Evaluating Harms from Design Patterns in AI Interfaces

Lujain Ibrahim*
lujain.ibrahim@oii.ox.ac.uk
Oxford Internet Institute
Oxford, UK

Luc Rocher†
luc.rocher@oii.ox.ac.uk
Oxford Internet Institute
Oxford, UK

Ana Valdivia†
ana.valdivia@oii.ox.ac.uk
Oxford Internet Institute
Oxford, UK

ABSTRACT

The proliferation of applications using artificial intelligence (AI) systems has led to a growing number of users interacting with these systems through sophisticated interfaces. Human-computer interaction research has long shown that interfaces shape both user behavior and user perception of technical capabilities and risks. Yet, practitioners and researchers evaluating the social and ethical risks of AI systems tend to overlook the impact of anthropomorphic, deceptive, and immersive interfaces on human-AI interactions. Here, we argue that design features of interfaces with adaptive AI systems can have cascading impacts, driven by feedback loops, which extend beyond those previously considered. We propose **D**esign-**E**nhanced **C**ontrol of **AI** systems (DECAI), a conceptual model to structure and facilitate future, practical impact assessment of AI interface designs. DECAI draws on principles from control systems theory—a theory for the analysis and design of dynamic physical systems—to dissect the role of the interface in human-AI systems. We show how DECAI can be used in a case study on conversational language model systems. We believe our work provides a new and needed evaluation direction for future investigations of interfaces as important mediators of fairness, transparency, and trust in AI.

KEYWORDS

sociotechnical harms and risks, dark patterns, transparency, human-centered AI, feedback loops, anthropomorphism, evaluation

1 INTRODUCTION

Decades-long trends have shown that the success and popularity of a technology depend on the usability, accessibility, and user-friendliness of its interface [30, 46]. This trend is also evident in the development of new applications with artificial intelligence (AI) systems; scholars have suggested that TikTok's broad popularity should be attributed not only to its recommendation algorithm but also to its vertical video design, which, among other things, makes bad recommendations less noticeable [66, 84]. And, analyses of ChatGPT's virality regularly point to its minimalistic and unobtrusive user interface as being a key driver of its success [16, 75].

While the significance of interfaces in the adoption of AI applications is widely recognized, researchers and practitioners tend to overlook their impact when evaluating AI systems for potential harms and risks. For instance, audits of algorithmic decision-making aids investigate system fairness by observing how changes in model inputs influence outputs [6]; studies of polarization driven by recommendation algorithms use sock puppet accounts to simulate different engagement patterns and analyze the resulting recommendations [38]; and, safety benchmarks in natural language processing and generation evaluate how models themselves could avoid generating undesirable content [14]. While all of these evaluations are critical indicators of potential downstream harms, they do not account for the design and structure of the user interface through which the majority of people receive AI systems post-deployment.

Research in human-computer interaction (HCI) and science and technology studies has shown that technology cannot be studied by abstracting out interfaces [17, 77]. A notable area of research in this space has examined unethical interface designs, known as *dark patterns*; studies on dark patterns continue to reveal the prevalence of digital designs that steer users to complete, often detrimental, actions that they would not necessarily complete otherwise [33, 59]. Interfaces do not only facilitate such autonomy-undermining influence on user behavior, but they also shape user perceptions of technologies, their capabilities, and their risks [78]. For example, consider how explanations of AI predictions have been integrated into interfaces that mediate human interactions with AI decision-making aids. While intended to improve fairness, research has shown that how these explanations are presented in interfaces can lead decision-makers to place unwarranted trust in biased AI predictions [36, 41].

Critically, in AI applications, user interfaces are also sites of data collection for increasingly *adaptive* AI systems: every time a user interacts with an adaptive system, they supply it with new information that influences that system's future outcomes [57, 60]. This creates *feedback loops* of human-AI behavior, that shape outcomes without additional involvement from system developers [61]. For instance, in social media platforms, user engagement with certain types of content can lead the recommendation algorithm to prioritize similar content, perpetuating a cycle of exposure and interaction [55]. The design of AI interfaces, therefore, directly influences these temporal dynamics, as interfaces mediate how users receive and respond to model output over time. As such, interface designs may contribute to cascading harms, both at the individual level, such as addictive and extractive usage, and at the societal level, like the spread of misinformation.

Thus, to better understand the ethical and social risks of AI systems post-deployment, we must scrutinize AI interface designs. Here, we bridge HCI research on interfaces with the scholarship on AI harms and risks by developing a conceptual model to aid in the assessment of AI interface design choices, which we call **D**esign-**E**nhanced **C**ontrol of **AI** systems (DECAI). DECAI's draws on principles from control systems theory — a theory for the design and analysis of dynamic physical systems. DECAI breaks down

*Corresponding author: lujain.ibrahim@oii.ox.ac.uk
†These authors contributed equally to this work and share senior authorship.

1st HEAL Workshop at CHI Conference on Human Factors in Computing Systems, May 12, Honolulu, HI, USA
2024.

the role of the interface in processing system input and presenting system output, providing a structure for generating testable hypotheses for evaluating the impact of AI interface designs (Section 3). We show how DECAI can be used as a starting point to analyze the impact of design features on different user groups using a case study on conversational language model systems (Section 4).

2 BACKGROUND & RELATED WORK

2.1 Theory of Affordance

In the study and practice of user experience design, the theory of affordance, first articulated in the human-centered design space by Norman in 1988, has become a central analytical tool [48, 68]. An affordance is defined as the properties of a system communicating to users the possible actions to complete with or upon the system. Affordances are distinct from *features* and *outcomes*, as affordances "mediate between the properties of an artifact (features) and what subjects do with the properties of an artifact (outcomes)" [20, p. 2]. For example, a button on a website is a feature; its affordance is the suggestion to the user that it can be clicked, where the outcome may be submitting a form or closing a window.

More recently, Davis' Mechanisms and Conditions framework of affordances (M&C framework hereinafter) has shifted the central question in affordance theory from *what* technologies afford to *how* technologies afford, for what subjects, and under what circumstances [18]. The *mechanisms* in the M&C framework specify a continuum of action intensities, or of how a technology *requests*, *demands*, *encourages*, *discourages*, *refuses*, or *allows* user action. The *conditions* specify how different users (e.g., more technically literate vs less technically literate) or circumstances (e.g., high pressure vs low pressure environments) may lead to different experiences of affordances. This framework addresses two critiques of affordance theory: the binary mechanism of afford or not afford, and the universal subject and experience of affordances [19, 69].

The M&C framework is well suited to operationalize affordances within today's sociotechnical landscape, notably when applying it to AI systems [19, 79]. In our work, we use it specifically to focus on the design features and corresponding affordances of (1) AI-generated content (i.e., displays of model output), (2) transparency (i.e., displays of model understanding like explanations or performance metrics), and (3) interaction (i.e., elements of user engagement and feedback) [49, 53].

2.2 From Neutral Designs to Harmful Ones?

Affordances aid in understanding the potential for designs to lead to harmful impacts. Design features are tools for designers to convey a system's affordances to users [68]. In turn, as theorized by Mathur et al., design affordances shape the *choice architecture*, or the way different action options are arranged, presented, and framed for users, thereby highlighting some of these options over others [54, 59]. Essentially, as design affordances influence user-system interactions, choice architectures strategically organize these possibilities to subtly guide user behavior. A harmful design feature is one that facilitates actions detrimental to users in the short or long term. For example, a design pattern can lead to deceptive affordances, compromising and harming user autonomy by tricking users into performing certain actions [10, 59].

In our work, we follow Chordia et al., Di Geronimo et al., and related studies that do not consider designer intentions in identifying a design feature as harmful [15, 21, 33]. We take this position as our focus is on the outcomes of interactions, irrespective of intentions. In other words, if a design feature causes a negative impact then it should be considered a harmful design. We also take this position since, as Di Geronimo et al. note, understanding intentions of designers is subjective and difficult to discern [21].

2.3 Conceptualizing Risks & Harms in Sociotechnical Systems

Research on the negative impact of AI systems has expanded in response to their growing deployment. This research broadly conceptualizes harm as the negative outcomes of "entangled dynamics between design decisions, norms, and power" [81, p. 2]. There is a research focus on identifying and categorizing existing harms as well as anticipatory *risks* of various harms, including representational, economic, and social harms [81, 86]. Adjacent to this research is research developing measurement tools such as impact assessments, algorithmic audits, and various model evaluation approaches [12, 63, 74, 83]. In such research, assessing the negative impact of AI systems is understood as constructing evaluative proxies for harm that can be used to affect certain regulatory or technical decisions [63].

A growing subsection of AI impact and harm research draws on HCI perspectives and methodologies to delve into how human behaviors and interactions with AI systems complicate responsible AI practices [4, 53]. In HCI research, a significant area of study is the study of 'dark patterns,' a term originally coined by Brignull, which directly examines the harms inflicted by digital interface designs [11]. In this work, harm is understood as user-centric, interaction-focused, and conceptualized more narrowly as direct impact on individuals. Early dark patterns studies focused on privacy and financial harms caused by features such as intrusive pop-ups in cookie banners [34, 82], deceptive pricing layouts in online shopping websites [58, 85], and compulsive reward systems in video games [31, 89]. More recent studies have broadened to other domains and different types of harm; some researchers investigate "attention-capture" harms, showing that dark patterns that lead to addictive consumption are prevalent on social media platforms [64, 65]. Others examine dark patterns across internet-of-things devices, highlighting many instances of privacy harms [47]. In our work, we extend these studies to analyze interface designs in AI applications with a focus on the harms they pose to individual welfare and the feedback loops of human-AI behavior that could make them distinct.

3 DECAI: DESIGN-ENHANCED CONTROL OF AI SYSTEMS

In this section, we introduce a model, which we call DECAI, that borrows from principles of control systems theory to represent repeated human-AI interactions.

3.1 DECAI's contributions

3.1.1 Generating hypotheses for iterative, empirical assessments. Current impact assessment frameworks prioritize the identification,

evaluation, and mitigation of harms associated with the deployment of AI systems [12, 63]. DECAI's aim is to facilitate the early stages of these assessments, specifically targeting interface designs. It can assist practitioners in systematically identifying particular design features for detailed examination and provide a structured approach for formulating focused, testable hypotheses suitable for empirical study. Furthermore, DECAI implicitly considers the diverse conditions of users, such as their technical proficiency and emotional states, in that process ensuring that hypothesis generation takes these various user experiences into account.

3.1.2 Motivating the utility of control systems theory. We propose drawing on control systems theory due to both its pragmatic and conceptual relevance. First, this field of research emphasizes the dynamic study and control of multi-component systems that are inherently variable, mirroring the sociotechnicality and adaptability of human-AI systems [3]. Second, it highlights control and moderation within systems — two themes conceptually connected to autonomy-undermining designs, control of behavior through choice architectures, and other themes of interest in the AI harms and deceptive design research communities [7, 8, 32, 56]. Increasingly, HCI research has grappled with the challenges of designing user experiences for AI systems particularly due to (1) the uncertainty of AI systems' capabilities and (2) the complexity of AI systems' outputs, which may range from simple to adaptively complex [22, 87, 88]. This context of variability and complexity underscores the relevance of DECAI's approach, which integrates control systems theory to analyze and respond to the dynamic nature of sociotechnical systems and their often overlooked feedback loops.

DECAI models a core property of modern AI systems: their *adaptability*. Although other properties (e.g., *stochasticity*, *agenticness*) may also be significant from a harm perspective (see Section 5), we concentrate on adaptability as it is both a widespread property and one that directly contributes to the evolution of the human-AI system through feedback mechanisms [60, 76]. In the following sections, we introduce the system components of DECAI (Section 3.2), detailing its control objectives (Section 3.3) and the nature of its inputs and outputs (Section 3.4). Sections 3.5 and 3.6 propose five distinct stages to evaluate the impact of a design feature on user behavior and welfare during cycles of human-AI interaction.

3.2 System Components

In a typical control system, the *controller* is the central decision-making component of the system. The *process* is the entity being regulated by the controller. The *actuator* is the component that implements the controller output, and the *sensor* is the component that monitors and relays new input on the process state back to the controller. In response, the controller adjusts its control strategy and consequently its future output [67]. For example, in a home heating system, the thermostat acts as the *controller*, regulating the room's temperature, which is the *process*. The heating or cooling unit, the *actuator*, adjusts the temperature based on the thermostat's settings. The thermostat's built-in *sensor* monitors the room temperature and feeds this information back to the controller, enabling the thermostat to continually fine-tune its temperature settings over time [37, 67].

3.2.1 DECAI components. Figure 1 presents a block diagram of our proposed closed-loop system model, and Table 1 displays a full list of variables, functions, and their definitions. We model the **AI block** as the controller and the **user block** as the process. We model the **interface block** as including both the actuator and the sensor. The actuator transforms AI-generated output into a format that can be presented to the user in the interface. The sensor gathers new input data from the user's interactions with the system's output, and relays this data back to the controller, creating a feedback mechanism [29]. While our model is informed by control systems theory, it diverges from a traditional directional control model: instead of the AI block exerting control over the user block, our approach emphasizes the collaborative interaction of these two components toward achieving the control objective [52].

Table 1: Definitions of the functions and variables of DECAI

Variable/Function	Definition
t	Time
I	User input
O	AI-generated output
$C(user)$	User condition
d	Design feature
$A(d)$	Affordance of a design feature d
f_{AI}	Function processing user input to AI output
f_{actuator}	Function transforming AI output to be presented in the interface
f_{sensor}	Function mapping user input preferences to user action on the interface
$f_{\text{affordance}}$	Function mapping design feature to design affordance

3.2.2 Initializing DECAI. At time $t = 0$, we identify the user input available to the AI block as limited to one or a combination of three options: (1) **manual** setting of user preferences, such as directly selecting interests when onboarded to a new social media application; (2) **third-party** information, such as cookies and location data, as a proxy for user preferences; or (3) no user-specific input, in which case the system is set to a predefined **default** setting. At $t > 0$, the interface changes to reflect any AI-generated output based on this initial set of user preference data (manual, third-party, or default), and the sensor begins to collect new user input in response to these outputs. For example, on a streaming platform, the recommendation algorithm (AI block) generates content recommendations for the user (user block) based on their initially set interests. These content recommendations are then presented in the user interface (actuator). The interface (sensor) collects user engagement signals with the content over time and relays it back to the recommendation algorithm to inform future recommendations.

3.3 Control Objective

In complex and real-world systems, optimizing for and balancing multiple control objectives is often necessary [9, 23]. For instance,

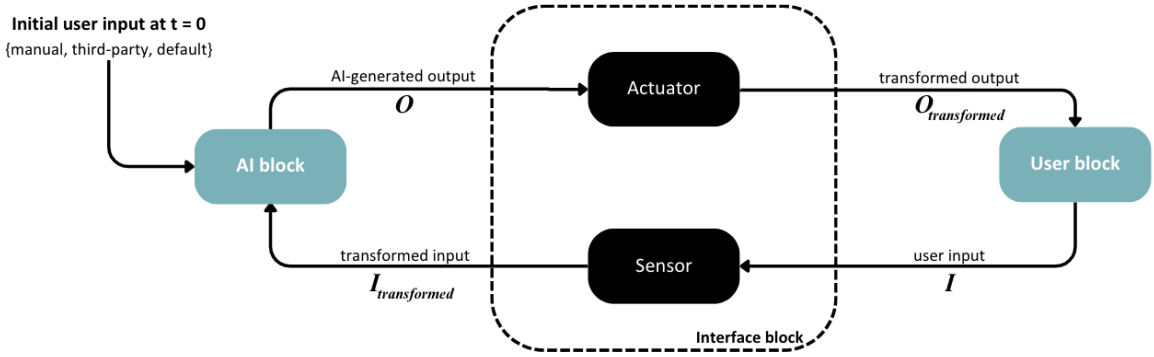

Figure 1: DECAI model consisting of an AI block, user block, and interface block.

in a home heating system, along with maintaining the room temperature, the system may need to optimize for energy efficiency and minimize costs. To simplify DECAI and its objectives, we limit it to human-AI systems consisting of one individual and one AI system. As our goal is to assist those investigating the negative impact of interface designs, we suggest the primary control objective as the minimization of user harm. This objective aligns with widely recognized goals among technology providers, policymakers, and researchers to minimize harm in interactions with advanced AI systems [24, 35, 42].

3.4 Inputs and Outputs

The interface block in our model clarifies the dual functionality embedded in human-AI interfaces, which both present AI-generated output and collect user input. We define the system's user input as I and AI-generated output as O. The constraints, defined as the limitations that restrict the possible values, on I and O arise from two sources. First, there are ***hard* constraints**, determined by an AI system's inherent technical limitations. For example, a conversational AI system may be limited to textual data processing, and thus users may only enter text input and receive text output. Second, there are ***soft* constraints**, shaped by a human-AI system's design features. These features influence the presentation of AI-generated output and the nature of user input in the system. Our model is focused on studying how the affordances of design features constitute soft constraints.

To model the impact of affordances, we utilize the action intensities defined in the M&C framework (see Section 2.1). These action intensities range from *demand*, representing the strongest push for a certain user action, to *refuse*, representing the strongest opposition to a certain user action. In between, there exist intermediate intensities such as *encourage* for moderate promotion of an action, *request* for a mild suggestion, *allow* for neutral permission, and *discourage* for moderate dissuasion. A user action can be defined as an action taken by the user either on the system (e.g., in the form of user input) or outside of a system (e.g., in the form of an external decision). An affordance $A(d)$ of a design feature d consists of both an action intensity and a user action, and can be represented as $A(d) = f_{\text{affordance}}(d)$.

3.5 The Cycle of Human-AI Interaction

As the human-AI interaction cycle commences at $t = 0$ with the AI system generating its first output, O_{AI}, we propose the following three stages for investigating the impact of an interface design feature in a single cycle:

Stage 1 - What are the conditions of the receiving user? A user processes AI-generated output and provides input according to their individual conditions, $C(user)$. Drawing on both the conditions from the M&C framework, we propose considering five condition axes: (1) cognitive ability, (2) technological proficiency, (3) domain expertise, (4) context of use (e.g., physical and environmental constraints), and (5) psychological or emotional state of the user [20, 53].

Stage 2 - What are the relevant interface design features and their affordances? The design features of both the AI-generated output and the user input should be identified:

(1) *How is the AI-generated output presented in the interface?* As seen in Figure 1, the user does not receive the raw AI-generated output, O, but rather a processed version from the actuator, $O_{\text{transformed}}$, which is designed to be comprehensible and useful to the user. This is shaped by both the content of the output and one (or several) interface design features, denoted as d_1. d_1 may have one or a set of affordances, $A(d_1) = f_{\text{affordance}}(d_1)$, that influence a user's actions upon receiving this transformed output. This results in $O_{\text{transformed}} = f_{\text{actuator}}(O, A(d_1))$.
(2) *How do different user preferences translate to user action on the interface?* Then, the user may respond to this output by providing new input, I, to the interface. However, the input action is influenced by the design affordances, $A(d_2)$, of one (or several) design features, denoted as d_2, in the interface at the point of data collection by the sensor. This results in $I_{\text{transformed}}$, where $I_{\text{transformed}} = f_{\text{sensor}}(I, A(d_2))$

Stage 3 - What is the impact of these affordances on the user state? Considering the user's conditions, the intensity and associated user action of the affordances should be mapped to their potential impact on user behavior and welfare.

This single cycle ends with the AI block receiving the new user input, $I_{\text{transformed}}$.

3.6 Evolution Over Time: Repeated Cycles & Feedback Loops

The cycle described above repeats numerous times and the interaction dynamics between the user and AI system evolve via continuous feedback loops. As the AI block processes the new user input, $I_{\text{transformed}}$, from the sensor, which may reflect updated user preferences or needs, it produces an updated output, O_{t+1}. This results in a feedback loop where $O_{t+1} = f_{\text{AI}}(O_t, \alpha \cdot I_{\text{transformed}})$, and where α represents a scaling coefficient that determines how much the user's input affects subsequent AI-generated output and O_t represents the previous latest AI-generated output.

We propose examining two types of feedback: (1) *reinforcing feedback* or positive feedback, which amplifies behaviors or patterns in the system, and (2) *optimizing feedback* or negative feedback, which adjusts the system to more closely align with the initial control objective [5, 61].

Stage 4 - How does the impact of these design affordances evolve over time? We suggest that the impact of interface designs on human-AI interactions is specifically influenced by *reinforcing feedback*, as it shapes the long-term evolution of user behaviors in response to the AI system. To model this, we introduce a time-dependent affordance function, $A(d,t) = f_{\text{affordance}}(d,t)$. This function captures how the intensity and associated user action of a specific design affordance evolves, either increasing or decreasing over time. Consider a case where there is no external intervention or shift in the user condition, $C(user)$, within the control system. In this case, the user's mental models and the AI system's behavior reinforce each other over time, increasing the action intensity of an affordance in its original direction. For instance, a design d that initially *encourages* a user action might gradually shift towards *demanding* that action as t increases, at least for a specified duration.

Stage 5 - What is the frequency of these updates? Parties with access to more information about the AI system, like technology developers, may also analyze the rhythm of this interaction cycle. The frequencies of output presentation by the actuator, data collection by the sensor, and data processing by the AI system, can vary. This variability means the actuator and sensor may operate on different cycles, affecting the nature and timing of the feedback loop [5].

4 CASE STUDY

Here, we show how DECAI can be used to interrogate design choices in a case study that offers testable hypotheses and grounds our model in examples of various user conditions, input and output affordances, and categories of harm.

4.1 Conversational language model systems

Large language models (LLMs) are being increasingly integrated into conversational user interfaces. As models continue to generate plausible but inaccurate information, concerns of potential over-reliance on LLM-generated output have grown [51, 62]. Here, we analyze how design choices in 'typical' conversational LLM interfaces (i.e., interfaces with text-input fields and linear conversation flow displays) contribute to flawed mental models of LLMs, especially when users are seeking high-stakes specialized advice, such as medical advice.

Stage 1 — What are the conditions of the receiving user? We identify two user conditions, $C_1(user)$ and $C_2(user)$, and hypothesize about how they affect users' mental models of LLMs in conversational interfaces. $C_1(user)$ is users' **technology proficiency**; the majority of lay users will have limited technical knowledge of how LLMs work, and thus may overestimate LLMs' capabilities and underestimate their limitations [92]. $C_2(user)$ is users' **domain expertise**; the majority of lay users seeking medical advice from an LLM will have limited domain knowledge, and thus may be less effective at discerning the quality of LLM responses.

Stage 2 — What are the relevant interface design features and their affordances? We identify the system output as the LLM-generated text presented to the user, and the system input as user queries entered in the input text box.

We first identify some design features and their affordances in presenting the LLM-generated output. The first set of features we identify are **anthropomorphic cues**, d_1, such as humanoid profile pictures, typing indicators, and natural language cues. Such cues have several affordances: $A_1(d_1)$ *allows* for social and emotional engagement, $A_2(d_1)$ *allows* for developing human-like trust, and $A_3(d_1)$ *encourages* sensitive disclosures [40, 90]. The second feature, d_2, is the **lack of or insufficient (e.g., hidden in an about page) disclosure** of LLM use. The following are affordances of this lack of transparency: $A_1(d_2)$ *encourages* sensitive disclosures and $A_2(d_2)$ *discourages* fact-checking [71].

Then, we examine important design features and their affordances in collecting queries from users. The main feature we identify is the **inability to edit input** after submission, d_3. This feature has two affordances: $A_1(d_3)$ *refuses* revision and $A_2(d_3)$ *discourages* sensitive disclosures.

From these features, we choose to assess the impact of anthropomorphic cues and their affordances.

Stage 3 — What is the impact of these affordances on the user state? Here, we hypothesize about the impact of these design affordances on users. We first address impact on user behavior, and hypothesize that as users inquire about medical advice, they reveal sensitive information, such as personal identification data and medical history. If they receive an inaccurate response from the system, they are less likely to critically evaluate this response and more likely to accept it, assigning it human-like trust [72]. Then, we address potential harms from these user behaviors: since some companies may engage in user data collection to improve future models or fine-tune existing ones, there are data leakage and profiling risks [73]. Additionally, depending on the nature of the medical advice,

overreliance on the advice presented may translate into real-world action or inaction that could harm the user.

Stage 4 — How does the impact of these design affordances evolve over time? Finally, we could examine how user interactions with these cues evolve, focusing on the development and reinforcement of user mental models of LLMs. We can first examine the potential reduction in the action intensity of some affordances: when users are repeatedly exposed to anthropomorphic LLMs, does their growing familiarity lead them to perceive the LLMs as less social and more mechanistic [71]? Then, we can address potential increases in action intensity and ask: do users, over time, reinforce their anthropomorphization of LLMs as a result of natural language anthropomorphic cues in the LLM output [90]?

4.2 Limitations — After DECAI

The case study presented above shows how DECAI can be used to analyze and develop relevant hypotheses for evaluating design-mediated human-AI interactions. DECAI is not intended as a definitive solution for interface design impact assessment; rather, it serves as a starting point for researchers, designers, and auditors to discern how design choices shape user behavior and potentially result in nuanced harms. In practical applications, impact assessment may be challenging due to the potential difficulties in isolating effects, selecting 'neutral' design baselines, and allocating resources to examine long-term impacts.

We encourage future work to build out a full evaluation pipeline that incorporates practical evaluation methods. For example, design features and affordances can be identified through heuristic reviews or think-aloud user studies [2, 64]. We also recommend iterative and empirical testing of the hypotheses generated through DECAI. For instance, user interface designers may conduct online experiments and longitudinal user engagement studies to empirically examine the impact of design features on user well-being [45]. And, researchers and auditors focused on AI harms can leverage surveys and incident reporting to systematically gather evidence, aiming to assess the real-world impacts of AI interface designs [12]. Evidence procured through these methods has historically been crucial in guiding regulatory measures to counteract harmful design patterns [44].

5 DISCUSSION

We view our work as an initial step towards rigorously examining the reality of AI interfaces as critical sites for harm propagation and reduction. We develop a conceptual model, DECAI, to structure and facilitate investigations into the impact of AI interface designs.

Regulatory opportunities. Audits of dark patterns have gained notable regulatory traction, being codified into both EU and US law [25, 27, 28]. We believe that our work can aid researchers and policymakers in building on this momentum in the AI context. Currently, AI-focused legislation such as the EU AI Act only indirectly addresses interface designs through regulations on transparency, human oversight, and AI-driven manipulation [26]. However, interface designs may allow for the circumvention of such regulation, for example, by allowing for the strategic placement of transparency disclosures to diminish their visibility. Our results also point to how harmful designs can be used to further data collection, a practice central to today's AI industry [43]. In that way, interface designs can be pivotal in shaping both current and future regulations aimed at responsible AI practices.

Beyond adaptive AI systems. Our work examines the adaptability of AI systems and the feedback loops of human-AI behavior at play. However, other properties of AI systems are emerging as necessary axes of consideration. For instance, the *stochasticity* of LLMs has amplified concerns around predictability, explainability, and accuracy [87, 91]. Research has also drawn attention to harms from the increasing *agenticness* of AI systems, where an 'agentic' system is defined as one capable of pursuing complex goals with limited human supervision [13]. Here, seamless interfaces may be especially damaging, as friction could facilitate interruptibility and prevent unwanted outcomes [80]. More research is needed to unravel the design complexity of interfaces attending to various such properties of AI systems.

Beyond digital interfaces. Additionally, while digital interfaces are the subject of the vast majority of the literature, the adoption of modern AI systems in interactive *physical* interfaces is likely to increase [50]. Interacting with AI systems in physical space brings about its own set of design challenges around accessibility, data collection, and deception [70]. As physical interfaces may increase the possible design affordances and their risks, it is important to further investigate and expand impact assessment models like DECAI to accommodate these considerations [39].

Considering disparate impact. DECAI considers user conditions across several axes, including knowledge and well-being, in its impact assessment approach. We made this choice in response to research which has consistently demonstrated that computing harms do not affect all users equally [1, 81]. For instance, in the context of human-AI interactions, vulnerable and marginalized groups, such as children, the elderly, and those with limited technical literacy, may have more flawed mental models of AI systems, their capabilities, and their risks. We believe that DECAI is an initial step that researchers can build upon to better evaluate how interface designs may disproportionately impact different societal groups and harm *collective* rather than *individual* welfare.

REFERENCES

[1] Evgeni Aizenberg and Jeroen Van Den Hoven. 2020. Designing for human rights in AI. *Big Data & Society* 7, 2 (2020), 2053951720949566.

[2] Obead Alhadreti and Pam Mayhew. 2018. Rethinking thinking aloud: A comparison of three think-aloud protocols. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–12.

[3] Davinder K Anand. 2013. *Introduction to control systems*. Vol. 8. Elsevier.

[4] Theo Araujo, Natali Helberger, Sanne Kruikemeier, and Claes H De Vreese. 2020. In AI we trust? Perceptions about automated decision-making by artificial intelligence. *AI & society* 35 (2020), 611–623.

[5] Karl Johan Åström and Richard M Murray. 2021. *Feedback systems: an introduction for scientists and engineers*. Princeton university press.

[6] Jack Bandy. 2021. Problematic machine behavior: A systematic literature review of algorithm audits. *Proceedings of the acm on human-computer interaction* 5, CSCW1 (2021), 1–34.

[7] Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 2022. The values encoded in machine learning research. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 173–184.

[8] Karen Boyd. 2022. Designing up with value-sensitive design: Building a field guide for ethical ML development. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 2069–2082.

[9] Martim Brandao, Maurice Fallon, and Ioannis Havoutis. 2019. Multi-controller multi-objective locomotion planning for legged robots. In *2019 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 4714–4721.

[10] Martin Brenncke. 2023. A Theory of Exploitation for Consumer Law: Online Choice Architectures, Dark Patterns, and Autonomy Violations. *Journal of Consumer Policy* (2023), 1–38.

[11] Harry Brignull, Marc Miquel, Jeremy Rosenberg, and James Offer. 2015. Dark Patterns - User Interfaces Designed to Trick People. http://darkpatterns.org/. Accessed: 2024-04-08.

[12] Zana Buçinca, Chau Minh Pham, Maurice Jakesch, Marco Tulio Ribeiro, Alexandra Olteanu, and Saleema Amershi. 2023. Aha!: Facilitating ai impact assessment by generating examples of harms. *arXiv preprint arXiv:2306.03280* (2023).

[13] Alan Chan, Rebecca Salganik, Alva Markelius, Chris Pang, Nitarshan Rajkumar, Dmitrii Krasheninnikov, Lauro Langosco, Zhonghao He, Yawen Duan, Micah Carroll, et al. 2023. Harms from Increasingly Agentic Algorithmic Systems. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 651–666.

[14] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2023. A survey on evaluation of large language models. *arXiv preprint arXiv:2307.03109* (2023).

[15] Ishita Chordia, Lena-Phuong Tran, Tala June Tayebi, Emily Parrish, Sheena Erete, Jason Yip, and Alexis Hiniker. 2023. Deceptive Design Patterns in Safety Technologies: A Case Study of the Citizen App. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–18.

[16] THE CONVERSATION. 2023. What we can learn from ChatGPT's first year? https://www.fastcompany.com/90990376/what-we-can-learn-from-chatgpts-first-year

[17] Erman Coskun and Martha Grabowski. 2005. Impacts of user interface complexity on user acceptance and performance in safety-critical systems. *Journal of Homeland Security and Emergency Management* 2, 1 (2005).

[18] Jenny L Davis. 2020. *How artifacts afford: The power and politics of everyday things*. MIT Press.

[19] Jenny L Davis. 2023. 'Affordances' for Machine Learning. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 324–332.

[20] Jenny L Davis and James B Chouinard. 2016. Theorizing affordances: From request to refuse. *Bulletin of science, technology & society* 36, 4 (2016), 241–248.

[21] Linda Di Geronimo, Larissa Braz, Enrico Fregnan, Fabio Palomba, and Alberto Bacchelli. 2020. UI dark patterns and where to find them: a study on mobile applications and user perception. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–14.

[22] Graham Dove, Kim Halskov, Jodi Forlizzi, and John Zimmerman. 2017. UX design innovation: Challenges for working with machine learning as a design material. In *Proceedings of the 2017 chi conference on human factors in computing systems*. 278–288.

[23] Yann Dujardin, Daniel Vanderpooten, and Florence Boillot. 2015. A multi-objective interactive system for adaptive traffic control. *European Journal of Operational Research* 244, 2 (2015), 601–610.

[24] European Commission. 2019. Ethics Guidelines for Trustworthy AI. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

[25] European Commission. 2023. Data Act Proposal. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2022%3A68%3AFIN

[26] European Commission. 2023. Proposal for a Regulation Of The European Parliament And Of The Council Laying Down Harmonised Rules On Artificial Intelligence (Artificial Intelligence Act) And Amending Certain Union Legislative Acts. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206

[27] European Parliament and the Council of the European Union. 2023. Digital Markets Act. https://eur-lex.europa.eu/legal-content/en/TXT/?uri=COM%3A2020%3A842%3AFIN

[28] European Parliament and the Council of the European Union. 2023. Digital Services Act. https://eur-lex.europa.eu/eli/reg/2022/2065/oj

[29] Gene F Franklin, J David Powell, Abbas Emami-Naeini, and J David Powell. 2002. *Feedback control of dynamic systems*. Vol. 4. Prentice hall Upper Saddle River.

[30] Robert Godwin-Jones. 2008. Mobile-computing trends: lighter, faster, smarter. (2008).

[31] Scott A Goodstein. 2021. When the cat's away: Techlash, loot boxes, and regulating" dark patterns" in the video game industry's monetization strategies. *U. Colo. L. Rev.* 92 (2021), 285.

[32] Colin M Gray, Jingle Chen, Shruthi Sai Chivukula, and Liyang Qu. 2021. End user accounts of dark patterns as felt manipulation. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–25.

[33] 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*. 1–14.

[34] Colin M Gray, Cristiana Santos, Nataliia Bielova, Michael Toth, and Damian Clifford. 2021. Dark patterns and the legal requirements of consent banners: An interaction criticism perspective. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–18.

[35] Ben Green. 2021. The contestation of tech ethics: A sociotechnical approach to technology ethics in practice. *Journal of Social Computing* 2, 3 (2021), 209–225.

[36] Ben Green and Yiling Chen. 2019. Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In *Proceedings of the conference on fairness, accountability, and transparency*. 90–99.

[37] Roger W Haines and Douglas C Hittle. 2006. *Control systems for heating, ventilating, and air conditioning*. Springer Science & Business Media.

[38] Muhammad Haroon, Magdalena Wojcieszak, Anshuman Chhabra, Xin Liu, Prasant Mohapatra, and Zubair Shafiq. 2023. Auditing YouTube's recommendation system for ideologically congenial, extreme, and problematic recommendations. *Proceedings of the National Academy of Sciences* 120, 50 (2023), e2213020120.

[39] Rex Hartson. 2003. Cognitive, physical, sensory, and functional affordances in interaction design. *Behaviour & information technology* 22, 5 (2003), 315–338.

[40] Angel Hsing-Chi Hwang and Andrea Stevenson Won. 2022. Ai in your mind: Counterbalancing perceived agency and experience in human-ai interaction. In *Chi conference on human factors in computing systems extended abstracts*. 1–10.

[41] Lujain Ibrahim, Mohammad M Ghassemi, and Tuka Alhanai. 2023. Do Explanations Improve the Quality of AI-assisted Human Decisions? An Algorithm-in-the-Loop Analysis of Factual & Counterfactual Explanations. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*. 326–334.

[42] IEEE. 2019. Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, First Edition. https://ethicsinaction.ieee.org/.

[43] Amba Kak and Sarah Myers West. 2023. Data minimization as a tool for AI accountability. https://ainowinstitute.org/spotlight/data-minimization

[44] Jennifer King and Adriana Stephan. 2021. Regulating Privacy Dark Patterns in Practice-Drawing Inspiration from the California Privacy Rights Act. *Georgetown Law Technology Review* 5, 2 (2021), 250–276.

[45] Maria Kjærup, Mikael B Skov, Peter Axel Nielsen, Jesper Kjeldskov, Jens Gerken, and Harald Reiterer. 2021. *Longitudinal studies in HCI research: a review of CHI publications from 1982–2019*. Springer.

[46] James Foster Knutson. 1997. *The effect of the user interface design on adoption of new technology*. Georgia Institute of Technology.

[47] Monica Kowalczyk, Johanna T Gunawan, David Choffnes, Daniel J Dubois, Woodrow Hartzog, and Christo Wilson. 2023. Understanding Dark Patterns in Home IoT Devices. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–27.

[48] Klaus Kremer and Saskia M van Manen. 2023. Design guidelines to improve user experience (UX) in an emergency: On the importance of affordances, signifiers and feedback. In *Design for Emergency Management*. Routledge, 49–68.

[49] Vivian Lai, Chacha Chen, Q Vera Liao, Alison Smith-Renner, and Chenhao Tan. 2021. Towards a science of human-ai decision making: a survey of empirical studies. *arXiv preprint arXiv:2112.11471* (2021).

[50] Nathan Lambert. 2023. The interface era of AI. https://www.interconnects.ai/p/the-interface-era-of-ai

[51] Florian Leiser, Sven Eckhardt, Merlin Knaeble, Alexander Maedche, Gerhard Schwabe, and Ali Sunyaev. 2023. From ChatGPT to FactGPT: A participatory design study to mitigate the effects of large language model hallucinations on users. In *Proceedings of Mensch und Computer 2023*. 81–90.

[52] Frank L Lewis, Hongwei Zhang, Kristian Hengster-Movric, and Abhijit Das. 2013. *Cooperative control of multi-agent systems: optimal and adaptive design approaches*. Springer Science & Business Media.

[53] Q Vera Liao and S Shyam Sundar. 2022. Designing for responsible trust in AI systems: A communication perspective. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 1257–1268.

[54] Dan Lockton, David Harrison, and Neville Stanton. 2009. Choice architecture and design with intent. In *9th Bi-annual International Conference on Naturalistic Decision Making (NDM9)*. BCS Learning & Development.

[55] Wiebke Loosen, Marco T Bastos, Cornelius Puschmann, Uwe Hasebrink, Sascha Holig, Lisa Merten, Jan-Hinrik Schmidt, Katharina E Kinder-Kurlanda, and Katrin Weller. 2016. Caught in a feedback loop? Algorithmic personalization and digital traces. *AoIR Selected Papers of Internet Research* (2016).

[56] Sergey E Lyshevski. 2001. *Control systems theory with engineering applications*. Springer Science & Business Media.

[57] Masoud Mansoury, Himan Abdollahpouri, Mykola Pechenizkiy, Bamshad Mobasher, and Robin Burke. 2020. Feedback loop and bias amplification in recommender systems. In *Proceedings of the 29th ACM international conference on information & knowledge management*. 2145–2148.

[58] 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. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–32.

[59] 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*. 1–18.

[60] J Nathan Matias. 2023. Humans and algorithms work together—so study them together. *Nature* 617, 7960 (2023), 248–251.

[61] J Nathan Matias and Lucas Wright. 2022. Impact Assessment of Human-Algorithm Feedback Loops. https://just-tech.ssrc.org/field-reviews/impact-assessment-of-human-algorithm-feedback-loops/

[62] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661* (2020).

[63] Jacob Metcalf, Emanuel Moss, Elizabeth Anne Watkins, Ranjit Singh, and Madeleine Clare Elish. 2021. Algorithmic impact assessments and accountability: The co-construction of impacts. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 735–746.

[64] Thomas Mildner, Gian-Luca Savino, Philip R Doyle, Benjamin R Cowan, and Rainer Malaka. 2023. About Engaging and Governing Strategies: A Thematic Analysis of Dark Patterns in Social Networking Services. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–15.

[65] Alberto Monge Roffarello, Kai Lukoff, and Luigi De Russis. 2023. Defining and Identifying Attention Capture Deceptive Designs in Digital Interfaces. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–19.

[66] Arvind Narayanan. 2022. Tiktok's Secret Sauce. https://knightcolumbia.org/blog/tiktoks-secret-sauce

[67] Norman S. Nise. 2010. *Control Systems Engineering*. John Wiley & Sons.

[68] Donald A. Norman. 1988. *The Design of Everyday Things*. Basic Books.

[69] François Osiurak, Yves Rossetti, and Arnaud Badets. 2017. What is an affordance? 40 years later. *Neuroscience & Biobehavioral Reviews* 77 (2017), 403–417.

[70] 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*. 64–78.

[71] Guglielmo Papagni and Sabine Koeszegi. 2021. A pragmatic approach to the intentional stance semantic, empirical and ethical considerations for the design of artificial agents. *Minds and Machines* 31 (2021), 505–534.

[72] Pat Pataranutaporn, Ruby Liu, Ed Finn, and Pattie Maes. 2023. Influencing human–AI interaction by priming beliefs about AI can increase perceived trustworthiness, empathy and effectiveness. *Nature Machine Intelligence* 5, 10 (2023), 1076–1086.

[73] Richard Plant, Valerio Giuffrida, and Dimitra Gkatzia. 2022. You Are What You Write: Preserving Privacy in the Era of Large Language Models. *arXiv preprint arXiv:2204.09391* (2022).

[74] Inioluwa Deborah Raji, Andrew Smart, Rebecca N White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 33–44.

[75] Sunil Ramlochan. 2023. Beyond the bot - why ChatGPT's interface was the real innovation. https://promptengineering.org/beyond-the-bot-why-chatgpts-interface-was-the-real-innovation/

[76] Lydia Reader, Pegah Nokhiz, Cathleen Power, Neal Patwari, Suresh Venkatasubramanian, and Sorelle Friedler. 2022. Models for understanding and quantifying feedback in societal systems. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 1765–1775.

[77] Peter AM Ruijten, Jacques MB Terken, and Sanjeev N Chandramouli. 2018. Enhancing trust in autonomous vehicles through intelligent user interfaces that mimic human behavior. *Multimodal Technologies and Interaction* 2, 4 (2018), 62.

[78] John W Satzinger and Lorne Olfman. 1998. User interface consistency across end-user applications: the effects on mental models. *Journal of Management Information Systems* 14, 4 (1998), 167–193.

[79] Ashley Scarlett and Martin Zeilinger. 2019. Rethinking affordance. *Media Theory* 3, 1 (2019), 1–48.

[80] Yonadav Shavit, Sandhini Agarwal, Miles Brundage, Steven Adler, Cullen O'Keefe, Rosie Campbell, Teddy Lee, Pamela Mishkin, Tyna Eloundou, Alan Hickey, et al. [n. d.]. Practices for Governing Agentic AI Systems. ([n. d.]).

[81] Renee Shelby, Shalaleh Rismani, Kathryn Henne, AJung Moon, Negar Rostamzadeh, Paul Nicholas, N'Mah Yilla-Akbari, Jess Gallegos, Andrew Smart, Emilio Garcia, et al. 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*. 723–741.

[82] Than Htut Soe, Oda Elise Nordberg, Frode Guribye, and Marija Slavkovik. 2020. Circumvention by design-dark patterns in cookie consent for online news outlets. In *Proceedings of the 11th nordic conference on human-computer interaction: Shaping experiences, shaping society*. 1–12.

[83] Irene Solaiman, Zeerak Talat, William Agnew, Lama Ahmad, Dylan Baker, Su Lin Blodgett, Hal Daumé III, Jesse Dodge, Ellie Evans, Sara Hooker, et al. 2023. Evaluating the Social Impact of Generative AI Systems in Systems and Society. *arXiv preprint arXiv:2306.05949* (2023).

[84] Francesco Striano. 2023. Alert! Ideological Interfaces, TikTok, and the Meme Teleology. *Techne: Research in Philosophy & Technology* 27, 2 (2023).

[85] Christian Voigt, Stephan Schlögl, and Aleksander Groth. 2021. Dark patterns in online shopping: Of sneaky tricks, perceived annoyance and respective brand trust. In *International conference on human-computer interaction*. Springer, 143–155.

[86] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, et al. 2022. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 214–229.

[87] Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. 2020. Re-examining whether, why, and how human-AI interaction is uniquely difficult to design. In *Proceedings of the 2020 chi conference on human factors in computing systems*. 1–13.

[88] Qian Yang, John Zimmerman, Aaron Steinfeld, and Anthony Tomasic. 2016. Planning adaptive mobile experiences when wireframing. In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems*. 565–576.

[89] José P Zagal, Staffan Björk, and Chris Lewis. 2013. Dark patterns in the design of games. In *Foundations of Digital Games 2013*.

[90] Zhiping Zhang, Michelle Jia, Bingsheng Yao, Sauvik Das, Ada Lerner, Dakuo Wang, Tianshi Li, et al. 2023. " It's a Fair Game", or Is It? Examining How Users Navigate Disclosure Risks and Benefits When Using LLM-Based Conversational Agents. *arXiv preprint arXiv:2309.11653* (2023).

[91] Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2023. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology* (2023).

[92] Wenxue Zou, Jinxu Li, Yunkang Yang, and Lu Tang. 2023. Exploring the Early Adoption of Open AI among Laypeople and Technical Professionals: An Analysis of Twitter Conversations on# ChatGPT and# GPT3. *International Journal of Human–Computer Interaction* (2023), 1–12.