Computers as Bad Social Actors: Dark Patterns and Anti-Patterns in Interfaces that Act Socially

LIZE ALBERTS, University of Oxford, United Kingdom and Stellenbosch University, South Africa
ULRIK LYNGS, University of Oxford, United Kingdom
MAX VAN KLEEK, University of Oxford, United Kingdom

Interfaces increasingly mimic human social behaviours. Beyond prototypical examples like chatbots, basic automated systems like app notifications or self-checkout machines likewise address or 'talk to' people in person-like ways. Whilst early evidence suggests social cues can enhance user experience, we lack a good understanding of when, and why, their use in interaction design may be inappropriate. We combined a qualitative survey (n=80) with experience sampling, interview, and workshop studies (n=11) to understand people's attitudes and preferences regarding how a range of automated systems talk to/at them. We thematically analysed examples of phrasings or conduct our participants disliked, their reasons, and how they would prefer to be treated instead. One category of inappropriate use we identified is when social design elements are used to manipulate user behaviour. We distinguish four such tactics: 'agents' playing on users' emotions (e.g., guilt-tripping, coaxing), being pushy, mothering users, or being passive-aggressive. Another category regards pragmatics: personal or situational factors that can make even a seemingly helpful or friendly message come across as rude, tactless, invasive, etc. These include contextual insensitivity (e.g., embarrassing users in public); expressing clearly false personalised care; or treating a user in ways they find misaligned with the system's role or the nature of their relationship. We discuss these inappropriate uses in terms of an emerging 'social' class of dark and anti-patterns. From participant suggestions, we offer recommendations for improving how interfaces treat people in interaction, including broader normative reflections on treating users respectfully.

CCS Concepts: • **Human-centered computing** → **Natural language interfaces**; **Empirical studies in interaction design**; **User centered design**; *Contextual design*; **Interaction paradigms**; • **Computing methodologies** → **Discourse, dialogue and pragmatics**.

Additional Key Words and Phrases: Dark Patterns; Deceptive Patterns; Computers Are Social Actors; Dialogue Agents; Conversational User Interface; Manipulation; Social Engineering; App Notifications; Chatbots; Respect; Interactional Ethics; Mixed Qualitative Methods

ACM Reference Format:
Lize Alberts, Ulrik Lyngs, and Max Van Kleek. 2024. Computers as Bad Social Actors: Dark Patterns and Anti-Patterns in Interfaces that Act Socially. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 202 (April 2024), 25 pages. https://doi.org/10.1145/3653693

1 INTRODUCTION

Automated systems increasingly interact with people as if they were thinking/feeling agents. This ranges from sophisticated (digital, virtual, or embodied) conversational agents (CAs) that maintain open-ended dialogues, to basic automated systems like app notifications or self-checkout machines

Authors' addresses: Lize Alberts, lize.alberts@cs.ox.ac.uk, University of Oxford, Department of Computer Science, Oxford, United Kingdom, OX1 3QG and Stellenbosch University, Department of Philosophy, Stellenbosch, South Africa, 7600; Ulrik Lyngs, ulrik.lyngs@cs.ox.ac.uk, University of Oxford, Department of Computer Science, Oxford, United Kingdom, OX1 3QG; Max Van Kleek, max.van.kleek@cs.ox.ac.uk, University of Oxford, Department of Computer Science, Oxford, United Kingdom, OX1 3QG.

ACM 2573-0142/2024/4-ART202
https://doi.org/10.1145/3653693

that frame messages as if they were a person directly addressing the user. Whereas prior research has highlighted some benefits of socially conforming, anthropomorphic interfaces, such as fostering a more intuitive and engaging user experience [1, 49] and increasing user trust [55], we do not yet have a good understanding of factors that can make the use of different social cues inappropriate. Beyond emerging concerns regarding agents based on large language models (LLMs) generating biased, toxic or misleading language [5, 57, 66], there is a lack of empirical data on how pragmatic factors, like the broader social context and other interactional particulars, affect how a given automated social action is received.

In this paper, we conducted a preliminary investigation into (a) the misuse and abuse of social cues in interfaces, (b) how users prefer to be spoken to, and (c) what it may mean for automated systems to treat them appropriately and respectfully in interaction. For this, we look beyond the more prototypical conversational user interfaces (CUIs) like chatbots, voice-assistants, and social robots, to any interface that behaves as a social actor: using social cues (e.g., affective expressions, first-person pronouns, etc.), or merely 'saying things' (e.g., in text, voice, gesture), as if addressing the user. We use the term *social interfaces*[1] to refer to this broader class.

Our first concern is the use of social cues as tools for manipulation. When interfaces mimic patterns in natural human-human social interaction, they encourage users to respond in social or emotional ways—as if it were a person talking to them—by tapping into known social/anthropomorphic biases [48, 49, 54]. In doing so, behavioural designers may unduly influence people's behaviour, in ways that bypass their more conscious/deliberate ('System 2' [30]) thinking, into doing things they may have otherwise been reluctant to. That is, functioning similarly to the infamous *dark patterns* in interaction design [40, 43, 44, 47].[2] So far, most of the dark patterns literature has focused on graphic user interface (GUI) elements (e.g., in shopping websites or cookie consent banners) that subtly 'trick', seduce, or pressure users into behaviours that serve other stakeholders' interests, with limited consideration of other interface modalities [26, 32, 51]. We characterise an emerging 'social' class of dark patterns that may appear in a range of emerging social interface types, be it graphic, voice-based, virtual or embodied.

Such socially manipulative tactics are already appearing, most commonly in smartphone notifications. Smartphone apps increasingly frame their notification messages in conversational (person- or even friend-like) ways to encourage users to stick to their goals, complete tasks, buy products, or otherwise increase app engagement. Some even pressure users by eliciting guilt or empathy, using emotive language like "I miss you", or seeming annoyed at users for "ignoring" them. Whilst using social cues in interaction design is not necessarily bad or unethical, there is potential for it to be exploited. We use 'social dark patterns' to refer to the inappropriate use of socially manipulative tactics: where designers try to manipulate people towards certain ends by encouraging them to treat an interface as an intentional agent—whose feelings and judgment they should care about—rather than an inanimate platform.

Beyond manipulation, we also consider the risk of social interfaces failing to exhibit appropriate nuance for a given social situation (which is notoriously difficult[3]) when they proactively 'speak to' people in everyday settings. Even a seemingly innocuous message may appear invasive, tactless, or even offensive in the wrong context—especially if it is scripted, repetitive or generic, and made to appear as if coming from an agent specifically addressing the user. Following the common

[1]This is not to be confused with the use of 'social interface' in development sociology as "a critical point of intersection between different lifeworlds, social fields or levels of social organization" [37, p.243]. Here, 'social' describes the behaviour of the interface itself.

[2]Broadly, these are tactics that apply insights from behavioural psychology to steer users' behaviour towards promoting another stakeholder's interests.

[3]See Suchman's discussion of interaction as the "contingent coproduction of a shared sociomaterial world" [59, p.23].

distinction between dark and *anti-patterns* [25, 42] (i.e., interface elements that fail to work as intended), we use 'social anti-patterns' to describe social interface elements that bother users for reasons the designers failed to anticipate, particularly for seeming somehow socially inappropriate.

Driven by these concerns, we used a combination of qualitative methods to explore the following research questions with end-users:

- **RQ1**: When do automated social behaviours seem manipulative or pressuring, and why?
- **RQ2:** When do automated social behaviours seem otherwise off-putting or inappropriate, and why?
- **RQ3:** How do people prefer automated systems "speak" to them?
- **RQ4:** What contextual factors affect their preferences?

Our study consisted of four complementary phases [41, 62]. The first was an exploratory survey (n=80) with adult English-speaking smartphone users, prompting them to describe occasions when they disliked how automated systems spoke to them. This was followed by three studies, using a subset of participants (n=11) from the first: (1) a week-long experience sampling study, capturing *in-situ* examples of notification messages that spoke to people in ways they disliked, annotated with details of what about them was undesirable, and why; (2) individual semi-structured interviews, allowing participants to elaborate on nuances in their preferences, and (3) a group-based exploratory workshop (n=3-4) with activities exploring variation in preferences between application domains. This combination of methods was designed to help us triangulate and refine our findings: whereas the first phase allowed us to gain an initial understanding of stand-out negative experiences across social interface types, the latter three helped us gain a better understanding of contextual, individual and domain variability.

Using an integrated, reflective approach to (iteratively) analyse our data into codes and themes [10, 50], we arrived at three sets of findings. The first set concerns the manipulative use of social cues in interface design, i.e., as dark patterns (RQ1). The second set regards counterproductive uses of social design elements, constituting social anti-patterns (RQ2). The final set regards participant suggestions for improved social acting (RQ3, RQ4), where we describe themes generated from participant suggestions regarding how they would prefer, and believe they deserve, to be treated by interfaces that "talk" to or at them.

Our primary contributions can be summarised as follows:

(1) We contribute to the literature on nudging and dark patterns by identifying and characterising an emerging *social* class of dark patterns, and how they may look on a range of interface modalities.
(2) We use our findings to critically engage with research in the Computers Are Social Actors paradigm by proposing when, and why, the use of social cues in interface design may be counterproductive—constituting 'social' anti-patterns.
(3) Based on our analysis of end-user preferences and suggestions, we offer recommendations to help prevent undesirable forms of automated social acting.
(4) We integrate normative considerations regarding how users feel they *deserve* to be treated by automated systems as specific duties for designers: laying the foundations for a framework unpacking what it means for interfaces to treat people respectfully in interaction.

Our findings highlight important factors that can affect the perceived appropriateness of automated social behaviours. The first is when an interface appears to use social/emotionally manipulative tactics: from our participants' examples and descriptions, we identify four such tactics that already appear in interface design. These are *agents playing on emotions* (e.g., guilt-tripping, coaxing, or eliciting empathy), *agents being pushy* (e.g., dictating, pressuring or nagging), *agents*

mothering (e.g., behaving like a concerned parent), and *agents being passive-aggressive* (e.g., conveying dissatisfaction or judgement)—all of which can undermine users' sense of autonomy. We also identify a range of factors that can make seemingly innocuous or well-meaning social behaviours seem inappropriate. This includes embarrassing users by addressing them in public spaces, or recommendations that, for contextual reasons, appear to be mocking users' weaknesses or insecurities. We also found that users can find certain social pleasantries off-putting for coming across as insincere, especially when incentives are obviously self-interested or messages generic. Moreover, using casual, friend-like language can likewise be off-putting when it feels misaligned with the agent's perceived role or relationship with the user.

Our work contributes to the literature on users' experiences of CUIs, particularly regarding expectation violations [29, 38, 45], as well as emerging research in Human-AI Interaction [3, 36, 65]. It also contributes to recent work anticipating harms related to LLM-based dialogue agents [5, 57, 66], which are expected to play increasingly active social roles in people's daily lives. However, ours is the first paper to treat, in a unified manner, the expanding set of systems that "speak" to and address users. Whilst prior work has mainly focused on the more overtly human-like versions of these, such as chatbots or social robots, our treatment admits a much larger set of interfaces that, while perhaps less socially capable, nonetheless create social situations through their communicative actions and use of social protocol. In doing so, we argue that they introduce kinds of risks that are best understood in social terms: as social actors failing to meet what someone might expect from a socially capable agent in the given social situation.

With the conceptual and empirical contributions of this paper, we hope to inspire a focused discussion on how automated systems treat people in interactions. This work forms part of a larger project to develop an ethics of interaction for mixed-initiative systems, centred on the concept of respect [2].

2 BACKGROUND AND RELATED WORK

2.1 When is social acting (in)appropriate?

The Computers Are Social Actors (CASA) research paradigm is centred on understanding people's anthropomorphic tendencies when interacting with computers. Starting with a series of simple lab studies in the 90s, it developed from the finding that people apply social norms and expectations (or *scripts* [54]) to their interactions with technology, even without being prompted by obvious anthropomorphic elements [48, 49].[4] According to Reeves and Nass [54], these responses occur even when people do not have corresponding beliefs that justify their behaviours, or offer rational arguments for why it would be silly to—they seem to apply such scripts *mindlessly*.

Complementing CASA, findings in neuroscience suggest that cues to human-like animacy can strengthen people's anthropomorphic tendencies, as they engage brain networks associated with social cognition. These can be bottom-up (e.g., how an entity looks/behaves), as well as top-down (e.g., beliefs of human-like commonalities) [19]. In interface design, common examples are *identity cues* (e.g., giving a virtual agent a name or face), *non-verbal cues* (e.g., affective speech, gendered voices or pausing as if thinking), and *verbal cues* (e.g., human-like mannerisms in responses) [27].[5] Such attributes help to make the system appear more like an intentional 'agent' speaking to the

[4]CASA is a specific application of the *Media Equation* [54], the idea that people behave as if they treat "mediated" life as real life; interacting with communication technologies in fundamentally social ways. This may include treating computers as if they have folk-psychological states (e.g., beliefs, intentions, or desires), or reacting to moving pictures on a screen as if the events were taking place in real life [48].

[5]Other (top-down) cues include personified descriptions of AI systems: beyond the countless AI personifications that populate the media (including the term *AI* itself), CUIs like social robots are often explicitly marketed as "friends" and "your next family member" that "can't wait to meet you" [22, p.60].

user, rather than a medium for people to communicate with each other [24], strengthening the user's social/emotional responses.

However, heightening expectations of a system's social capabilities, and then falling short of them, may severely harm users' overall experience and satisfaction [15, 39]. Moreover, there is a growing body of research on individual factors (e.g., users' cognitive styles [34], personality traits [31], age [23, 33]) and contextual factors (e.g., the domain of application [60]) that may affect users' evaluation of social interfaces,[6] raising the question of what social cues are desirable, and in which contexts. Using certain social cues in the wrong situations may even have backfiring effects, e.g., making angry customers angrier [29, 38], or stressed users more stressed [45]. In an analysis of 500 customer reviews of chatbot apps, Svikhnuskina *et al.* [61] found that, whilst current chatbots offer certain benefits to users, they fail to meet various contextual expectations, including repeating responses, going off-topic, and being perceived as "rude". Moreover, they found frequent mentions of chatbots being "unnaturally supportive" to the point of discomfort.

Whilst studies like these point to factors that may render social cues ineffective, undesirable and even offensive in certain contexts, studies of users' experience with social interfaces are generally limited in terms of sampling and ecological validity (e.g., using lab-based/wizard-of-oz techniques). User experience (UX) studies with conversational systems also tend to focus on the usability of specific systems (and often focus on basic metrics like 'enjoyment' or 'satisfaction' [1]). Instead, we wanted to delve deeper into users' more general experience with the range of social interfaces they encounter in their daily lives, even ones that are not typically considered 'conversational'.

Beyond social acting generally, we were also interested in the potentially inappropriate use of social cues in behavioural design. To lay the groundwork for this, the next subsection briefly introduces the concepts of nudging, dark patterns, and anti-patterns.

2.2 Distinguishing nudges, dark and anti-patterns

2.2.1 Nudges and dark patterns. Behavioural design involves applying knowledge of how certain (psychological, social, material) factors influence how people respond to elements in interface design. In the case of nudges and dark patterns, this involves knowledge of people's general cognitive heuristics and biases,[7] which designers may either try to counter (by prompting careful reflection) or harness (utilising biases strategically). Both typically involve manipulating choice architectures—how options are framed or presented—such that certain options are more likely to be chosen. The key difference is that nudges are typically aimed at promoting the interests of the person being nudged (i.e., *the user*), whereas dark patterns typically aim to serve other stakeholders.

However, this boundary can be fuzzy. Originally, *nudging* was proposed as an unobtrusive and ethical form of paternalism, "without forbidding any options or significantly changing their economic incentives" [35, p.6]. This idea has been applied broadly in HCI, from preventing unwanted mistakes to helping users stick to their goals (see Caraban's [13] review). Yet, the term is increasingly used to describe a wider range of strategies for 'beneficent' behaviour manipulation, including more or less overt/intrusive, and sometimes controversial means [9, 20]. This can involve inducing fear, using deceptive tactics, or invoking a sense of shame or social pressure towards some desired end [13, 21]. Dark (design) patterns,[8] on the other hand, may harness various nudging (or even coercive) tactics to get users to do things they may otherwise have actively avoided [40, 43]. This typically

[6] For example, in a focus group exploring older adults' perspectives on social robot carers/companions, Laitinen *et al.* [33] found that several participants expressed concern that toy-like robots may be perceived as patronising/infantilising, a form of stigmatising that older adults already struggle with.

[7] That is, systematic deviations from rational judgment in human behaviour [63].

[8] These are also sometimes called *deceptive patterns,* to avoid racial connotations. However, as the tactics we consider do not necessarily involve deception, we find this term misleading.

involves inducing some misleading mental model by modifying choice architectures—*manipulating the decision space* (e.g., by placing unequal burdens on choices or strategically omitting options) and/or the *flow of information* (e.g., by deceptive framing or hiding key information), thereby [44]—encouraging users to make certain inferences and take certain actions.

'Dark pattern' has been used to describe multiple related, but not collectively required, aspects of objectionable interfaces [44]. Whilst some characterise it in terms of the *intention* to exploit users towards self-interested goals [17, 25] (contrary to 'beneficent' nudging), others find it objectionable for reasons beyond intention. According to Mathur *et al.*'s [44] review, this includes facts about the *interface* (e.g. being deceptive, coercive, or manipulative), the *mechanisms* of influence (e.g., subverting user intent or preferences), and the *effects* of the interface design (e.g., benefiting other stakeholders and/or harming users).

With these more nuanced characterisations, it is not always clear when a nudge counts as a dark pattern, as even well-intentioned forms of behavioural design (e.g., reminders to eat healthily) can meet some of these criteria (e.g., harming users by shaming them, or undermining their sense of autonomy). Moreover, intentions are rarely clear-cut,[9] and what is in someone's 'best interest' is often a matter of framing. In consideration of these ambiguities, we use the term 'dark patterns' in a broad sense, regardless of intention, to refer to design patterns that involve inducing certain emotions (e.g., guilt, fear, shame) or misleading mental models (e.g., a misconception of the system or their actions) to manipulate user behaviour towards specific ends—involving mechanisms or goals that the user is unaware of or did not consent to.[10]

As mentioned earlier, most of the dark pattern literature has focused on how information is phrased or presented in GUIs, most prominently cookie consent banners [6, 7, 18], ads on online platforms [28, 67], and e-commerce websites [43, 46, 58]. Recently, a handful of papers have started considering dark patterns in other interface modalities like robots [32, 56], auditory interactions with voice user interfaces (VUIs) [51], and proxemic interactions with movable platforms [26]. In section three, we extend this by characterising a *social* class of dark pattern that may appear, in some form, in any of these interface types.

Whereas dark patterns may cause negative user experiences due to their sneaky, manipulative or deceptive nature, other design patterns may have negative effects purely because of oversight or bad design, known as anti-patterns.

2.2.2 Anti-patterns. The term 'anti-pattern' was introduced by Brown *et al.* to describe a design pattern, i.e., a commonly occurring solution to a problem, that "generates decidedly negative consequences" [12, p.7]. Rather than strategically manipulating design elements towards certain outcomes, here negative consequences result *unexpectedly* due to oversight on the designer's part (e.g., making the wrong button too easy to press, or the right button too difficult). Possible causes include a lack of knowledge, insufficient experience in solving a particular type of problem (e.g., implementing a pattern/theoretical approach poorly), or applying a useful or trendy pattern in an inappropriate context [8]. Another common cause is designers failing to consider situational aspects like different users' experiences or contextual needs and goals [8] (e.g., image recognition software that only recognises certain skin tones, or text that colourblind users find unreadable).

Using this terminology, the next section introduces the terms 'social' dark and anti-patterns to characterise a particular class of risks relating to social interfaces.

[9]That is, a behaviour that serves another stakeholder's interests can still be framed as, in some regards, serving the user, e.g., encouraging the purchase of a gym membership or healthier (but more expensive) alternatives.

[10]We add the latter, as we would not consider mechanisms for behaviour change that people purposefully use to help them keep to their goals as 'dark', but we may consider it a dark pattern if a mental health app encourages engagement in unpleasant (e.g., fear-inducing or autonomy undermining) ways as an added feature by default.

3 CHARACTERISING 'SOCIAL' FORMS OF DARK AND ANTI-PATTERNS

When conducting our literature review, we found two papers relating to social cues being used manipulatively by interface designers [32, 56]. Yet, both limit their focus to social robots, and neither empirically engage with user experience. Lacey and Caudwell argue that the 'cute' aesthetic features of home robots should be considered a dark pattern in that they elicit a "powerful affective bond" from users, masking their potentially harmful (i.e., privacy-imposing) features [32, p.374]. Moreover, by leading the user to assume the role of "caregiver", they suggest that the robot's childlike cuteness gives the user a false sense of authority that may obscure the powers of influence the technology has over them [32, p.378]. This is echoed by Shamsudhin and Jotterand [56], who frame social robot design as an "inherently persuasive project", as users are led to "believing, at least temporarily, that the robot is human-like, has life-like properties, can be trusted, and there is value in the creation and maintenance of this human–robot relationship" [56, p.95].

However, a (cute) appearance is only one of many social cues in robots that can elicit social/emotional responses:

> Robots that incorporate social cues such as gaze, proximity, and facial expressions, push our Darwinian buttons ... and effectively coerce us into interacting with them socially [11, p.1915].

Any combination of verbal or nonverbal cues could potentially be used in manipulative ways [56], such as suggesting disappointment or anger with body language, or tone and facial expressions to elicit guilt or shame. On their own, text or voice can still use sentiment or tone to convey judgment or emotion strategically (e.g., "It makes me sad that you would not…", or "It would make me happy if you would…")—especially when coupled with expressions of affection or familiarity, like calling the user their "best friend" or addressing them by name. Although, in some contexts, such phrases may be more appropriate (e.g., in robot/chatbot companions or toys), they may be objectionable in others (e.g., to get users to do or consent to something they are reluctant to). Such social design patterns can be implemented in any range of interfaces, not just sophisticated CAs. However, risks may be amplified when systems behave in more convincingly human-like ways, or if a user has built a trusting 'relationship' with a specific system-as-agent over time [56]. These are merely a few examples of how interface designers may use social cues in manipulative and/or exploitative ways.

We characterise *social dark patterns* as the use of *social* design patterns (i.e., social cues or interaction patterns) for inducing certain emotions or misleading mental models (e.g., a misconception of the system's capacities) to manipulate user behaviour towards certain ends; particularly, involving mechanisms or goals they are not aware of or did not consent to. As in GUI dark patterns, this involves utilising/exploiting knowledge of specific cognitive biases in people's behaviour: in this case, social/anthropomorphic biases that lead people to treat an interface as a thinking/feeling agent, whose "feelings" and "judgment" they are cued to care about, rather than a platform [48].

Apart from unduly influencing people's behaviour, social behaviours may also be inappropriate for contextual reasons, such as exhibiting a lack of tact or situational sensitivity—although these are not mutually exclusive. We characterise *social anti-patterns* as social design patterns that are applied poorly or in an inappropriate context, such that they negatively impact user experience. For instance, a designer could think users may enjoy it when a system addresses them by a pet name (e.g., *bestie* or *sweetie*), but certain users may find this off-putting, or even offensive (e.g., being deemed patronising or sexist). Such risks are especially high in less sophisticated forms of automated social acting, like generic messages that are made to seem like it is personalised to the user, or addressing them directly, at that moment, when it is not.

We introduce this vocabulary to talk about particular risks that come with raising expectations of interfaces as social actors. With our conceptual and empirical contributions, we hope to encourage further research on risks of this nature, that is, how interfaces treat people in social interaction.

4 METHODS

Our study had four phases: an exploratory survey (n=80), followed by a week-long study using an experience sampling method (ESM) [64], individual semi-structured interviews, and a group-based exploratory workshop. The latter three phases used a subset of participants from Phase 1 (n=11). A summary of the phases is shown in Figure 1. All interviews and workshops took place online.

Using a mixed qualitative methods [41, 50, 62] (or *intra-paradigm* [52]) design, we integrated qualitative data from the four phases to triangulate our findings. This combination of methods was designed to counteract the limitations of each phase, and help us to uncover nuances in our participants' experiences and preferences. Data was thematically analysed in an iterative, integrated way using a combination of Braun and Clarke's [10, 16] reflexive approach and O'Reilly *et al.*'s [50] integrated analytic approach, described below. Our aim was exploratory: to start identifying socially manipulative tactics that are used in interfaces, as well as everyday situations in which people can find certain automated social behaviours otherwise inappropriate.

This study was approved by our university's Central University Research Ethics Committee (CUREC, ref CS_C1A_22_022). Publicly shareable data, coding, and study materials are available on the Open Science Framework (OSF).[11]

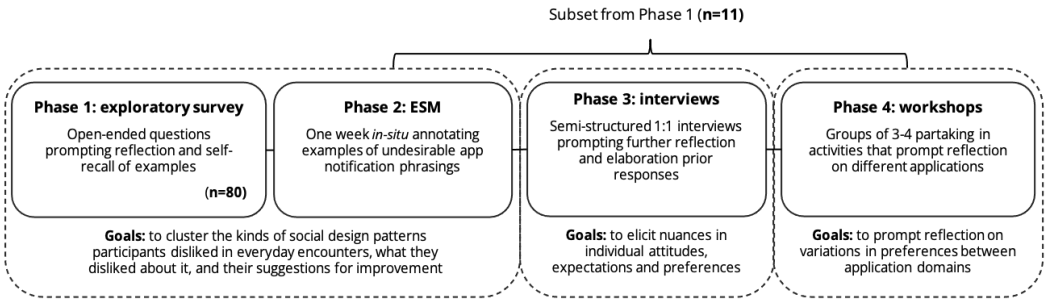

Fig. 1. Flowchart of study design and methods

4.1 Phase 1: Exploratory survey of 80 smartphone users

The first phase was an online survey consisting mainly of free-text questions. The aim was to collect examples of automated social behaviours that respondents found bothersome or off-putting, and to start identifying patterns in what they disliked about those encounters.

4.1.1 Recruitment. Recruitment was done using Prolific Academic, as well as a Twitter post using our university's departmental research group account, with the following inclusion criteria:

- Being from/primarily residing in a predominantly English-speaking country
- Being 18 years or older
- Owning a smartphone

We focused this initial exploration on adults, as we suspected younger participants may experience social interfaces differently.[12] We chose the second criterion as the researchers spoke English

[11] https://osf.io/t7z6n/?view_only=072754264d2947d0ad62f46dff4cd161

[12] E.g., children may be less likely to dislike interfaces talking to them in playful or paternalistic ways, or "playing pretend".

Table 1. Demographics for Study 1.

Country of residence	UK (54%), Canada (24%), Ireland (14%), South-Africa (6%), US (3%)
Age range	21-30 (35%), 31-40 (28%), 41-50 (14%), 51-60 (13%), 18-20 (8%), 61-70 (4%)
Gender	Male (53%), Female (46%), Prefer not to disclose (1%)

as a common language. We intended to recruit all participants via Twitter, but included Prolific when we had a smaller turnout than expected.[13] We then recruited around 20 participants at a time and analysed the data until we started seeing saturation in our findings.

A total of 84 participants filled in the Phase 1 survey: 19 recruited via Twitter, and 65 via Prolific. Of our Twitter responses, we excluded three for not meeting the location criteria, and one for not consenting to our data usage terms, leaving 80 participants in the final data set. 53% identified as male, 46% as female, and 1% preferred to not disclose their gender. Over half (54%) were based in the UK, and 63% were between the ages of 21 and 40. Table 1 summarises the demographic information. As the sorts of social interfaces people commonly encountered at the time were limited,[14] we found 80 participants sufficient. This relatively low number was also more feasible for qualitative analysis.

4.1.2 Procedure. To administer the survey, we used *Jisc online surveys*,[15] a platform that had data protection agreements in place with our institution. We designed the survey to capture a broad collection of examples, and explanations, of encounters with interfaces 'speaking' to participants in ways they found bothersome or off-putting. We were also interested in their general attitudes to different forms of treatment by automated systems: for instance, using more or less formal tones; expressing more or less emotion; or treating them more as a friend or a customer. Our aim was primarily to *understand reasons* for negative experiences, rather than making inferences about the *relative frequency* of negative experiences. As such, the questions were designed for qualitative analysis: capturing users' examples/experiences of undesirable automated social behaviours and their reasons, in their own words.

As the most common type of social interface, the survey first asked about negative encounters with smartphone notifications: whether participants had ever been "*particularly annoyed or put off by the tone or phrasing of an app notification*". If so, they were asked to list any examples they could recall, followed by an explanation of what it was about it that they disliked. We then asked the same question regarding social interfaces more broadly: whether participants had been "*annoyed or put off by other automated systems (recordings, chatbots, self-checkout machines, etc.) 'speaking' to you as if they were a real person*".

Afterwards, to understand their general preferences, respondents were asked to rate agreement on a five-point Likert scale with the following statements:

I like it when an automated system...

- *...talks to me in a friendly/chatty tone (e.g. greets me, addresses me by name, uses conversational language)*
- *...speaks like it has thoughts or feelings of its own (e.g. "I think/feel...")*
- *...has other human-like qualities (e.g. uses a human-sounding voice, has a name, has a face)*
- *...talks to me as if we're friends*

[13] Prolific survey respondents were compensated at a rate of £10 an hour, with an estimated completion time of 21 minutes (£3.50 base pay), although most took less than 10 minutes. Twitter respondents completed the survey voluntarily, but all were compensated equally for partaking in the further phases.

[14] This study was conducted in August, 2022.

[15] https://www.onlinesurveys.ac.uk/

Table 2. Demographics for Phases 2-4.

Participant	Country of residence	Age range	Gender	Workshop group
P1	UK	31-40	Female	G2
P2	Ireland	31-40	Female	G1
P3	Canada	21-30	Female	G3
P4	UK	51-60	Male	G1
P5	Canada	31-40	Female	G2
P6	UK	21-30	Male	G3
P7	South-Africa	21-30	Male	G2
P8	South-Africa	21-30	Male	G1
P9	UK	41-50	Female	G2
P11	Ireland	41-50	Male	G3
P12	UK	31-40	Male	G3

This was followed by a question to "*elaborate a bit on the above (e.g. when you do or don't, and why)*". Given our qualitative focus, our analysis focused on the patterns in respondents' elaboration on the extent to which they liked different manners of speaking, rather than on the Likert scores *per se*: their purpose was primarily to prompt reflection.

Finally, the survey asked participants if they had "*any further thoughts to share on the topic of an automated system or artificial agent acting as if it were a person*". The survey concluded by asking respondents if they were interested in participating in an upcoming study on the same topic, which we used for recruitment for the other study phases.

4.2 Phase 2: Experience sampling method

Whereas the survey allowed us to rapidly capture data, from a broader sample of participants, regarding stand-out negative encounters with social interfaces, the second phase used an ESM to capture *in-situ* examples of undesirable social behaviours as they happened. These were annotated with participant's descriptions of what bothered them, factors that contributed to their experiences and suggestions for improvement. This gave us access to relevant contextual information about particular experiences that the survey was less likely to capture, and circumvented the survey's reliance on respondents' ability to recall experiences from memory. For this, we focused on app notifications as a basic form of social interface with which most smartphone users would be familiar (and likely encounter multiple times daily).

4.2.1 Recruitment. For the next set of studies (Phases 2-4), we invited 24 participants who completed Phase 1 and expressed interest in further participation, using no further exclusion criteria. 16 agreed to participate, but five dropped out before starting, leaving 11 (demographics in Table 2).

4.2.2 Procedure. At the start of Phase 2, participants were emailed a consent form and information sheet detailing the goals, requirements and data processing procedures for Phases 2-4. The first study required participants to enable all app notifications on their smartphones (with the exception of personal messaging apps like WhatsApp or email, which were irrelevant to us). For one week (at least five, at most seven days), participants were asked to capture and annotate all the notification message phrasings that bothered them, as they occurred. The information sheet, and prior survey, conveyed that we were mainly interested in *"the way in which such systems 'speak' to them"*, rather than other aspects of the notifications.

Table 3. Headings in the daily diary tables for the ESM.

Quote (or image/ description/ paraphrase) **of notification message**	**App that sent it** (optional)	**What was it about this example that bothered you?**	**Please describe any contextual factors** (e.g., your current mood, task, environment, or personal background) **that you think may have contributed to your bad experience**	**If you could change the notification in any way, how would you improve it?**

Each participant was given access to a set of editable Microsoft Word tables (one for each day), located in a private folder in the lead author's university Nexus 365 OneDrive SharePoint account. They were asked to add entries whenever they encountered an undesirable notification (or take screenshots during the day and add them by the end, which we instructed them on). Table 3 shows the column headings of the daily entry form. Participants were offered five spaces for daily entries but were instructed on the process of adding more if needed.

4.3 Phase 3: Semi-structured interviews

As the format of Phases 1 and 2 (boxes to fill out) may have incentivised overly concise responses, this phase enabled participants to elaborate on their responses and related questions.

4.3.1 Procedure. After the ESM, we conducted 1:1 semi-structured interviews (15-30 min) with each participant over Microsoft Teams. We adapted the questions to each participant's survey responses and ESM entries, asking them to elaborate and reflect deeper on the causes of their dissatisfaction with specific notifications, as well as nuances or exceptions to their preferences. For instance, if a participant stated in the survey that they dislike it when a system talks to them as a person, we asked them if they could think of any context where they might feel differently. We also tried to unpack the source of any dissatisfaction with interfaces behaving in social ways: whether they believed it was down to a current lack of technical sophistication, or something more fundamental. If participants were bothered by, e.g., obviously generic messages, we asked about the extent to which they would prefer if technologies used personal data to tailor to them.

4.4 Phase 4: Exploratory workshops

The aim of the final phase, exploratory workshops (45 min), was to prompt group discussion and reflection on how preferences may vary between domains and applications. Another goal was to elicit more information about how participants felt users *deserved* to be treated in different domains, exploring what appropriate or respectful treatment might mean to them.

4.4.1 Procedure. We invited all participants from the previous two phases to one of three online workshops, using Microsoft Teams and the Miro online whiteboard tool.[16] Through a series of four Miro activities (mainly involving generating sticky notes) workshop participants (n=3-4) speculated together on their preferences regarding different social interfaces. First, they were told to select two interfaces out of a choice of six (7 min): *an app that motivates you to stick to your goals*, *a virtual personal receptionist (for managing emails and calendar)*, *a smart home assistant (Alexa/Google Home)*, *a tutoring chatbot for small children*, *a robot carer/companion for the elderly*, or *a chatbot therapist*

[16] https://miro.com

app. We then asked them to generate sticky notes on (a) "how you think the system should talk to people (i.e., in a way that is respectful/appropriate) in each context", and (b) "ways the system shouldn't talk to or treat people (including what may be off-putting, disrespectful, unethical, etc.)" (10-15 min). They were then asked to reflect on specific rights they believe users should have for how they are treated in each domain (15 min). They were then asked to briefly reflect on how their preferences may differ in other contexts and to cluster similar expectations together (5-8 min). Here, we gave them all the initial options, as well as eight other applications (e.g., a talking fridge, an automated vehicle, etc.) as examples. After describing each task, the researcher left the call to allow participants to discuss amongst themselves.

4.5 Data and analysis

4.5.1 Data preparation. At the end of Phase 2, we collated each participant's (5-7) daily forms into a single PDF document for analysis. During Phase 3, we took audio and video recordings of each (15-20 min) interview via Microsoft Teams and stored them in a secure institutional Nexus 365 OneDrive SharePoint folder. The lead author then downloaded the auto-generated transcripts from each interview (done by Microsoft Teams) and, following the recordings, edited them for correctness and de-personalisation. The transcripts were then thematically analysed. Finally, in Phase 4, we took audio and video recordings of each (45 min) workshop, and downloaded the completed anonymous Miro worksheets (containing sticky notes filled with text) as PDFs. We used the recordings of group discussions to help us analyse the worksheets more accurately.

4.5.2 Analytic approach. Rather than treating each phase separately, as in similar HCI studies [41, 62], our data was analysed using O'Reilly *et al.*'s [50] integrated analytic approach: using a mixture of inductive and deductive enquiry to iteratively refine themes and codes generated from multiple qualitative data sets [50]. We chose this approach as our methods were designed to complement and enhance each other, offering different perspectives and nuances on users' experiences with social interfaces. Not only did this enable a deeper and richer engagement with our issues of interest [50], but also helped us to find patterns of inconsistencies and similarities between contexts through critical comparison (similar to grounded theory approaches [14]).

To code and generate themes from our data, we used Braun and Clarke's [10, 16] reflexive approach. In what follows, we highlight a few assumptions underlying our analysis [50]. As our study focused on a largely unexplored area of dark patterns, our analysis was neither completely inductive (data-driven), nor deductive (theory-driven). On the one hand, we were, to some extent, driven by theory, as our particular understanding of dark patterns drove the kinds of research questions we posed and helped us to recognise similar tactics. However, rather than using a pre-specified codebook, we expected the types of dark patterns that would emerge in a social context to differ from those found in other kinds of interfaces, and so we were committed to understanding what *users* found manipulative or otherwise off-putting in these kinds of interfaces and why. We also did not want to assume that a design pattern is 'dark' purely because it seems to fit our theoretic criteria, and so we actively encouraged our participants to repeatedly reflect on the extent to which different patterns may be positive in different contexts.

4.5.3 Analysis process. We started our analysis by open-coding the survey results from Phase 1, guided by our research questions. In particular, we distinguished examples of social acting that contained social design patterns that seemed manipulative (i.e., framing things in ways that could be construed as pressuring users to take certain actions) (RQ1) and reasons that users found certain social design patterns otherwise off-putting (RQ2, RQ4). We also started coding patterns in suggestions from participants on how to improve features they disliked (RQ3). However, we also inductively coded patterns that went beyond our research questions. In particular, we noticed

patterns in how participants feel they deserve to be treated on more principled grounds, which we decided to incorporate into the remaining study designs. By the end of this process, we started combining codes into low-level themes.

During Phase 2, our codes and themes from Phase 1 helped us to deductively code similar extracts, but were refined or updated when patterns of nuances or exceptions were found. The ESM helped us collect more examples of manipulative tactics, richer data on contextual reasons why users found social design patterns off-putting, as well as more specific suggestions for improvement. As a part of this process, we iterated back over our prior codes and started clustering them into higher-level themes.

The codes and themes generated thus far helped us to plan the interview questions that would be most useful to ask in Phase 3, so as to further test and refine our findings. This phase followed a similar process of iteration, abstraction, and critical comparison.

Finally, the Phase 4 workshop data were coded separately as they were most usefully clustered by application domain, since each workshop focused on two particular social interfaces. As the workshop was more future-facing (anticipating differences in preferences between hypothetical social interfaces), rather than looking for things participants disliked and would like to improve, we looked for social design patterns that groups decided were desirable (RQ3) and undesirable/inappropriate (RQ2, RQ4) relative to each domain. Following our inductive coding of more normative/principled considerations for how users feel they should be treated, we asked groups to reflect on this for each application domain and thematically coded this as well. We used these results on domain variability to expand on/add nuance to our prior findings during the write-up stage.

By the end of this process, our codes and themes were ordered within three sets: *social dark patterns*, *social anti-patterns*, and *suggestions for improved social acting*. Thematic coding was conducted using NVivo 1.7.1.

5 RESULTS

5.1 Data collected

In total, survey participants contributed 84 examples of notifications that bothered them (median number of examples per participant = 1, min = 0, max = 3) and 62 examples of bothersome social behaviours by other automated systems (median = 1, min = 0, max = 3). In Phase 2, the 11 participants collectively contributed 125 diary entries about notifications they found bothersome. Two people participated for five days, one for four, and the rest for all seven. Of these, only one participant experienced no bothersome notifications (stating this in their form for every day). Seven participants had entries of bothersome notifications on all of the days they participated, two on 71% of their days, and one on 51%. The median number of entries per participant was 8 (min = 0, max = 24).

In Phase 4, we conducted two workshops with four participants (G2, G3), and one with three (G1). The application domains chosen by the groups to discuss were: a smart home assistant (G3), a tutoring chatbot for small children (G2), a robot carer/companion for the elderly (G1), and a chatbot therapist app (G1, G2, G3).

The next section summarises the themes and sub-themes we arrived at by the end of our integrated analysis process.

5.2 Social dark patterns: manipulative uses of social cues (RQ1)

Under this theme, we coded excerpts relating to social behaviours our participants found "manipulative or pressuring" (RQ1). Our participants' examples and descriptions showed clear overlap with known dark pattern tactics (e.g., being dishonest, adding pressure or burdens to choices, inducing guilt/shame), but were typically framed as the (social) behaviour of an *agent*, rather than elements

on a user interface (although they knew the messages were scripted by a person). One of our participants explained, "*I anthropomorphise more than I would like. When a machine talks to me, I can't help but think of it as a person and how I'm relating to it, and treating it, as a person.*" (P50, **S**[17]). We generated four themes to distinguish between tactics our participants described: "agents playing on emotions", "agents being pushy", "agents mothering', and "agents being passive-aggressive".

5.2.1 Agents playing on emotions. The first theme (for which contributing codes applied to 21% of survey respondents, 10% of all ESM entries, and 5/11 interviews) was the use of emotional manipulation by systems that behave as social actors (e.g., using first-person pronouns or addressing the user directly) and express certain emotions (especially disappointment or sadness) in order to get users to *do* something. This either involved appealing to their empathy (making the user feel bad for the 'agent') or appealing to their self-image (e.g., coaxing them or making them feel bad about themselves). Describing their general experiences with bothersome app notifications, one participant wrote: "*Some [say] they're sorry to see me go or that they will miss me, which is false. Some try to make me feel bad for ignoring them*" (P32, **S**). Another described how the system tries to steer them against their own desires, illustrating the manipulative aspect: "*Feeling sad when I 'hurt' it, or worse, ignore it. But I WANT to be able to ignore my phone and my stupid apps.*" (P50, **S**). Even if participants did not find this sort of tactic persuasive, there was general consensus that it should not be used as they found it annoying and/or unethical. Several participants criticised such tactics for being dishonest about the system's capacities, expressing emotions that were "false" or a "lie": "*We all know it is not a person with emotions, so whatever human-like qualities are used, must be a kind of manipulation to get or keep you interested*" (P13, **S**).

5.2.2 Agents being pushy. This theme contained references (26% **S**; 20% **E**; 1/11 **I**) to an automated system using dictating language (e.g., giving orders), aggression (e.g., shouting, using all caps or exclamation marks) or repetitive nagging to convey a sense of urgency to pressure a user to do a task. These kinds of tactics were found both in self-interested (e.g., marketing) messages and "beneficent' messages meant to help the user (e.g., reminders or instructions). For instance, one participant disliked the urgent tone of a notification from the popular language learning app Duolingo: "*Hi, it's Duo. Reminding you to practice French. Got 3 minutes now?*", explaining that "*the sense of obligation that is created is unpleasant*" (P12, **E**). Some survey respondents also expressed annoyance at self-checkout machines for being "pushy":

> *She comes across as very aggressive and loud. It feels like she's shouting at you. And she's very pushy and impatient. For example, when completing your purchase, she immediately and continuously reminds you to remove all scanned items from the bagging area* (P5, **S**).

Whilst such prompts are meant to be helpful, a rapid succession of orders felt to some participants like it was pressuring them to do the task faster, along with inducing a sense of being watched or judged by the 'agent'.

5.2.3 Agents mothering. This theme contained references (8% **S**; 5% **E**; 6/11 **I**) to tactics that felt paternalistic or overly helpful in a way that undermined participants' sense of autonomy: "*It's kind of like a parent you can't chat back to*" (P8, **S)**. We chose the term 'mothering' based on one of our participant responses: "*Apps regularly make me feel like they are trying to mother me*" (P34, **S**), because it was often the *tone* of the message (e.g., overly helpful or concerned), that made participants feel inappropriately controlled. In a rather poignant example, one respondent

[17]We will mark the data source with the following acronyms: **S** = survey (Phase 1), **E** = experience sampling (Phase 2), **I** = interview (Phase 3), **W** = workshop (Phase 4). For data from Phase 1-3, quotes are accompanied by an anonymous participant ID, e.g. "P50"; for data from Phase 4, quotes are accompanied by the number of the workshop group where the data was generated e.g., on a Miro worksheet.

complained about their phone's bedtime reminder stating that their bedtime is approaching and that they should wind down soon: "*I feel like it is too dictating and even though I feel like I am not particularly sleepy or outside having fun I feel bad not to be ready to go to bed at that moment*" (P70, **S**). Whilst, in some cases, these are similar to some 'pushy' tactics described above, we distinguished this theme for being specifically described in overly protective or parent-like terms. During interview discussions, a few participants highlighted the importance of respecting user autonomy, even if they want to act outside of their best interests: "*you have to empower the person to make that decision. You can't make too many assumptions about their well-being on their behalf.*" (P12, **I**). However, there were also examples where this tactic was used for more self-interested means: "*Stressed? (With tear face, worried-looking emoji). Why not take a break and play?*" (P5, **E**), in the context of a game app prompting a user to open it.

5.2.4 Agents being passive-aggressive. Rather than being overtly pushy or dictating, this theme regarded automated systems using more covert means to convey dissatisfaction or judgment through tone or implication (e.g., sarcasm) as a person would (6% **S**; 0% **E**; 0/11 **I**). A few of the excerpts referenced Duolingo: "*Hey it's Lily, Duo says you're ignoring him so he's sent me*" (P13, **S**). However, most mentions here did not include concrete examples.

5.3 Social anti-patterns: counterproductive uses of social cues (RQ2)

As opposed to the behavioural tactics above that strategically induce certain emotions or mental models to manipulate, another set of themes (guided by RQ2) related to social design patterns that upset users for reasons that the designers failed to anticipate. At a high level, we distinguished between the themes *contextually insensitive or tactless*, *inappropriate use of tone* (e.g., talking to users in inappropriately friend-like, parental, or childlike ways), and *obviously faking capacities* (i.e., systems pretending to be aware/sincere when it is clear they are not).

5.3.1 Contextually insensitive or tactless. The first theme contained references (50% **S**; 30% **E**; 10/11 **I**) to automated systems interrupting users or saying something (seemingly harmless) at an inappropriate time, such that it seems insensitive or offensive by implication. For example, one participant mentioned self-checkout machines telling them to "'*remove the item from the area' ... sometimes it's a false flag and the system doesn't realize but keeps repeating as if accusing you of stealing something*" (P65, **S**). Several excerpts mentioned systems being oblivious to contextual factors that make a suggestion irrelevant (e.g., suggesting they start something they are already doing), or even hurtful. For instance, one ESM participant was bothered by a notification telling them "you've achieved 73% of your step goal", explaining: "*I went mountain climbing. First time in my life and I felt proud of myself*" (P8, **E**). Another participant received a notification for ordering alcoholic drinks at 3 p.m., which they found insensitive given that they "*used to drink a bit more than I like to*" (P3, **E**).

5.3.2 Inappropriate use of tone. This theme contained references (35% **S**; 22% **E**; 8/11 **I**) to tones or *ways of speaking* that participants found off-putting for not being aligned with the role of, or nature of their relationship with, the system-as-agent. The first was an inappropriate use of a friend-like tone, conveying a sense of unwarranted closeness (e.g., addressing the user by name or pet names, or using affectionate emojis). This sometimes made participants feel uncomfortable, especially in marketing or professional service contexts. In such contexts, perceived incentives seemed to taint how agreeable and friendly behaviours seemed: "*...what are the incentives? Why is this thing doing this? What is it trying to accomplish? Is it actually trying to help me? Is it trying to help the company?*" (P7, **I**). Some participants also described it as "crossing a boundary" as they felt that is something that needs to be "earned": "*It's a term of endearment that you get over time, you know*" (P5, **I**). To

explain the strangeness, two participants compared such behaviours to a random person *"coming up to them in the street"* or *"in the shop"*, saying it would be strange even if a person spoke in such a level of familiarity "out of nowhere": "*There is definitely a line of friendliness that is sometimes crossed and it usually comes across as fake/try-hard/creepy.*" (P7, **S**).

There was also a dislike among our participants of interfaces treating them as if they were children (e.g., pitying them, praising them, or explaining more than necessary). Examples include being put off by "*an investing app telling me 'well done' and 'good job' etc when I was using it*" (P42, **S**) or a self-checkout machine "*saying obvious things like 'don't forget your receipt!"'* (P28, **S**). Multiple participants commented that they find such behaviours "patronising". This also included frustrations around being treated as if one needs special treatment, e.g., "*[the self-checkout machine's] loudness makes me think that she thinks I have a hearing problem or something*" (P5, **S**). Such patterns often overlapped with the 'mothering' tactics described above, which seemed to put participants off regardless of manipulative use. Again, this bore on a lack of appropriateness given the nature of their relationship with the agent, as some participants explained it may be appropriate if someone you are "close to" or "care about" speaks to you in a certain way, but not an interface: "*It's so different, it doesn't know me, like, 'Who are you to have that right?'*" (P8, **I**).

Rather than sounding condescending, another common complaint was social interfaces using child or teen-like language (e.g., emojis, *netspeak* or other infantile behaviour—particularly in the marketing contexts). Such behaviours came up a few times in the smartphone notifications that participants captured in the ESM (12% **E**), and were described as "try-hard", "childish", and "annoying". As one participant elaborated in the interview, "*I think basic pleasantries like, I don't know, just like, 'please' whatever is fine, but when it's trying to, like, be your friend or trying to be, like, too relatable ... that type of thing is just kind of cringy*" (P3, **I**).

5.3.3 Obviously faking capacities. A bothersome behaviour that was mentioned in all phases of study (39% **S**; 15% **E**; 5/11 **I**, as well as 3/3 of the workshop groups) was when an interface clearly "fakes" capacities for comprehension, care, or concern, typically when it is noticeably a generic or pre-programmed message. A common complaint was that it comes across as "condescending" when participants are expected to be easy to fool, or willing to "play make-believe". As one participant stated, "*I don't want to play make-believe with something ... we know you're robot!*" (P1, **I**). Several participants were also bothered by systems "acting" like they care about them in a personalised way, when they clearly do not: "*I just know it's just like programmed to say, like, 'Hi, are you having a good day?' and it's just like, no one's actually there asking me that. That's just everyone who opens this app or whatever is going to see that.*" (P3, **I**).

Contrary to early CASA research that was taken to support the use of baseless flattery and pleasantries in systems [48], some of our participants were put off by the obvious insincerity of such acts: "*This is coming from nowhere. This is not coming from a person who cares about me and wants me to feel good or whatever, or cares about anything. It doesn't have anything. It's empty, it's a machine.*" (P12, **I**). Multiple participants expressed a strong desire for "machines"/automated systems to just act *like what they are*—without trying to seem human-like in any way. Beyond being misleading, some suggested that what counts as appropriate for automated systems may differ from what is appropriate for people:

> *I generally dislike automated systems posing as real human beings as it decreases their authenticity. If I engage with an actual human, I'm expecting realistic human responses. If my conversation is carried out with an automated service I rather expect them to give me straight answers and simple choices, there is no need to introduce elements like pretending they have feelings as it dehumanises the whole experience even more by introducing fake genuine interest on their part* (P28, **S**).

5.4 Participant suggestions for improved social acting (RQ3, RQ4)

Along with our investigation into dark and anti-patterns in social interfaces, we also wanted to explore with our participants, at each phase of study, how they would improve the design of these interfaces if they could. This set of themes was centred on answering our final two research questions (RQ3, RQ4). Whilst we expect variation in preferences between different demographics, we wanted to highlight the suggestions on which there was some consensus. As the ESM was limited to preferences in the context of smartphone notifications, we used the workshops to explore preferences across other application domains, which we discuss under the relevant themes.

5.4.1 Machines should "stay in their role". There was a common desire (44% **S**; 1% **E**; 9/11 **I**) among participants for automated systems to "stay in their role" in terms of acting like a tool, a service, or a lifeless machine (as opposed to a human): "*It's a bot, I don't need it to pretend and be human, a friend or anything else other than a bot*" (P66, **S**). This includes putting effectivity before "flourishes" (e.g., being chatty, personable, or making small talk), as it can detract from the immediate needs of the user. A few participants said that they would not necessarily mind simple conversational pleasantries, as long as it does not detract from the actual purpose of the system: "*The more human-like a system is, the more likely I am to feel like I'm supposed to treat it like a human. This is often at odds with the job of whatever the machine is*" (P49, **S**). Relatedly, some participants expressed a desire for automated systems generally "keeping the relationship professional". During the interviews, we asked participants to reflect on whether they might feel differently if systems were more sophisticated in the future. However, a few considered it something they would feel uncomfortable with regardless: "*I just don't think computers will ever be so much like a human that I would feel comfortable with them interacting with me like a human.*" (P5, **I**).

During the workshops, we identified some contexts in which preferences may differ. One group considered human-likeness more appropriate in the context of smart home devices or robot carer/companions for the elderly (G2, **W**). They also suggested that a tutoring bot for young children would warrant more of a "casual", "friendly and fun" (G2, **W**) tone than a formal/professional one.

5.4.2 Make the baseline 'neutral'. This theme related to participants' descriptions of what may be a good way of speaking/treating users for automated systems generally. We coded several excerpts preferring 'to-the-point" and "neutral" language (61% **S**; 10% **E**; 36 **I**). That is, some "medium" between sounding too formal/monotonous and informal/excited: "*The exaggerated fake happiness in the voices is patronising to me*" (P70, **S**), "*Just cut the overexcitement*" (P1, **S**). This also included preferences for speaking in a clear, concise and transparent way: "*not like, 'Ohh, today's gonna be a great day. The weather's shining outside' ... just tell me the weather. We don't need the fluff.*" (P7, **I**).

5.4.3 Anticipate relevant contextual factors. This theme contained references to specific contextual factors that participants believed may affect how social behaviours are received (48% **S**; 26% **E**; 9/11 **I**). One factor was individual differences, in which participants mentioned their level of extroversion, personality, mood, age, current activity, and neurodivergence as possibly contributing to their preferences. For instance, a few participants complained about being spoken to/addressed without their consent, due to being shy or private people: "*Self-checkout machines volume is way too loud and draws attention. I'm shy in public and it gives me anxiety to use self-checkouts*" (P19, **S**). Workshop participants considered some aged-related differences: whilst one group suggested children may not find overly excited/helpful tones as patronising as adults (G2, **W**), another mentioned that elderly participants may be extra sensitive to feeling patronised or infantilised, as they are commonly subject to demeaning treatment (G1, **W**). Several ESM participants also mentioned more complex personal situations affecting how notifications are received: "*I don't have a right to tell my health app like, 'No. I'm going through an emotional time' or whatever*" (P3, **I**).

Another factor was the passing of time, which may change the nature of the relationship the user has with an agent, thus making more familiar tones more appropriate, or, conversely, the same behaviours less appealing: "*...its first notifications tend to be more useful. And the more you get them, it's just like, 'Oh, screw it, I'm over that.'*" (P8, **I**).

5.4.4 Offer means for customisation. Multiple participants expressed a need for customisation, to have the means to exert more control over how they are "spoken to" (13% **S**; 9% **E**; 5/11 **I**). A few comparisons were made to human-human contexts, stating that usually, when interacting with a person, people have the ability to negotiate how they prefer to be treated, whereas automated systems (like notifications) tend to afford more of a "take-it-or-leave-it" approach: "*[With a friend or a partner] at the very least you would have a conversation and deal with it in some way ... Whereas, with the app, there's no way of, kind of, saying, 'No, sorry, please just remind me and that's enough'*" (P12, **I**). In some cases, being able to modify how one is addressed can be especially important, as in the case of dead-naming trans people: "*as a trans person, names [are] a bit fraught, especially when my bank app insists on using my legal name which is not the one I go by in daily life*" (P13, **S**).

5.4.5 General normative considerations. Finally, as we coded data from different study phases, we started noticing excerpts that took more principled stances on how participants believe users *deserve* to be treated, as basic normative standards for *all* social interfaces. Some of these give further justification for the preferences and frustrations raised above.

One theme was respecting the user's autonomy (8% **S**; 2% **E**; 7/11 **I**, 3/3 **W**) by not letting them asymmetrically bend to the system's needs. In the case of automated systems giving instructions or reminders, one participant complained "*there's no interlocutor with whom you can have a conversation to try to come to an accommodation or any form of compromise with*" (P12, **I**). Specific examples from the workshops included allowing users to choose not to respond to a therapy bot's questions (G2) or to pause the session (G1). In principle, this means treating people as rational agents, rather than objects to control or manipulate: "*We need to treat people as agents, not as input and output and... processing machines*" (P12, **I**), or "*I'd resent the app for playing on my feelings and maybe be spiteful, not do something on purpose*"(P7, **I**).

A related theme was to design systems in ways that make their intentions and capacities transparent (i.e., not misleading people) (40% **S**; 14% **E**; 11/11 **I**, 2/3 **W**). This also relates to the theme of respecting user intelligence (3% **S**; 5% **E**; 1/11 **I**, 1/3 **W**) by not treating them as incapable or incompetent (e.g., over-explaining or withholding important information).

Another theme was to not "pigeonhole" users (14% **S**; 6% **E**; 5/11 **I**, 2/3 **W**) by assuming too much about who they are or what they like. On the topic of personalising services with user data, one interview participant expressed a need for being treated as dynamic and unpredictable: "*The whole point of being human is that you can act in unpredictable ways, and you can reinvent yourself all the time. And algorithmic profiling does not allow for that*" (P12, **I**). Practically, one workshop group suggested that a therapy chatbot should be able to truly "listen to" or accommodate individual concerns, rather than just offering advice (G3).

The final theme was respecting people's sense of personal boundaries (11% **S**; 1% **E**; 4/11 **I**, 1/3 **W**). This included concerns around agents "knowing too much" about users in a way that feels "invasive" or "creepy". One participant suggested that the *sense* of privacy/secrecy can be at least as important as what the system actually knows: "*I know that you're listening to me, but don't make it quite so obvious. Like, you know, hide in the bushes over there rather than the bushes directly in front of me. Give me at least some kind of figment of privacy*" (P2, **I**).

6 DISCUSSION AND IMPLICATIONS

In this work, we integrated four qualitative methods to elicit user experiences and preferences regarding how a range of automated systems 'talk' to them: from app notifications, to self-checkout machines, to chatbots. Our findings highlight important factors that can affect the extent to which social design patterns (i.e., social cues or interaction patterns) are deemed appropriate. We distinguish between social *dark patterns*, social design patterns that are used to manipulate user behaviour towards certain ends, and social *anti-patterns*, where social design patterns are applied poorly or in an inappropriate context, such that they bother users for reasons the designers failed to anticipate.

From iteratively analysing and coding participants' survey responses, ESM entries, interviews, and workshop group discussions, we generated four themes that capture types of social dark pattern tactics already appearing in interfaces, most prominently in smartphone notifications. We also generated three themes that describe situational factors that can make certain social design patterns be received badly, such as seeming insensitive, insincere, or invasive. Finally, we constructed five themes that represent specific suggestions, from our participants, on how to improve how social interfaces treat users, including considerations regarding the sorts of design choices that they find unethical or disrespectful in more principled terms.

As an exploratory study, our aim was not to give a definitive overview of all the tactics that could be used as dark patterns in social interfaces, but to characterise an emerging class of dark and anti-patterns, and to understand user attitudes towards examples that already exist in common automated systems. We also do not mean to posit these as people's attitudes towards social interfaces generally. Rather, we want to shine a lens on everyday contexts in which automated forms of social acting (particularly when talking to/at users proactively) can be received badly, and start to identify contributing contextual factors. Ultimately, our aim is to better understand what it means for an interface to be a *good* (tactful, respectful, constructive) social actor.

6.1 The misuse and abuse of 'social' design patterns

To our knowledge, this work is the first that treats in a unified manner the expanding set of systems that behave as social actors—not necessarily for their use of overt social/anthropomorphic cues, but for proactively "saying things" (in text, sound, or gesture) as if addressing the user. Thereby, even very basic interactive systems like notifications are able to *create* social situations [59], leading to frustrations when these actions seem socially inappropriate: whether it is for certain facts about the system, such as their role or relationship with the user, or facts about the user and their current situation. Knowing well that a system is merely a platform, our participants still tended to describe their frustrations in terms of human-like attributes (e.g., "*it*" or "*she*" seeming passive-aggressive, insincere, judgmental, or "knowing" too much), as they were judging the *behaviour* of the 'agent' as indicative of such attributes. In line with findings in the CASA paradigm, the interface is apparently treated as a social actor, as its actions are nevertheless treated (described/experienced) as if coming from an intentional agent—given what might be expected from an intentional agent in that context.

We anticipate a risk that such intuitive responses—arising from known social/anthropomorphic biases in people–may be harnessed to manipulate user behaviour in typical dark pattern fashion [43, 44]. That is, by encouraging people to treat the interface as a thinking/feeling agent, whose feelings or judgment they should care about, as a means of pressuring them to act. Our preliminary findings, mainly in the context of smartphone notifications, indicate that such socially manipulative tactics are already emerging, such as eliciting empathy for the 'agent', expressing judgment on its behalf, or acting as a concerned parent that controls the user out of care for them. Whilst the deception risk is relatively low in the context of notifications, we believe it may increase with more

sophisticated forms of social interfaces (e.g., chatbots or robots)—especially when interacting with vulnerable populations like children. This expands on Lacey and Caudwell's [32] use of *cuteness* as a dark pattern, by positioning it within a broader class of dark patterns that involve leveraging social cues in agent-like systems, offering a vocabulary for future HCI researchers to help identify/talk about similar phenomena.

Our findings also offer further insight into how raising expectations of a system *as a social actor* can backfire. Prior UX research has mainly considered expectation violations for chatbots [29, 38, 45], where certain social design patterns (like using affectionate/friend-like language) may make more sense, as users may have the opportunity to build some form of rapport or relationship with the agent (depending on its role). However, when applying the same design patterns in contexts where such familiar or affectionate terms have not been "earned", our participants expressed feeling tricked or even "dehumanised"—especially when the message is obviously generic.

Following the recent successes/hype surrounding dialogue agents based on LLMs, there has been a spike in interest in developing interfaces that behave as social actors. Whilst researchers have started exploring risks related to such technologies, most of the focus thus far has been on the level of semantics: making outputs as "helpful, honest, and harmless" as possible [53], by, for instance, guarding against biased, misleading, toxic, or inaccurate language [5, 57, 66]. However, our findings show that even seemingly innocuous (helpful, friendly, 'harmless') social actions can cause offence. This pragmatic dimension of social-interaction harms has not yet been empirically explored, as interfaces that are more commonly considered "conversational", like chatbot apps, social robots, and LLM agents, have so far mainly been engaged *with* by users at times of their choosing, rather than taking a more active social role across different contexts. As this is likely to change in future, we offer preliminary empirical evidence of situational risks that are harder to mitigate on a semantic/language level alone.

Whilst there were areas where we identified relative consensus (e.g., participants preferring more professional, emotionless tones for automated systems), this may be due, in part, to the particular domain of focus (i.e., app notifications), and demographic similarities (e.g., predominantly English-speaking adults). However, the fact that there was such an overwhelming backlash to certain ways that social design patterns have been used, shows that more attention needs to be given to how different people prefer to be spoken to/addressed in different contexts, rather than assuming 'the friendlier the better', or that what works well in one domain will work in another. With our findings, we hope to encourage a discussion on risks regarding particularly social-interactional harms (e.g., feeling disrespected, judged, or offended) that increase with interfaces behaving as social actors, which we build on in a recent paper [2].

6.2 Towards a more respectful approach to CASA

Whilst it may be alluring to apply CASA insights to behavioural design, our findings suggest that people are much more socially discerning than a designer may hope—especially after becoming increasingly accustomed to interfaces behaving in social ways. Rather than "mindlessly" treating a social interface as sentient or finding it agreeable, people may, contrarily, be immediately suspicious *because* they suspect that someone is trying to "get something" from them. An important general disanalogy with *people* as social actors, is that we expect people to act in service of their own needs and desires, whereas, with a platform, there is the knowledge that it is an artefact designed to serve external interests. Thus, rather than seeming fun or engaging, friend-like or coaxing behaviours can easily appear insincere or patronising. Even when systems are designed to help users—proactively motivating or suggesting/reminding them of activities that benefit them—users can get frustrated if this is done in a way that is undeservedly familiar, parental, or merely for knowing that the system *doesn't actually care about them.*

More fundamentally, this points to a critique of common assumptions underlying behavioural design, in social interfaces or otherwise. In applying general design patterns to steer user behaviour, expecting them to be "easy" to manipulate using predefined strategies, it treats the user not as an intelligent, self-defining agent, but as a "mindless" collection of biases and predictable responses. Rather, interfaces should treat people respectfully: in ways that show "a commitment to core values that make someone a person" [4, p.1], such as their intelligence, rational agency, and sense of self-worth [2]. Some of the normative considerations our participants raised start to paint a picture of what this may look like, which we integrate and summarise below.

6.2.1 Be transparent about intentions and capacities. Interfaces should be transparent, not only about their machinelike nature, but about their actual capacities and the intentions of their stakeholders. Even if participants are not effectively deceived, the very act of trying to deceive them can feel patronising (i.e., as if undermining their intelligence).

6.2.2 Allow people to negotiate how they are treated. Rather than treating users as 'things' to be steered, they should have the opportunity to negotiate how they are addressed/spoken to. More than telling them what they should do, they deserve options to express whether they wish to do things differently or to talk to a person who will understand.

6.2.3 Do not assume help is needed. Even with good intentions, assuming users need help without their explicit request (e.g., micromanaging them or proactively explaining instructions) can feel patronising for failing to support their sense of competence and dignity.

6.2.4 Do not treat individuals as the sum of their parts. As applications of machine learning become more commonplace, it becomes all the more alluring to predict user preferences in terms of other users they cluster with. Whilst this may be useful to some extent, the very premise that people can be clustered into types or computationally modelled/predicted is dehumanising, as it undermines their agency as self-defining individuals.

6.2.5 Respect personal boundaries. Even if a system has access to enough data to make inferences about a person, personalising responses to certain details about them can come across as invasive or "creepy". With the development of ubiquitous technologies, it becomes increasingly important to respect people's personal boundaries by not making them feel too "watched" (even if they are).

These principles, which we aim to incorporate into a normative framework for what counts as "good" automated social acting, consist of what we could discern from our participants' responses. Fleshing it out will require further engagement with different demographics, in different domains, and with different modalities of social interfaces, which we aim to do in future work.

7 LIMITATIONS AND FUTURE WORK

Among the limitations of our approach, our participant pool was limited to those aged 18-60 in English-speaking countries. We would like to repeat this in other regions and domains, where both cultural norms, as well as differences in system roles, could yield substantial differences in perspectives. We did not include children, nor older adults in our sample, who may have very different views on the (in-)appropriate contexts for social acting. Another potential limitation stems from anchoring a large part of our investigation on smartphone notifications, which we chose for being a highly common, yet under-discussed platform behaving as a social actor. Although we asked our participants about other modalities and domains, the focus on notifications may have primed participants to focus overly on this (and, as such, the social contexts of marketing and receiving reminders to do unwanted tasks). However, notifications have not yet been analysed in this way, while the ethics of more prototypical CUIs have so far dominated debates. As we

emphasised, our study was centred on *negative* experiences, in order to identify anti-patterns and dark patterns, and should not be interpreted as an unbiased view of social acting in general. Finally, as with any qualitative study, there is a risk of investigator bias. We tried to reduce this risk with our phased experimental design, testing our interpretations by asking participants for clarification on comments made during earlier phases.

This work is part of a larger project to develop an ethics of interaction for mixed-initiative systems, centred on the concept of respect. Our recent paper builds on this preliminary framework and highlights some further areas where it may be applied, including LLM dialogue agents [2].

8 CONCLUSION

Early CASA research has encouraged interaction designers to make systems act in increasingly socially-conforming (chatty, even friend-like) ways to elicit favourable responses from users. However, we still have a limited understanding of when the use of certain social cues in interfaces is inappropriate. This paper contributes to the literature on dark patterns/nudging by identifying and characterising an emerging 'social' class of dark patterns. Drawing from a series of qualitative engagements with end-users, we also critically engage with CASA research by proposing when, and why, even seemingly innocuous automated social behaviours may bother or even offend users—constituting 'social' anti-patterns. Based on end-user preferences and suggestions, we offer user-led recommendations to help interface designers prevent undesirable forms of automated social acting, and treat users in more appropriate and respectful ways. Overall, we hope our work inspires critical reflection on how automated systems treat people in interactions.

ACKNOWLEDGMENTS

This work was funded by a Lighthouse Graduate Scholarship from the University of Oxford, supported by a gift from Amazon Web Services [CS2020_Lighthouse_1376707], and a grant from the UK's Engineering and Physical Sciences Research Council [EP/S035362/1].

REFERENCES

[1] Jordan Abdi, Ahmed Al-Hindawi, Tiffany Ng, and Marcela P Vizcaychipi. 2018. Scoping review on the use of socially assistive robot technology in elderly care. *BMJ open* 8, 2 (2018), e018815.

[2] Lize Alberts, Geoff Keeling, and Amanda McCroskery. 2024. What makes for a 'good' social actor? Using respect as a lens to evaluate interactions with language agents. arXiv:2401.09082 [cs.CL]

[3] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300233

[4] Dina Babushkina. 2022. What Does It Mean for a Robot to Be Respectful? *Techné: Research in Philosophy and Technology* 26, 1 (2022), 1–30.

[5] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 610–623. https://doi.org/10.1145/3442188.3445922

[6] Benjamin Maximilian Berens, Heike Dietmann, Chiara Krisam, Oksana Kulyk, and Melanie Volkamer. 2022. Cookie Disclaimers: Impact of Design and Users' Attitude. In *Proceedings of the 17th International Conference on Availability, Reliability and Security* (Vienna, Austria) *(ARES '22)*. Association for Computing Machinery, New York, NY, USA, Article 12, 20 pages. https://doi.org/10.1145/3538969.3539008

[7] Carlos Bermejo Fernandez, Dimitris Chatzopoulos, Dimitrios Papadopoulos, and Pan Hui. 2021. This Website Uses Nudging: MTurk Workers' Behaviour on Cookie Consent Notices. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 346 (oct 2021), 22 pages. https://doi.org/10.1145/3476087

[8] Sandro Javier Bolaños-Castro, Rubén González-Crespo, and Víctor Hugo Medina García. 2011. Antipatterns: a compendium of bad practices in software development processes. *International Journal of Interactive Multimedia and Artificial Intelligence* 1, 4 (2011), 41–46.

[9] Tara Brabazon. 2015. Digital fitness: Self-monitored fitness and the commodification of movement. *Communication, Politics & Culture* 48, 2 (2015), 1–23.

[10] Virginia Braun and Victoria Clarke. 2019. Reflecting on reflexive thematic analysis. *Qualitative research in sport, exercise and health* 11, 4 (2019), 589–597.

[11] Cynthia Breazeal, Kerstin Dautenhahn, and Takayuki Kanda. 2016. Social Robotics. In *Springer Handbook of Robotics*, Bruno Siciliano and Oussama Khatib (Eds.). Springer International Publishing, Cham, 1935–1972. https://doi.org/10.1007/978-3-319-32552-1_72

[12] William H. Brown, Raphael C. Malveau, Hays W. "Skip" McCormick, and Thomas J. Mowbray. 1998. *AntiPatterns: Refactoring Software, Architectures, and Projects in Crisis* (1st ed.). John Wiley & Sons, Inc., USA.

[13] Ana Caraban, Evangelos Karapanos, Daniel Gonçalves, and Pedro Campos. 2019. 23 Ways to Nudge: A Review of Technology-Mediated Nudging in Human-Computer Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3290605.3300733

[14] Ylona Chun Tie, Melanie Birks, and Karen Francis. 2019. Grounded theory research: A design framework for novice researchers. *SAGE open medicine* 7 (2019), 2050312118822927.

[15] Leon Ciechanowski, Aleksandra Przegalinska, Mikolaj Magnuski, and Peter Gloor. 2019. In the shades of the uncanny valley: An experimental study of human–chatbot interaction. *Future Generation Computer Systems* 92 (2019), 539–548.

[16] Victoria Clarke, Virginia Braun, and Nikki Hayfield. 2015. Thematic analysis. *Qualitative psychology: A practical guide to research methods* 222, 2015 (2015), 248.

[17] Gregory Conti and Edward Sobiesk. 2010. Malicious Interface Design: Exploiting the User. In *Proceedings of the 19th International Conference on World Wide Web* (Raleigh, North Carolina, USA) *(WWW '10)*. Association for Computing Machinery, New York, NY, USA, 271–280. https://doi.org/10.1145/1772690.1772719

[18] Lorrie Faith Cranor. 2022. Cookie Monster. *Commun. ACM* 65, 7 (jun 2022), 30–32. https://doi.org/10.1145/3538639

[19] Emily S Cross, Richard Ramsey, Roman Liepelt, Wolfgang Prinz, and Antonia F de C Hamilton. 2016. The shaping of social perception by stimulus and knowledge cues to human animacy. *Philosophical Transactions of the Royal Society B: Biological Sciences* 371, 1686 (2016), 20150075.

[20] Laura Dodsworth. 2021. *A state of fear: How the UK government weaponised fear during the Covid-19 pandemic.* Pinter & Martin, London, UK.

[21] Paul Dolan, Michael Hallsworth, David Halpern, Dominic King, and Ivo Vlaev. 2010. *MINDSPACE: influencing behaviour for public policy.* Technical Report. Institute of Government.

[22] Judith Donath. 2020. 5253Ethical Issues in Our Relationship with Artificial Entities. In *The Oxford Handbook of Ethics of AI.* Oxford University Press, Oxford, UK. https://doi.org/10.1093/oxfordhb/9780190067397.013.3 arXiv:https://academic.oup.com/book/0/chapter/290656277/chapter-ag-pdf/44521956/book_34287_section_290656277.ag.pdf

[23] Asbjørn Følstad and Petter Bae Brandtzaeg. 2020. Users' experiences with chatbots: findings from a questionnaire study. *Quality and User Experience* 5, 1 (2020), 1–14.

[24] Andrew Gambino, Jesse Fox, and Rabindra A Ratan. 2020. Building a stronger CASA: Extending the computers are social actors paradigm. *Human-Machine Communication* 1, 1 (2020), 5.

[25] Colin M. Gray, Yubo Kou, Bryan Battles, Joseph Hoggatt, and Austin L. Toombs. 2018. The Dark (Patterns) Side of UX Design. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3174108

[26] Saul Greenberg, Sebastian Boring, Jo Vermeulen, and Jakub Dostal. 2014. Dark Patterns in Proxemic Interactions: A Critical Perspective. In *Proceedings of the 2014 Conference on Designing Interactive Systems* (Vancouver, BC, Canada) *(DIS '14)*. Association for Computing Machinery, New York, NY, USA, 523–532. https://doi.org/10.1145/2598510.2598541

[27] G Mark Grimes, Ryan M Schuetzler, and Justin Scott Giboney. 2021. Mental models and expectation violations in conversational AI interactions. *Decision Support Systems* 144 (2021), 113515.

[28] Hana Habib, Sarah Pearman, Ellie Young, Ishika Saxena, Robert Zhang, and Lorrie FaIth Cranor. 2022. Identifying User Needs for Advertising Controls on Facebook. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW1, Article 59 (apr 2022), 42 pages. https://doi.org/10.1145/3512906

[29] Rhonda Hadi. 2019. When Humanizing Customer Service Chatbots Might Backfire. *NIM Marketing Intelligence Review* 11, 2 (2019), 30–35. https://doi.org/10.2478/nimmir-2019-0013

[30] Daniel Kahneman. 2011. *Thinking, fast and slow.* Farrar, Straus and Giroux, New York. https://www.amazon.de/Thinking-Fast-Slow-Daniel-Kahneman/dp/0374275637/ref=wl_it_dp_o_pdT1_nS_nC?ie=UTF8&colid=151193SNGKJT9&coliid=I3OCESLZCVDFL7

[31] Rafal Kocielnik, Raina Langevin, James S. George, Shota Akenaga, Amelia Wang, Darwin P. Jones, Alexander Argyle, Callan Fockele, Layla Anderson, Dennis T. Hsieh, Kabir Yadav, Herbert Duber, Gary Hsieh, and Andrea L. Hartzler. 2021. Can I Talk to You about Your Social Needs? Understanding Preference for Conversational User Interface in Health. In

Proceedings of the 3rd Conference on Conversational User Interfaces (Bilbao (online), Spain) *(CUI '21)*. Association for Computing Machinery, New York, NY, USA, Article 4, 10 pages. https://doi.org/10.1145/3469595.3469599

[32] Cherie Lacey and Catherine Caudwell. 2019. Cuteness as a 'Dark Pattern' in Home Robots. In *14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, Daegu, Korea (South), 374–381. https://doi.org/10.1109/HRI.2019.8673274

[33] Arto Laitinen, Marketta Niemelä, and Jari Pirhonen. 2016. Social robotics, elderly care, and human dignity: a recognition-theoretical approach. In *What social robots can and should do*. IOS Press, Amsterdam, The Netherlands, 155–163.

[34] Eun-Ju Lee. 2010. The more humanlike, the better? How speech type and users' cognitive style affect social responses to computers. *Computers in Human Behavior* 26, 4 (2010), 665–672.

[35] Thomas C Leonard. 2008. *Richard H. Thaler, Cass R. Sunstein, Nudge: Improving decisions about health, wealth, and happiness*. Yale University Press, New Haven, CT.

[36] Tianyi Li, Mihaela Vorvoreanu, Derek Debellis, and Saleema Amershi. 2023. Assessing Human-AI Interaction Early through Factorial Surveys: A Study on the Guidelines for Human-AI Interaction. *ACM Trans. Comput.-Hum. Interact.* 30, 5, Article 69 (sep 2023), 45 pages. https://doi.org/10.1145/3511605

[37] Norman Long. 2001. *Development sociology: actor perspectives*. Routledge, New York, NY.

[38] Gale M. Lucas, Jill Boberg, David Traum, Ron Artstein, Jon Gratch, Alesia Gainer, Emmanuel Johnson, Anton Leuski, and Mikio Nakano. 2017. The Role of Social Dialogue and Errors in Robots. In *Proceedings of the 5th International Conference on Human Agent Interaction* (Bielefeld, Germany) *(HAI '17)*. Association for Computing Machinery, New York, NY, USA, 431–433. https://doi.org/10.1145/3125739.3132617

[39] Ewa Luger and Abigail Sellen. 2016. "Like Having a Really Bad PA": The Gulf between User Expectation and Experience of Conversational Agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 5286–5297. https://doi.org/10.1145/2858036.2858288

[40] Kai Lukoff, Alexis Hiniker, Colin M. Gray, Arunesh Mathur, and Shruthi Sai Chivukula. 2021. What Can CHI Do About Dark Patterns?. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI EA '21)*. Association for Computing Machinery, New York, NY, USA, Article 102, 6 pages. https://doi.org/10.1145/3411763.3441360

[41] Kai Lukoff, Ulrik Lyngs, Himanshu Zade, J. Vera Liao, James Choi, Kaiyue Fan, Sean A. Munson, and Alexis Hiniker. 2021. How the Design of YouTube Influences User Sense of Agency. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 368, 17 pages. https://doi.org/10.1145/3411764.3445467

[42] Diana MacDonald. 2019. *Anti-patterns and dark patterns*. Apress, Berkeley, CA, 193–221. https://doi.org/10.1007/978-1-4842-4938-3_5

[43] Arunesh Mathur, Gunes Acar, Michael J. Friedman, Eli Lucherini, Jonathan Mayer, Marshini Chetty, and Arvind Narayanan. 2019. Dark Patterns at Scale: Findings from a Crawl of 11K Shopping Websites. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 81 (nov 2019), 32 pages. https://doi.org/10.1145/3359183

[44] Arunesh Mathur, Mihir Kshirsagar, and Jonathan Mayer. 2021. What Makes a Dark Pattern... Dark? Design Attributes, Normative Considerations, and Measurement Methods. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 360, 18 pages. https://doi.org/10.1145/3411764.3445610

[45] Jingbo Meng and Yue (Nancy) Dai. 2021. Emotional Support from AI Chatbots: Should a Supportive Partner Self-Disclose or Not? *Journal of Computer-Mediated Communication* 26, 4 (05 2021), 207–222. https://doi.org/10.1093/jcmc/zmab005 arXiv:https://academic.oup.com/jcmc/article-pdf/26/4/207/40342390/zmab005.pdf

[46] Carol Moser, Sarita Y. Schoenebeck, and Paul Resnick. 2019. Impulse Buying: Design Practices and Consumer Needs. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3290605.3300472

[47] Arvind Narayanan, Arunesh Mathur, Marshini Chetty, and Mihir Kshirsagar. 2020. Dark Patterns: Past, Present, and Future: The evolution of tricky user interfaces. *Queue* 18, 2 (2020), 67–92.

[48] Clifford Nass and Youngme Moon. 2000. Machines and mindlessness: Social responses to computers. *Journal of social issues* 56, 1 (2000), 81–103.

[49] Clifford Nass, Jonathan Steuer, and Ellen R. Tauber. 1994. Computers Are Social Actors. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Boston, Massachusetts, USA) *(CHI '94)*. Association for Computing Machinery, New York, NY, USA, 72–78. https://doi.org/10.1145/191666.191703

[50] Michelle O'Reilly, Nikki Kiyimba, and Alison Drewett. 2021. Mixing qualitative methods versus methodologies: A critical reflection on communication and power in inpatient care. *Counselling and Psychotherapy Research* 21, 1 (2021), 66–76.

[51] Kentrell Owens, Johanna Gunawan, David Choffnes, Pardis Emami-Naeini, Tadayoshi Kohno, and Franziska Roesner. 2022. Exploring Deceptive Design Patterns in Voice Interfaces. In *Proceedings of the 2022 European Symposium on Usable Security* (Karlsruhe, Germany) *(EuroUSEC '22)*. Association for Computing Machinery, New York, NY, USA, 64–78. https://doi.org/10.1145/3549015.3554213

[52] Michelle O'Reilly and Nikki Kiyimba. 2015. *Advanced qualitative research: A guide to using theory*. Sage, New York, NY.

[53] Liangming Pan, Michael Saxon, Wenda Xu, Deepak Nathani, Xinyi Wang, and William Yang Wang. 2023. Automatically Correcting Large Language Models: Surveying the landscape of diverse self-correction strategies. arXiv:2308.03188 [cs.CL]

[54] Byron Reeves and Clifford Nass. 1996. *The media equation: How people treat computers, television, and new media like real people*. Cambridge university press, Cambridge, UK.

[55] Anna-Maria Seeger, Jella Pfeiffer, and Armin Heinzl. 2017. When do we need a human? Anthropomorphic design and trustworthiness of conversational agents. In *SIGHCI 2017 Proceedings*. Association for Information Systems, Seoul, Korea, 15 pages.

[56] Naveen Shamsudhin and Fabrice Jotterand. 2021. Social Robots and Dark Patterns: Where Does Persuasion End and Deception Begin? In *Artificial Intelligence in Brain and Mental Health: Philosophical, Ethical & Policy Issues*, Fabrice Jotterand and Marcello Ienca (Eds.). Springer International Publishing, Cham, 89–110. https://doi.org/10.1007/978-3-030-74188-4_7

[57] Renee Shelby, Shalaleh Rismani, Kathryn Henne, AJung Moon, Negar Rostamzadeh, Paul Nicholas, N'Mah Yilla-Akbari, Jess Gallegos, Andrew Smart, Emilio Garcia, and Gurleen Virk. 2023. Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society* (Montréal, QC, Canada) *(AIES '23)*. Association for Computing Machinery, New York, NY, USA, 723–741. https://doi.org/10.1145/3600211.3604673

[58] Ray Sin, Ted Harris, Simon Nilsson, and Talia Beck. 2022. Dark patterns in online shopping: do they work and can nudges help mitigate impulse buying? *Behavioural Public Policy* (2022), 1–27. https://doi.org/10.1017/bpp.2022.11

[59] Lucy Suchman. 2007. *Human-machine reconfigurations: Plans and situated actions*. Cambridge University Press, Cambridge, UK.

[60] Nina Svenningsson and Montathar Faraon. 2020. Artificial Intelligence in Conversational Agents: A Study of Factors Related to Perceived Humanness in Chatbots. In *Proceedings of the 2019 2nd Artificial Intelligence and Cloud Computing Conference* (Kobe, Japan) *(AICCC 2019)*. Association for Computing Machinery, New York, NY, USA, 151–161. https://doi.org/10.1145/3375959.3375973

[61] Ekaterina Svikhnushina, Alexandru Placinta, and Pearl Pu. 2021. *User Expectations of Conversational Chatbots Based on Online Reviews*. Association for Computing Machinery, New York, NY, USA, 1481–1491.

[62] Jonathan A. Tran, Katie S. Yang, Katie Davis, and Alexis Hiniker. 2019. Modeling the Engagement-Disengagement Cycle of Compulsive Phone Use. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300542

[63] Amos Tversky and Daniel Kahneman. 1974. Judgment under Uncertainty: Heuristics and Biases. *Science* 185, 4157 (1974), 1124–1131. https://doi.org/10.1126/science.185.4157.1124 arXiv:https://www.science.org/doi/pdf/10.1126/science.185.4157.1124

[64] Niels Van Berkel, Denzil Ferreira, and Vassilis Kostakos. 2017. The experience sampling method on mobile devices. *ACM Computing Surveys (CSUR)* 50, 6 (2017), 1–40.

[65] Niels van Berkel, Mikael B. Skov, and Jesper Kjeldskov. 2021. Human-AI Interaction: Intermittent, Continuous, and Proactive. *Interactions* 28, 6 (nov 2021), 67–71. https://doi.org/10.1145/3486941

[66] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2022. Taxonomy of Risks Posed by Language Models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 214–229. https://doi.org/10.1145/3531146.3533088

[67] Eric Zeng, Miranda Wei, Theo Gregersen, Tadayoshi Kohno, and Franziska Roesner. 2021. Polls, Clickbait, and Commemorative $2 Bills: Problematic Political Advertising on News and Media Websites around the 2020 U.S. Elections. In *Proceedings of the 21st ACM Internet Measurement Conference* (Virtual Event) *(IMC '21)*. Association for Computing Machinery, New York, NY, USA, 507–525. https://doi.org/10.1145/3487552.3487850

Received January 2023; revised October 2023; accepted January 2024