray et al., 2021). Surprisingly, the interviewees barely mentioned the role of a conversational designer, indicating that conversational design's role within CA development is still largely unexplored. Additionally, we contribute to existing information systems development literature by indicating that time pressure fosters the integration of dark patterns. In turn, we contribute to existing evidence that time pressure leads to quality issues by incorporating the potentially manipulative and deceptive implications of dark patterns (Austin, 2001).

In addition, we enrich the understanding of team characteristics that lead to dark patterns in CAs in the first place. Our findings align with existing research by indicating that dark patterns can be prevented with the appropriate experience and ethical responsibility of those involved in the development team (Gray et al., 2018). Moreover, our results suggest that regardless of the platform, the drivers of dark patterns in CAs are similar. Consequently, we show that dark patterns and their drivers are a concerning theme across all CA development projects, which demonstrates the severe challenge of dark patterns across all platforms. Existing research proposes several design principles for CAs that can act as countermeasures for adopting dark patterns. This involves proactive guidance of the user during the conversation (Diederich et al., 2020, Feine et al., 2019b) and functional transparency about the capabilities of the CA (Strohmann et al., 2019). Nevertheless, existing design principles are only attributed to the characteristics of CAs. Our study pinpoints explicitly the essential role of organizational and team drivers that foster dark patterns. Thus, we recommend future research to actively investigate design principles that affect the development process to mitigate dark patterns. An example of a design principle within the development process is to actively conduct user research in all stages of the development process.

5.3 Implications for Practice

Besides the provided implications for research, our study also provides fruitful implications for practitioners. Dark patterns have increasingly infiltrated information systems over the last few years (Mathur et al., 2019). In response, governmental institutions first introduce regulations to prevent dark patterns (Akhtar, 2021). However, the regulations only address specific manifestations of dark patterns in user interface designs. Concerning CAs, existing regulations only partially target the integration of dark patterns. For instance, in the European Union introduced the general data protection regulation (GDPR) to reduce and regulate the collection of personal information by organizations (European Parliament, 2016). In turn, this affects the technical driver of data collection potential in CAs. Nevertheless, our study suggests that additional regulations are needed to prevent the incorporation of dark patterns within CAs. Within our study, we identified six independent drivers for adopting dark patterns in CAs. Instead of focusing on the integration dark patterns, governments might emphasize the causes of dark patterns as well to generate a more substantial regulation.

Furthermore, our study demonstrates the importance of the right team setup and skillset for the user-centered development of CAs. Our results suggest that experts in conversational design, such as experienced user experience designers, desperately need to prevent dark patterns in CAs. In addition, the overall team setup and autonomy are essential. Our study shows that other departments and stakeholders often influence CA development to have adverse and dark outcomes because of alternative objectives.

Only when the CA development team is independent enough to prevail the incorporation of dark patterns in CA can be prevented. On the one hand, dark patterns in CAs can lead to short-term benefits for the organization, such as increased revenues, additional data collection, and steering users' attention (Narayanan et al., 2020). On the other hand, dark patterns can lead to adverse long-term consequences due to the trust loss of the user (Narayanan et al., 2020). In the same vein, we identified that experience in CA development influences the likelihood of adopting dark patterns. Therefore, we recommend that practitioners involve experienced professionals within the CA development team. Additionally, our study further emphasizes the need to practice user research actively and to collect constant feedback from the users to avoid permanent dark patterns in the CAs. In sum, organizations need to prevent dark patterns in their applications. Our study guides practitioners by depicting organizational and team drivers for the integration of dark patterns in CAs.

5.4 Limitations and Future Research Directions

Although our study provides valuable insights regarding the drivers of dark patterns in CAs, there are still some limitations in the selected research approach that offer valuable avenues for future research. First, for our data collection, we conducted a semi-structured interview. Although this qualitative approach enabled us to gather a unified understanding of dark pattern drivers, it comes with the following limitation. Due to the existent COVID-19 restrictions, we had to conduct the interviews online, limiting the assessment of the interviewees' body language.

Furthermore, our proposed model was derived from qualitative interviews, and it is crucial for future research to conduct subsequent quantitative studies to address these limitations. For instance, our study could be used to derive hypotheses for a quantitative survey that could further explain and validate the drivers of dark patterns in CAs.

Lastly, our research was motivated by explicitly identifying the drivers for adopting dark patterns in CAs. However, dark patterns will arise in other novel technological advancements, such as virtual reality and the metaverse (Wohlgenannt et al., 2020, Spiekermann et al., 2022). While we only explain drivers for adopting dark patterns in CA, future research needs to examine the transferability of our identified drivers to prevent dark patterns in other technologies.

6 Acknowledgements

The present study was funded by the Tiroler Wissenschaftsförderung (F.45047/8-2022, “Mitigation of Dark Patterns in Socio-Technical Systems”).

References

Adler, R. F., Iacobelli, F. and Gutstein, Y. (2016). Are you convinced? A Wizard of Oz study to test emotional vs. rational persuasion strategies in dialogues. *Computers in Human Behavior,* 57**,** 75-81.

Akhtar, A. (2021). *California is banning companies from using 'dark patterns,' a sneaky website design that makes things like canceling a subscription frustratingly difficult*. URL: https://www.businessinsider.com/what-are-dark-patterns-2021-3 (visited on 16.11.2022).

Araujo, T. (2018). Living up to the chatbot hype: The influence of anthropomorphic design cues and communicative agency framing on conversational agent and company perceptions. *Computers in Human Behavior,* 85**,** 183-189.

Austin, R. D. (2001). The effects of time pressure on quality in software development: An agency model. *Information systems research,* 12**,** 195-207.

Banks, J. (2019). A perceived moral agency scale: development and validation of a metric for humans and social machines. *Computers in Human Behavior,* 90**,** 363-371.

Barrett, L. F., Tugade, M. M. and Engle, R. W. (2004). Individual differences in working memory capacity and dual-process theories of the mind. *Psychological bulletin,* 130**,** 553.

Chen, J. V., Le, H. T. and Tran, S. T. T. (2021). Understanding automated conversational agent as a decision aid: matching agent's conversation with customer's shopping task. *Internet Research*.

Chin, H. and Yi, M. Y. (2021). Voices that Care Differently: Understanding the Effectiveness of a Conversational Agent with an Alternative Empathy Orientation and Emotional Expressivity in Mitigating Verbal Abuse. *International Journal of Human–Computer Interaction***,** 1-15.

Chung, M., Ko, E., Joung, H. and Kim, S. J. (2020). Chatbot e-service and customer satisfaction regarding luxury brands. *Journal of Business Research,* 117**,** 587-595.

Ciechanowski, L., Przegalinska, A., Magnuski, M. and Gloor, P. (2019). In the shades of the uncanny valley: An experimental study of human–chatbot interaction. *Future Generation Computer Systems,* 92**,** 539-548.

De Cicco, R., Silva, S. and Alparone, F. R. (2021). "It's on its way": Chatbots applied for online food delivery services, social or task-oriented interaction style? *Journal of Foodservice Business Research,* 24**,** 140-164.

Diederich, S., Brendel, A. B. and Kolbe, L. M. (2020). Designing anthropomorphic enterprise conversational agents. *Business & Information Systems Engineering,* 62**,** 193-209.

Diederich, S., Brendel, A. B., Morana, S. and Kolbe, L. (2022). On the design of and interaction with conversational agents: An organizing and assessing review of human-computer interaction research. *Journal of the Association for Information Systems,* 23**,** 96-138.

European Parliament (2016). 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).

Feine, J., Gnewuch, U., Morana, S. and Maedche, A. (2019a). Gender bias in chatbot design. *International Workshop on Chatbot Research and Design*, 2019a. Springer, 79-93.

Feine, J., Gnewuch, U., Morana, S. and Maedche, A. (2019b). A taxonomy of social cues for conversational agents. *International Journal of Human-Computer Studies,* 132**,** 138-161.

Glaser, B. and Strauss, A. (1967). The constant comparative method (Chapter 5). From The discovery of grounded theory. Chicago: Aldine.

Glaser, B. G. (1992). *Basics of grounded theory analysis: Emergence vs forcing*, Sociology press.

Gray, C. M., Chen, J., Chivukula, S. S. and Qu, L. (2021). End user accounts of dark patterns as felt manipulation. *Proceedings of the ACM on Human-Computer Interaction,* 5**,** 1-25.

Gray, C. M., Kou, Y., Battles, B., Hoggatt, J. and Toombs, A. L. (2018). The dark (patterns) side of UX design. *Proceedings of the 2018 CHI conference on human factors in computing systems*, 2018. 1-14.

Harjunen, V. J., Spapé, M., Ahmed, I., Jacucci, G. and Ravaja, N. (2018). Persuaded by the machine: The effect of virtual nonverbal cues and individual differences on compliance in economic bargaining. *Computers in Human Behavior,* 87**,** 384-394.

Hix, D. and Hartson, H. R. (1993). *Developing user interfaces: ensuring usability through product & process*, John Wiley & Sons, Inc.

Hoch, S. J. and Schkade, D. A. (1996). A psychological approach to decision support systems. *Management Science,* 42**,** 51-64.

Hummel, D. and Maedche, A. (2019). How effective is nudging? A quantitative review on the effect sizes and limits of empirical nudging studies. *Journal of Behavioral and Experimental Economics,* 80**,** 47-58.

Johnson, D. (2022). *How to turn off Siri on your iPhone and prevent the virtual assistant from listening to you*. URL: https://www.businessinsider.com/how-to-turn-off-siri (visited on 14.06.2022).

Kollmer, T. and Eckhardt, A. (2023). Dark Patterns: Conceptualization and Future Research Directions. *Business & Information Systems Engineering,* 65**,** 201-208.

Kundinger, T., Wintersberger, P. and Riener, A. (2019). (Over) Trust in automated driving: The sleeping pill of tomorrow? *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019. 1-6.

Lacey, C. and Caudwell, C. (2019). Cuteness as a 'dark pattern'in home robots. *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2019. IEEE, 374-381.

Landis, J. R. and Koch, G. G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics***,** 363-374.

Lehto, T. and Oinas-Kukkonen, H. (2015). Examining the persuasive potential of web-based health behavior change support systems. *AIS Transactions on Human-Computer Interaction,* 7**,** 126-140.

Locke, K. (2000). Grounded theory in management research. *Grounded Theory in Management Research***,** 1-160.

Luger, E. and Sellen, A. (2016). " Like Having a Really Bad PA" The Gulf between User Expectation and Experience of Conversational Agents. *Proceedings of the 2016 CHI conference on human factors in computing systems*, 2016. 5286-5297.

Maedche, A., Legner, C., Benlian, A., Berger, B., Gimpel, H., Hess, T., Hinz, O., Morana, S. and Söllner, M. (2019). AI-based digital assistants. Business & Information Systems Engineering, 61 (4), 535–544.

Marshall, B., Cardon, P., Poddar, A. and Fontenot, R. (2013). Does sample size matter in qualitative research?: A review of qualitative interviews in IS research. *Journal of computer information systems,* 54**,** 11-22.

Matavire, R. and Brown, I. (2013). Profiling grounded theory approaches in information systems research. *European Journal of Information Systems,* 22**,** 119-129.

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

Mathur, A., Kshirsagar, M. and Mayer, J. (2021). What makes a dark pattern... dark? Design attributes, normative considerations, and measurement methods. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021. 1-18.

Murtarelli, G., Gregory, A. and Romenti, S. (2021). A conversation-based perspective for shaping ethical human–machine interactions: The particular challenge of chatbots. *Journal of Business Research,* 129**,** 927-935.

Narayanan, A., Mathur, A., Chetty, M. and Kshirsagar, M. (2020). Dark Patterns: Past, Present, and Future: The evolution of tricky user interfaces. *Queue,* 18**,** 67-92.

Osman, M. (2004). An evaluation of dual-process theories of reasoning. *Psychonomic bulletin & review,* 11**,** 988-1010.

Pietrantoni, N., Greulich, R. S., Brendel, A. B. and Hildebrandt, F. (2022). Follow Me If You Want to Live-Understanding the Influence of Human-Like Design on Users' Perception and Intention to Comply with COVID-19 Education Chatbots.

Rheu, M., Shin, J. Y., Peng, W. and Huh-Yoo, J. (2021). Systematic review: Trust-building factors and implications for conversational agent design. *International Journal of Human–Computer Interaction,* 37**,** 81-96.

Saunderson, S. and Nejat, G. (2020). Investigating strategies for robot persuasion in social human-robot interaction. *IEEE Transactions on Cybernetics*.

Schuetzler, R. M., Grimes, G. M. and Giboney, J. S. (2019). The effect of conversational agent skill on user behavior during deception. *Computers in Human Behavior*, 97**,** 250-259.

Schuetzler, R. M., Grimes, G. M., Giboney, J. S. and Rosser, H. K. (2021). Deciding Whether and How to Deploy Chatbots. *MIS Quarterly Executive*, 20.

Schulman, D. and Bickmore, T. (2009). Persuading users through counseling dialogue with a conversational agent. *Proceedings of the 4th international conference on persuasive technology*, 2009. 1-8.

Song, C. S. and Kim, Y.-K. (2022). The role of the human-robot interaction in consumers' acceptance of humanoid retail service robots. *Journal of Business Research*, 146**,** 489-503.

Spiekermann, S., Krasnova, H., Hinz, O., Baumann, A., Benlian, A., Gimpel, H., Heimbach, I., Köster, A., Maedche, A. and Niehaves, B. (2022). Values and Ethics in Information Systems. *Business & Information Systems Engineering*, 64**,** 247-264.

Strauss, A. and Corbin, J. M. (1997). *Grounded theory in practice*, Sage.

Strohmann, T., Höper, L. and Robra-Bissantz, S. (2019). Design guidelines for creating a convincing user experience with virtual in-vehicle assistants. *In: Proceedings of the 52nd Hawaii International Conference on System Sciences, ,* 11**,** 54–78.

Thaler, R. H. and Sunstein, C. R. (2008). Nudge: improving decisions about health. *Wealth, and Happiness*, 6**,** 14-38.

Thomson, S. B. (2010). Grounded theory-sample size. *Journal of Administration and Governance*, 5**,** 45-52.

Westhoven, M. and Tegtmeier, P. (2021). Influence of Personality Traits on Helping Behaviour in Human-Robot Interaction. *2021 IEEE International Conference on Advanced Robotics and Its Social Impacts (ARSO)*, 2021. IEEE, 78-84.

Westin, F. and Chiasson, S. (2021). "It's So Difficult to Sever that Connection": The Role of FoMO in Users' Reluctant Privacy Behaviours. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021. 1-15.

Wohlgenannt, I., Simons, A. and Stieglitz, S. (2020). Virtual reality. *Business & Information Systems Engineering*, 62**,** 455-461.

Xu, K. and Lombard, M. (2017). Persuasive computing: Feeling peer pressure from multiple computer agents. *Computers in Human Behavior*, 74**,** 152-162.

Yang, X. and Aurisicchio, M. (2021). Designing Conversational Agents: A Self-Determination Theory Approach. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021. 1-16.

You, S. and Robert, L. (2018). Trusting robots in teams: Examining the impacts of trusting robots on team performance and satisfaction. *Proceedings of the 52th Hawaii International Conference on System Sciences, Jan*, 2018. 8-11.

Zierau, N., Wambsganss, T., Janson, A., Schöbel, S. and Leimeister, J. M. (2020). The anatomy of user experience with conversational agents: a taxonomy and propositions of service clues. *International Conference on Information Systems (ICIS).-Hyderabad, India*, 2020.

Złotowski, J., Sumioka, H., Nishio, S., Glas, D. F., Bartneck, C. and Ishiguro, H. (2016). Appearance of a robot affects the impact of its behaviour on perceived trustworthiness and empathy. *Paladyn, Journal of Behavioral Robotics*, 7.