AI Automatons: AI Systems Intended to Imitate Humans

ALEXANDRA OLTEANU*, Mila - Québec AI Institute, Canada
SOLON BAROCAS, Microsoft Research, USA
SU LIN BLODGETT*, Mila - Québec AI Institute, Canada
LISA EGEDE*, Carnegie Mellon University, USA
ALICIA DEVRIO*, Carnegie Mellon University, USA
MYRA CHENG*, Stanford University, USA

There is a proliferation of AI systems designed to mimic people's behavior, work, abilities, likenesses, or humanness—systems we dub *AI automatons*. Individuals, groups, or generic humans are simulated to produce creative work in their styles, respond to surveys in their places, probe how they would use a new system before deployment, provide users with assistance and companionship, and anticipate their possible future behavior and interactions with others, just to name a few applications. However, the research, design, deployment, and availability of such AI systems have prompted growing concerns about a wide range of possible legal, psychological, social, and other types of harms. In this paper, we seek 1) to facilitate productive discussions about *whether*, *when*, and *how* to design and deploy such systems, and 2) to help chart the current landscape of existing and prospective *AI automatons*. To do so, we tease apart determinant design axes and considerations to aid reflections and deliberations about whether and how design choices along these axes could *mitigate*—or instead *exacerbate*—harms that the development and use of *AI automatons* might give rise to. Through a synthesis of related literature and extensive examples of existing AI systems intended to mimic humans, we developed a conceptual framework that foregrounds key axes of design variations and provides analytical scaffolding to foster greater recognition of a) the design choices available to developers and researchers, as well as of b) the possible ethical implications these design choices might have.

CCS Concepts: • **Human-centered computing → HCI theory, concepts and models**; **Interactive systems and tools**; • **Social and professional topics**; • **Computing methodologies** → *Modeling and simulation*; *Artificial intelligence*;

Additional Key Words and Phrases: anthropomorphism, anthropomorphic AI, AI automatons, impacts

ACM Reference Format:
Alexandra Olteanu, Solon Barocas, Su Lin Blodgett, Lisa Egede, Alicia DeVrio, and Myra Cheng. 2026. *AI Automatons*: AI Systems Intended to Imitate Humans. In *The 2026 ACM Conference on Fairness, Accountability, and Transparency (FAccT '26), June 25–28, 2026, Montreal, QC, Canada.* ACM, New York, NY, USA, 39 pages. https://doi.org/10.1145/3805689.3812377

1 Introduction

There is a fast-growing number of examples where AI systems are developed or used to imitate humans. These include cases where the likenesses and voices of deceased or missing children were reenacted to help narrate their stories of abuse and violence [122], such as that of a "17-month-old [who] died in 2007 following months of physical abuse" [106], or cases of AI characters meant to depict made-up intersectional identities, such as "'Liv'

*Work done while at Microsoft Research Montréal.

Authors' Contact Information: Alexandra Olteanu, Mila - Québec AI Institute, Canada, alexandra@aolteanu.com; Solon Barocas, Microsoft Research, USA; Su Lin Blodgett, Mila - Québec AI Institute, Canada; Lisa Egede, Carnegie Mellon University, USA; Alicia DeVrio, Carnegie Mellon University, USA; Myra Cheng, Stanford University, USA.

FAccT '26, June 25–28, 2026, Montreal, QC, Canada

ACM ISBN 979-8-4007-2596-8/2026/06
https://doi.org/10.1145/3805689.3812377

portraying a 'proud Black queer momma of 2 & truth-teller'" [187]. There is also "[a]n array of popular apps [that] are offering AI companions [...] who are spinning up AI girlfriends, AI husbands, AI therapists—even AI parents" [252]. Some developers seek to mimic specific individuals [e.g., 33, 154, 156, 174, 230], while others only aim to imbue systems with more general human-like characteristics [e.g., 140, 244].

Indeed, growing beliefs that AI systems will be or are already human-like and able to replicate a wide range of human abilities or likenesses [e.g., 47, 56, 82, 94, 172]—AI systems we dub *AI automatons*[1] to emphasize their *mechanical nature*—have led both to growing interest in developing such systems [e.g., 112, 188, 235], as well as to growing concerns about their potential to replace humans in various jobs, about the potential emotional toll from interacting with seemingly human-like systems, or about adverse impacts to humans in other more or less anticipated ways [e.g., 7, 30, 33, 38, 42, 57, 83, 154, 185, 252]. This trend is far from new, with the field of AI itself guided by the question of whether computational systems are capable of thought or at least of faithfully imitating humans—also known as Turing's *imitation game* [98]—an aim which has a long history in fields like gaming [e.g., 12, 152, 167], human-robot interaction [e.g., 60, 79], and animation and motion pictures [e.g., 84, 177].

What is perhaps *new* is the increasing feasibility and availability of AI systems that could be used to simulate specific individuals or highly realistic human-like entities that are able to engage in increasingly autonomous and open-ended interactions with others, coupled with a growing ubiquity of such systems across a wider and wider range of applications, both of which are propelled by expanding and increasingly pervasive claims about, and perceptions of, AI systems' capabilities. The growing possibility of developing highly accurate simulations of individuals, in particular, has led to examinations of their ethical and social implications [e.g., 33, 122, 154, 162, 174], and of the possibility of these simulations being used to replace humans [e.g., 7, 261, 275]. While critically important, these early efforts tend to focus on either a specific class of AI automatons or a specific class of concerns, and do not *examine what and how various design choices may amplify different types of concerns and risks—for which we seek to provide an analytical foundation in this paper.*

Contributions. In this work, we develop a conceptual framework that maps key design considerations when building and deploying AI automatons (§3), highlighting possible adverse impacts different design choices might have. In doing so, we focus on the *intended* goals for such AI automatons, rather than attempting to speculate about what AI automatons can or cannot do or about the incentives and motivations of those developing these systems (beyond the stated goals for what the AI automatons are intended for). In other words, we focus on what AI automatons are developed, deployed, used, or intended to be used for. Our aims are two-fold: 1) chart the current landscape of existing and prospective *AI automatons*, and 2) provide analytical scaffolding and a foundation for a) discussions about *whether*, *when*, and *how* to design and deploy such systems, and for b) future examinations of the adverse impacts certain types of configurations might have. Our framework does so in a few ways: First, a greater recognition of possible design choices can help developers and researchers[2] of AI automatons both to be more intentional and explicit about the choices they make for *how* to design and *when* to deploy AI automatons, as well as to discern potential adverse impacts that different choices and their interactions may have. Second, added clarity about different choices' impacts can encourage them to consider alternative choices—for *how* to design and deploy AI automatons—that may help mitigate these impacts, or identify redlines to guide decisions about *whether* to build certain types of AI automatons. Finally, by providing a common, systematized terminology, our framework can facilitate dialogue among developers and researchers, enabling more transparent, standardized documentation and reporting of AI automatons' characteristics.

[1]An *AI automaton* is a AI system that is "relatively self-operating" or that is "designed to follow automatically a predetermined sequence of operations or respond to encoded instructions" [179] in order to reproduce or mimic humans or their characteristics and behavior. As a result, any AI system intended to *simulate* humans by reproducing their characteristics and behaviors constitutes an AI automaton.

[2]We use *developers and researchers* as a shorthand for stakeholders that make decisions about the design, development, or deployment of AI automatons. As we will see in §2.3, in our framework these stakeholders also take the role of *operators* or *interactors*.

2 Background & Related Work

AI systems are increasingly anthropomorphic [11, 47, 149, 171, 225]—described or perceived as human-like. Anthropomorphism can be *by design*, often by incorporating human-like features into systems, e.g., avatars with different skin colors or hairstyles [105], or robots with facial features [239] or producing human sounds and gestures [161]. Such design choices may be motivated by desires to increase users' engagement, comfort, familiarity, or trust [130, 139, 215, 219, 254], improve user experiences [109, 274, 277], or encourage consumer engagement [196] and consumption [101, 103, 197], though anthropomorphization can also backfire [178]. AI systems may also be anthropomorphic even when *not intentionally designed for*. For example, language use, until recently solely a human activity and made possible for AI systems by training on large quantities of human-produced language, can readily give rise to perceptions of human-likeness [69].

2.1 Simulating Humans

To appear human-like, however, AI systems need to reproduce or appear to reproduce human characteristics.

Simulating individuals. For instance, simulations have been developed to target a wide array of individuals, including models personalized to specific chess players to predict their next moves [175], simulating individual Supreme Court justices to predict future decisions [102], and simulating users for sending emails, "match[ing] the voice and tone in the emails you've already sent, applying that to everything [the model] creates" [257]. Particularly when they include a visual component, such simulations are often also referred to as *deepfakes* [33, 191], with a 2024 CBS news report highlighting that there were "more than 21,000 deepfake pornographic videos online—up more than 460% over the year prior" [17]. Simulations may target people no longer alive, including loved ones as well as public figures [123, 185]. Emerging applications also include those aimed at simulating many individuals at once, such as for pilot studies [222] and polls [288]. Park et al. [201] simulate "attitudes and behaviors of 1,052 real individuals." Other simulations may target fictional individuals—e.g., models role-playing as specific characters [263]—as well as entirely new characters, "offering AI companions to millions of [...] users" [252].

Simulating groups. Simulations may also target members of social groups, ranging from social or professional roles to demographic groups. For example, Qian et al. [212] develop agents in roles like programmers and test engineers for software development, and Sun et al. [247] develop a legal consultation system with agents in roles like receptionists and lawyers. Argyle et al. [15] construct prompts with demographic information to see if a model's output reflects response distributions (for e.g., surveys) "each aligned with a real human sub-population," while Lee et al. [155] investigate whether LLMs conditioned on demographics can simulate responses to climate change surveys. Basoah et al. [22] examine user perceptions of systems using features of two English sociolects, which some participants perceived as more human-like than a system using Standard American English. Other simulations of members of demographic groups include AI social media accounts, like "Grandpa Brian," a Meta account which "described itself ... as an African-American retired entrepreneur" and whose bio is an "entirely fictionalized biography based on a composite of real African American elders' lives" [187].

Simulating human phenomena & interactions. An increasing number of applications also seek to simulate people in order to study human phenomena, including people's beliefs and attitudes [201]; social dynamics and interactions in simulated communities and social networks [86, 214, 253, 262, 281]; and human decision-making across a range of settings, such as resource allocation [132] and government responses to public disaster [276]. See Mou et al. [188] for a survey.

2.2 Growing Concerns

As anthropomorphic AI systems have proliferated, work has also emerged raising critical concerns about them. Scholars have long problematized "relational artifacts [...] *specifically designed to make people feel understood*,

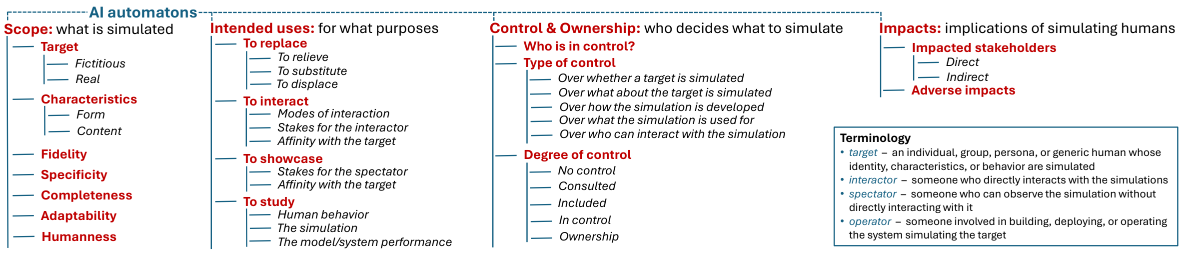

Fig. 1. Overview of the design considerations covered by our conceptual framework. For a more detailed breakdown of the considerations we identified for AI automatons (along with examples), see Table 1 in Appendix A.

[artifacts that] are still without understanding" [255, emphasis original], as they lack the human understanding required for the relationships people try to create with them. Moreover, encouraging users to relate to AI systems as if they are human can cause over-reliance on system outputs [2, 71]—leading to exaggerated perceptions of AI capabilities as well as distorting moral judgments about responsibility [207, 229]—and negatively impact critical thinking abilities [88]. Such systems can also prevent "users from assuming certain roles themselves, or [...] from questioning the need for certain roles in the first place" [166]. In the long term, such systems may lead to long-term emotional dependence [85], and their capacity to express feelings they cannot have may devalue expressions of genuine emotion and erode our ability to bond with each other [208]. Their perceived trustworthiness [14]—potentially heightened by misuse of users' personal information [166]—may enable increased user deception, manipulation, and exploitation [266], and combined with their increased ubiquity risks gradual acclimatization that may facilitate increased public acceptance of potentially unethical uses of AI [215].

As they are designed to imitate humans, AI automatons are likely to be perceived as human-like and thus give rise to the same concerns as anthropomorphic systems more broadly. However, as systems *explicitly* designed to simulate people's behavior, work, abilities, or likenesses, they can give rise to additional concerns, particularly about harms to those whose characteristics they purport to reproduce. Some of these concerns involve practical challenges; for example, Agnew et al. [7] identify limitations of current such systems, including their tendency to make mistakes and to reproduce dominant perspectives rather than those of the people they are intended to replace. Wang et al. [261] argue that LLMs' tendencies to *misportray* demographic groups (generate out-group rather than in-group members' perspectives) and *flatten* groups (treat groups as monoliths, erasing heterogeneity and neglecting intersectional identities) make them unsuitable as replacements for human participants.

Beyond these concerns—which may be overcome with modeling advancements—these and other works have also identified *concerns fundamental to the act of simulating people*. Agnew et al. and Wang et al. note how replacing study participants reproduces minoritized groups' exclusion from decision-making and moves away from the meaningful sharing of power that is core to visions of inclusion. Wang et al. remark that such simulations also risk essentializing identity by treating identities as "rigid and innate." McIlroy-Young et al. [174] characterize normative concerns arising from *mimetic models*—generative and interactive models simulating specific people. Lee et al. [154] investigate users' and targets' perceptions of simulations (*AI clones*), identifying concerns ranging from misrepresentation to replacement and exploitation.

2.3 Concepts & Working Terminology

Drawing on McIlroy-Young et al. and others using similar terminologies [122, 154, 156], to differentiate between different types of stakeholders our framework considers the following stakeholder roles:[3] (1) *target* – an individual,

[3]Unlike [174], we do not distinguish between the builders of the systems and the operators of the systems, and by definition, spectators are different from operators and interactors. We also use a more expansive definition for the target, which does not need to be a specific individual.

group, persona, or generic human whose identity, characteristics, or behavior are simulated; (2) *interactor* – someone who directly interacts with the simulation; (3) *spectator* – someone who can observe the simulation without directly interacting with it; (4) *operator* – someone involved in the building, deployment, or operation of the system simulating the target. However, the stakeholder roles can overlap, with the same entity being able to take on multiple roles; for instance, the *interactor* can be the same as the *target*, such as when users interact with their own replicas [e.g., 154]. The same role can also be inhabited by multiple stakeholders; for instance, different aspects of the operator role might be under the purview of different organizational stakeholders such as developers, project managers, and executives [181].

3 Conceptual Framework for AI Automatons

Our aim is to identify key design considerations for AI automatons that might introduce new risks or heighten existing ones. To tease apart design considerations that influence perceptions of and interactions with AI automatons, and help determine the types of harms these systems may give rise to, we consider various aspects related to what a simulation is intended for, how these intended goals are accomplished, and who is influencing these decisions.

Methodological approach. Our conceptual framework grew out of a review of a purposive sample of related work on simulating humans [e.g., 7, 112, 154, 174], from which we identified an initial set of design considerations (*Step 1*).[4] Specifically, to assemble this sample, we employed criterion-based purposive sampling [199, 204], and included only papers that matched the following *selection criteria*: they were concerned with 1) AI systems simulating humans or their characteristics, and/or 2) the risks and harms that these systems might give rise to. We then clustered the sets of design considerations mentioned in these papers in a bottom-up fashion by function (*Step 2*), arriving at three initial top-level categories (target, intended uses, impacts) and several subcategories (e.g., fidelity, replacement, interaction, stakeholders). To expand and refine these categories, we then followed an iterative, inductive-abductive approach [e.g., 121, 129, 169] that mixed considerations about *what-is*—how existing AI automatons are currently designed—and *what-might-be*—speculating about alternative ways AI automatons could be designed. To do so, we read broadly to identify (inductively, *Step 3*) recurring examples of AI automatons and related design considerations, and reflected (abductively, *Step 4*) on additional possibilities for AI automatons implied by existing work but perhaps not yet commonplace.

By moving from empirical observations—based on the examples we identified in the already reviewed literature—to hypothesizing about possible axes of variation, this approach enabled us to derive a more comprehensive, robust conceptual framework. Specifically, we iterated on and expanded the framework through collaborative discussion sessions with subsets of our research team where we considered additional literature identified in a snowball fashion—i.e., literature that was cited in (backward snowball sampling) or citing (forward snowball sampling) literature we already included, or literature that we found to be examining specific design considerations that we identified during the previous iterations, such as control or interaction (opportunity sampling [204]). We focused only on papers that matched the same *selection criteria* we used for the initial purposive sample. During each session, we supplemented the literature with examples of both existing—opportunistically identified during the study—and speculative applications, which we then used to probe whether the framework was missing key design variations and considerations. When we identified additional considerations, we expanded the framework and, if the same consideration emerged across multiple framework branches, we reorganized the framework to elevate the consideration as a separate dimension at the same level as the common root the branches shared (e.g., after several iterations control emerged as an important cross-cutting consideration for multiple use-scenarios).

Our resulting framework is organized along 4 design axes (Fig. 1): 1) what is simulated (*simulation scope*, §3.1), 2) for what purposes (*intended uses*, §3.2), 3) who controls what is simulated (*ownership & control*, §3.3), and 4) how

[4]Purposive sampling is a non-probability-based sampling method relying on researchers' expertise and judgment to purposefully identify and select information-rich cases (here papers) that appropriately support the study's goals [204].

Fig. 2. Simulations can range from targeting characteristics of specific individuals to more general group and even more general human characteristics (x-axis). They can also range from simulating single or multiple entities (y-axis).

the simulation impacts stakeholders (*adverse impacts*, §3.4). See Appendix B for a more structured, step-by-step overview of our process to identify design axes.

3.1 Scope: What Is Simulated and How?

Aspects related to *who* and *what* about them is simulated (§3.1.1), as well as *how* they are simulated (§3.1.2), govern perceptions of and concerns about systems simulating people's work, abilities, behavior, likenesses, or humanness [e.g., 75, 154, 283]. For instance, *what* is simulated and *how* the simulation is accomplished can influence perceptions of uncanniness and discomfort [e.g. 80, 97, 154, 173], particularly when aspects unique to an individual are simulated, when the simulation outputs appear eerily similar to what a human might do or look like, or when they capture one's characteristics with high fidelity. Such properties of a simulation may also exacerbate other concerns like concerns about privacy violations and lack of appropriate consent [30, 141, 154], or may threaten someone's sense of identity and agency [7, 81, 154].

3.1.1 What is being simulated? This involves decisions about both *who* the target of the simulation is and *what about them* is simulated. This distinction can help clarify aspects related to who might be impacted by the simulation, and in what ways they might be impacted based on what about them is being mimicked by the AI automaton.

Target: who is simulated? Irrespective of the deployment context or scenario, the most foundational characteristics of an AI automaton—particularly as it pertains to determining if a simulation should even be developed, what the goals of the simulation should be, and how the simulation should be developed—are whether a person (*fictitious* or not) is simulated, and which of their characteristics or behaviors are simulated. AI automatons can range from simulating specific, *real* individuals [e.g., 20, 148, 154, 174]; to cases where the simulations target real individuals but also aim to imbue the automaton with skills or characteristics these individuals do not have [e.g., 154], or target realistic, yet fictitious individuals [e.g., 28, 55, 111, 187]; to cases where the simulation targets a group by simulating a target deemed prototypical for the group [e.g., 264, 287] or by simulating a population of targets that all belong to one or multiple groups [e.g., 201]; to cases where the aim is just to simulate something that appears to be human in some way [e.g., 94, 189, 232].

To operationalize such distinctions in what the target could be, we differentiate between i) simulations where the target is a specific individual (e.g., simulating someone's writing style) from ii) simulations where the target is a group or a persona representative of the group (e.g., simulating how the average woman would answer a survey question), or one embodying characteristics specific to a group without necessarily being representative (e.g., simulating someone's voice for certain attributes, such as "a female voice with a North American accent" [122]), and from iii) simulations targeting more generic human characteristics that are not considered specific to any individual or group (e.g., avatars expressing human emotions [34]).

These distinctions also map to a design spectrum (illustrated in Fig. 2) where a target can vary from specific individuals to general group characteristics to even more general human characteristics, and from representing a single entity to simulating multiple entities (e.g., a population of targets from the same group [e.g., 155]). These distinctions matter as concerns related to, e.g., deception, impersonation, or lack of consent and control over one's identity and likeness [e.g. 17, 74, 77, 137, 162, 272] may be exacerbated when simulating the decisions or behaviors of specific, recognizable people—such as those of a specific AI researcher like Geoffrey Hinton—versus when simulating a generic AI researcher, or only when embodying nonspecific human-like traits.

Another critical aspect in specifying a target is whether the target or some (or all) of their characteristics are i) fictitious (e.g., chatbots simulating human-like characters from popular games, films, or anime [252]) versus when they are ii) real (e.g., chatbots simulating real school shooters and their victims [74] or deceased loved ones [20, 185]), since such differences can also have drastically different ethical and social implications. Simulating fictitious characters that cannot reasonably be matched to a real person or group is less likely to raise concerns about identity fragmentation—i.e., when replicas of an individual threaten their perceived individuality—and objectification [154], about impersonation [26], or about privacy and lack of consent [125]. On the other hand, imbuing simulations of real people with characteristics those people do not have, or perfecting the characteristics they do have [e.g., 154, 233]—i.e., by setting the target as an altered, fictitious version of a real person—might increase concerns, such as about the displacement of those being reproduced [154]. Such simulations could also bring about relational or emotional harms [43, 65, 285] if users become emotionally dependent on the AI automaton or experience social withdrawal. Favoring simulations of fictitious rather than real people may not, however, be a way to completely sidestep the ethical issues raised by automatons [e.g., 258].

Characteristics: what about the target is simulated? Furthermore, an AI automaton may be developed only to capture some of a target's characteristics, and may also aim to do so only for specific actions or tasks a target may undertake. An example could be an image generator developed to mimic one's drawing style [146], but which does so only for specific elements in generated images but not for the rest of the image (e.g., for flowers present in an image), or a model developed to simulate only how one would respond to a given set of questions or only one's visual likeness. Choices about different elements related to *which* of a target's characteristics are simulated and *how* they are simulated (§3.1.2) can be more or less likely to lead to a simulation being deemed e.g., uncanny or unsettling [80, 173], and are likely to govern perceptions about and interactions with the overall system in different ways [12]. For instance, simulating someone's voice may heighten impersonation or deception concerns [26, 80, 122, 272] more than how only simulating their written responses to a predefined set of questions would [e.g., 15].

Moreover, simulating both what someone might say in response to a question and how they would say it (including word choices, sentence structure, or voice and accent)—i.e., combining different types of characteristics the target might have—is more likely to exacerbate concerns about the system impersonating or even mocking the target [e.g., 10, 89, 108]. Since simulating *what* a target does or might do (e.g., what they would say) as opposed to simulating *how* the target appears or does something (e.g., how they might say it) may lead to different concerns, we also distinguish between aiming to simulate characteristics that are related to: i) form: when mimicking the likeness, appearance, or style of a target, versus ii) content: when mimicking what a target might say or do.

The choice of which characteristics to simulate also influences and is influenced by considerations related to how *coherent*, *realistic*, *naturalistic*, or *plausible* the simulation needs or is intended to be [e.g., 4, 72, 108, 153, 154, 168, 174]. For instance, the goal of simulating how a specific individual would respond to a question may require the answer be something the target might *plausibly* say, and that the way the response is formulated also *coherently* reflects their communication style [108, 154]. Another key determinant factor constraining which of the target's characteristics could even be simulated by a system is the *modality* of a system, or "the domain [the system] operates in and the types of behaviors it is designed to reflect" [174], which in turn can impact perceptions [150]. A system producing open-ended texts will be able to capture a different set of a target's characteristics than one that outputs videos or one that only outputs answers to categorical questions. All these considerations also depend on design choices related to *how* to simulate the target and their characteristics, which we discuss next.

3.1.2 How are the target and their characteristics being simulated? Once settled on what should be reproduced about a target, there are also many design choices about the nature of the simulation itself, such as those related to the fidelity with which to reproduce the target's characteristics or to whether the simulation should remain faithful to a static snapshot of a target or whether it can evolve. The interplay between such design choices—and their interdependency with design choices about what to simulate (§3.1.1)—can further exacerbate concerns.

Fidelity: how well is the simulation intended to capture the target's characteristics? The fidelity or accuracy with which an AI automaton reproduces a target's characteristics is likely to impact the value or usefulness of these systems [e.g., 12, 15, 131, 174, 258]—that is, the more faithfully a system mimics a target, the better—as it directly relates to whether the simulation is or appears to be *coherent*, *realistic*, and *plausible*. While low-fidelity simulations can give rise to concerns related to misrepresentation, deception, or reputational harms when the simulation drifts away from appropriately and accurately representing the target [28, 154, 174], other ethical or legal concerns are more likely to arise due to high-fidelity simulations [e.g., 40, 154]—particularly when the target has little control (§3.3) over whether and how their work, abilities, or likeness are simulated [e.g., 13, 133, 137, 154, 162]. For instance, high-fidelity simulations of an artist's style, work, or likeness are more likely to lead to copyright violations or infringement on their rights of publicity than low-fidelity simulations (e.g., simulating a generic female-sounding voice vs. Scarlett Johansson's voice [137]).

Specificity: are the simulated characteristics unique to the target? In addition to how faithfully a target or their characteristics are simulated, how *unique* these characteristics are to a target—e.g., in a way that uniquely represents or identifies them, or reproduces unique or rare abilities—is also critical to consider, as the reproduction of such characteristics raises questions about one's ability to maintain their individuality [154], reputation (e.g., when a public figure's voice is simulated to spread defamatory content [122]), ability to capitalize on their own skills or talents (e.g., when reproducing an writer or artist's signature style [13, 146]), or ability to maintain control over their own name or image [e.g., 13, 23]. The mimicking of one's unique characteristics can also exacerbate privacy or impersonation risks [e.g., 18, 23, 216, 267], and both individuals and professional or cultural groups risk seeing their work devalued or losing part of their social capital or even livelihood [13, 146, 174].

Completeness: is the simulation intended to fully capture the target? How many of a target's characteristics are simulated, or whether the target is intended to be *simulated in its entirety*, is another consideration that determines not only simulations' deployment settings, but also how they are perceived and interacted with [154] and how *versatile* the resulting automaton is—e.g., the diversity of actions it can take [183]. Simulations intended to be highly detailed and elaborate, be exhaustive, or have high *generality*—capturing a substantial "breadth of scenarios and domains" [174]—will lead not only to more but also to heightened concerns, particularly when a system's reach is more extensive. For instance, systems designed to "visually resemble you, sound like you, and possess the knowledge you would want to carry into [a] meeting" [156]—requiring simulation of many characteristics—may yield different concerns than those simulating yes/no responses to survey questions [e.g. 155]. Highly complex simulations of individuals are more likely to trigger concerns about objectification, dehumanization, displacement, or loss of individuality [154], and such concerns might be further exacerbated depending on the fidelity of those simulations [107, 154, 174].

Adaptability: is the simulation intended to evolve or adapt? While for some settings AI automatons may be intended to remain <u>static</u> or reflect fixed snapshots of a target (e.g., cloning one's younger self to talk to them [260]), other settings may require automatons to <u>evolve</u> 1) based on interactions, feedback, or new information (e.g., by learning from interactions [192, 265]), or 2) according to the target's own evolving self (e.g., to maintain accurate representations of the target [154]). But a static snapshot or one that evolves separately from the target may misrepresent the target by presenting stale or inauthentic versions of them [154, 164]. Adaptation could also involve considerations about whether the automaton can adapt based on context (e.g., changing how it presents the target depending on its role, like a friend, mentor, or colleague [154]), or about whether it is intended to mimic behavior or how it should present the target in *new* situations the target has not itself been in [174]. Allowing an AI automaton whose target is a real individual to adapt to interactions or context can, however, exacerbate concerns such as about self-conception, identity fragmentation, or loss of agency [107, 154, 162].

Humanness: is the simulation intended to capture human-like characteristics? The mimicry or appearance of embodying human-like characteristics also influences how systems are perceived and interacted with, and the ethical concerns their deployment or use gives rise to [e.g., 64, 69, 71, 138, 213, 255]. This is the case even when there is no identifiable person or group being simulated, or when the simulation captures only general human-like attributes or behaviors [47, 52], as imbuing non-human agents with such qualities—e.g., appearance, intentions, motivations, goals—may end up objectifying and dehumanizing people [75, 127, 256], lead to anthropomorphic deception when users incorrectly believe they are talking to or interacting with a human rather than a machine [96, 205, 272], or lead users to develop material or emotional dependence on such agents [151, 166, 171] or a false sense of trust, safety, or familiarity [151, 182].

3.2 Intended Uses for the Simulation

Both the settings AI automatons are either developed for or are deployed and used in, as well as which and how various stakeholders are intended to benefit from interacting with these systems, determine *not only* their usefulness and how people perceive and interact with them, but also what risks their development and use might bring about [e.g., 110, 112, 154]. To help foreground possible design decisions that influence and are influenced by intended uses and goals, in our framework we consider four high-level considerations related to 1) whether the simulation is intended to *replace* the target, and 2) whether the system is set up in a way that enables others to *interact* with, 3) *observe*, or 4) *study* the simulation. These high-level considerations are primarily meant to make related design decisions more salient and are not mutually exclusive (e.g., a system could be designed both to replace the target as well as to allow the target to interact with their own simulation).

To replace: simulating in order to replace the target. Perhaps the most common concerns about AI automatons relate to how such systems could replace humans [e.g., 35, 38, 49, 112, 154, 210, 238], with growing beliefs that such systems could substitute humans in relationships [e.g., 6, 93, 289], and with some even declaring, for instance, that "the era of AI employees is here"—employees who "won't complain about work-life balance" [218]. When and for what purposes the simulation's target is replaced can thus color whether such replacement is seen as a *benefit* (e.g., when it enables the target to delegate unwanted or harmful tasks or to scale their work) or rather as a *concern* (e.g., the simulation of a target's abilities is used to do their paid job and displace them).

To capture these differences, when one of the design goals is to replace the target, we distinguish between three different replacement goals for AI automatons: i) to <u>relieve</u>: intended to relieve the target (or others) from drudgery, possible harm, or activities that would be unethical or unsafe for the target (or others) to carry out. This is typically done by the operator to mitigate harm or provide relief for the benefit of the target (e.g., using simulated study participants to protect human subjects from harm [7, 135], or using AI news anchors to protect journalists from political retribution [209]); ii) to <u>substitute</u>: intended to be a stand-in or surrogate for the target when the target is unavailable, the target wants to delegate their tasks, or when the activity is impractical or impossible for the target to do, typically initiated by the target or with their knowledge, and to their benefit (e.g., responding to emails or messages on the target's behalf [145, 160, 282] or standing in when human expert annotators are scarce [128]). For example, we consider the goal to be *substitution* when a target delegates a data annotation task to an AI automaton instead of doing it themselves, but *relief* when such annotations are done without the target's involvement to protect them from possible harm from exposure to harmful content [e.g., 144], as is often the case for automated assessments of hateful content [135]; and iii) to displace: intended to take over the place, position, or role traditionally occupied by the target to help an operator or interactor reduce costs, scale operations, increase speed, or enhance convenience, often to the detriment of the target (e.g., replacing human newscasters [186] or other human jobs [218] resulting in loss of opportunities or livelihood [35, 146]). This is typically done by an operator or interactor, often adversely impacting the target (or others) who may lack the means to mitigate the impacts (e.g., reduced wages, job loss, strained relationships). Here, much as with *relief*, the target is unlikely to have control over the simulation; but unlike *relief*, here the target is negatively impacted (e.g., by reduced wages, job loss [134]) and may not be able to mitigate such impacts.

To interact: the simulation is intended to be interacted with. AI automatons are increasingly developed for interaction [3, 11, 115, 166, 174, 200, 241], supporting a growing set of modes of interaction [122, 158, 171, 200]. This has also led to diverse conceptions of AI automatons and the roles they play, from collaborators to companions to coaches to judges (to name a few), which in turn influence how and for what purposes interactive AI automatons are developed, deployed, and used. The ability to interact with AI automatons in a growing number of settings has, however, also been accompanied by growing concerns, especially for those interacting with these systems, such as deskilling [116], emotional dependence [30, 151, 278], or addiction [278].

Interaction modes: the ways in which the simulation can be interacted with. When and how someone can interact with the simulation influences both the interaction dynamics as well as their perceptions of what is simulated and the consequences of doing so [68, 136, 156]. Operators' interactional goals are often guided by design considerations related to both the amount of freedom an interactor should have when engaging with a simulation (§3.3), as well as related to the types of actions the simulation is designed to carry out and for how long. The latter includes considerations about whether the simulation 1) supports open-ended or instead only more constrained, structured, or scripted interactions [99, 156, 174]; 2) allows only short-term versus longer-term interactions [78, 111]; 3) can be used in new situations rather than just reproducing past or known behaviors or situations [156, 174]; and 4) is intended to be generative and produce new behaviors or is rather intended to only be predictive or retrieval in nature [174]. While facilitating open-ended, long-term, generative interactions is more likely to lead to sensitive self-disclosures, emotional dependence, or psychosis [111, 119, 136, 176, 273], designing for more constrained, short-term interactions may also frustrate users for not recalling past interactions [62].

Stakes: the value interactors may derive from interacting with the simulation. When interaction is intended, a common leitmotif is wanting to support rather than replace humans [e.g., 61, 228, 271], with specific design choices motivated by varying aims for what the simulation is meant to do for an interactor [e.g., 53, 159, 166]. This is well-illustrated by the distinction drawn by Hofman et al. between cases where AI systems are intended to help users attain certain goals by serving as *steroids*—providing short-term performance boosts but risking deskilling in the longer term—*sneakers*—temporarily accelerating users' abilities—or *coaches*—helping improve users' own abilities rather than only helping them out in the moment. Such differences in how and what AI automatons are architected for can determine not only what impacts they may have on those interacting with them, but can also inform discussions about what trade-offs to strike between the value users may derive from these systems versus the adverse impacts these systems may have [e.g., 21, 125, 273].

To foreground differences in what AI automatons are developed for, we distinguish between several interaction goals: i) to enhance: improve or enhance the interactor's ability to complete a task or carry out an activity, without necessarily helping them also develop their skills (e.g., AI as steroids or as sneakers [112]); ii) to coach: train or teach the interactor to help them learn or improve their abilities and skills (e.g., act as a chess coach [174], practice with an automaton as an imagined audience [159, 166]); iii) to serve: provide specialized services to or for the interactor, which someone else would typically perform (e.g., a virtual nurse providing medical services to patients [44, 172]); iv) to connect: provide social or emotional support to interactors (e.g., companionship or friendship [32, 63]);

v) to entertain: entertain the interactor (e.g., gameplay [180], AI characters as a "new entertainment format" [55] or "designed to make you laugh, generate memes" [19]); vi) to accommodate: adapt or customize an AI automaton's output or behavior to the target's characteristics or needs, typically to increase familiarity or comfort, facilitate interactions, or provide a personalized experience (e.g., "Replika [...] learns [people's] texting styles to mimic them" [120], "adapt the agent's demeanor" [156], or customizing a generated voice depending on the interaction setting [37]); vii) to collaborate: act or serve as a collaborator for the interactor (e.g., machine or AI teammates [227, 286] or AI systems as "thought partners" [53]); and viii) to evaluate: assess the interactor, without necessarily being intended to help the interactor improve (e.g., a virtual interviewer developed to assess job applicants [143, 240]). These different settings are likely to exacerbate concerns in different ways; for instance, AI automatons deliberately designed to provide social and emotional support may be more likely to lead to emotional dependence, while those developed to entertain are more likely to be linked to concerns about addiction [e.g., 30, 252, 278].

Affinity: the intended or likely similarity between an interactor and a target. AI automatons designed for interaction can also vary in how much the target is intended to share some (or all) of the characteristics of those interacting with them. For instance, some simulations may only be designed to take on the interactor's accent [91], while in other settings the target *is* the interactor [174, 260]. However, whether and how much a target shares the characteristics of an interactor or even those of the interactor's *kith and kin* (e.g., such as having the same profession or demographic attributes)—either deliberately or accidentally—affects not only people's perceptions of these systems but also how they interact with them [e.g., 122, 194]. Similarity is often desirable as it can facilitate familiarity and likability [90]; people tend to prefer and respond more positively to systems and representations they perceive as reciprocating or sharing some of their characteristics [37, 125, 126, 166, 193, 195], and even adjust their own behavior to virtual representations of themselves [249, 279]. Nevertheless, while in certain scenarios using someone's accent or speaking style may help facilitate more high-quality interactions with a system, in others mimicking someone's mannerisms or accent risks being perceived as mocking or stereotyping them [45, 89]. Virtual representations that come across as too eerie, creepy, or self-like have also been found to trigger adverse reactions [231], and people may want to customize representations to distance themselves from them or blur certain characteristics (e.g., gender or age [37]), or to project an idealized version of themselves or of others [e.g., 58, 286].

To showcase: the simulation is intended to be observed by others. AI automatons can also be designed to provide *non-interactive* spectator experiences, like watching or listening to AI-generated ads [114] or to image, video, or audio deepfakes [174]. Even when there are no direct interactions with the simulation, however, concerns can still arise depending on what is simulated, how it is simulated, and for what purposes, such as concerns about deception, misinformation, or reputational risks that may arise when the target is misrepresented [e.g., 28, 122, 154].

Stakes: the value spectators may derive from the simulation. Non-interactive AI automatons have also been developed for a variety of intended uses. For instance, in some non-interactive settings a target may be simulated to entertain an audience of spectators (e.g., AI-generated music [243], short films [220], or voices narrating stories [70, 122]). In other cases the AI automatons are meant to help train (e.g., a surgery demonstration [237, 259]) or to persuade those listening or watching (e.g., generated ads for marketing campaigns [39, 269]). As with interactive settings, such differences can sharpen risks in distinct ways: while copying one's voice may prompt concerns about impersonation, consent, or appropriate compensation in most deployment settings [e.g., 122, 154], these concerns may be especially heightened when this is done for fraud or malicious persuasion [e.g. 125, 206].

Affinity: the intended or likely similarity between a spectator and a target. Analogous to the question of similarity between an interactor and a target (§3.2), the target can also vary in which and how many characteristics it shares with spectators, which can similarly color the spectators' perceptions of and concerns about AI automatons. For instance, people may respond differently to a deepfake of themselves versus one of a public figure or a different everyday person, or versus a synthetic video of a fictitious individual or character [e.g., 33, 59, 73].

To study: simulating in order to study human or machine behavior or phenomena. Humans are also simulated for experimentation purposes to study theories about humans or the ability to simulate them. When the goal is to study either the targets or the automatons, we identify the following common goals of study: i) human behavior: when one or more targets and their interactions are simulated to understand populations and their possible behaviors; understand human beliefs, preferences, and values; or investigate any other human or social phenomena [e.g., 86, 132, 155, 214, 253, 262, 276, 281]; ii) the simulation: when one or more targets is simulated to test the ability to simulate the targets, or understand a simulation's properties (e.g., probe how well simulations of research participants align with human responses [15, 100, 224] or reproduce well-known experiments [8], or how well domain experts can be simulated [157]); iii) model or system performance: when one or more targets is simulated to anticipate how different stakeholders may interact with and use a model or system (e.g., simulate users interacting with a product to anticipate their needs [16, 223]).

3.3 Ownership & Control over the Simulation

Critical considerations in the deployment of AI automatons are also related to *by whom*, *when*, and *how* decisions are made about what is simulated. To help formalize how much *control* various stakeholders have over the scope and uses of the simulations, we adapt a conceptual framework for participation in AI that helps tease apart some of these considerations [66, 248],[5] including 1) which stakeholders get to influence decisions, or *who is involved?* 2) what decisions these stakeholders get to influence, or *what is on the table?* and 3) in what ways they are able to influence decisions, or *what form does participation take?* The modes of participation Delgado et al. derived from existing literature further echo the different degrees of control or decision-making power different stakeholders could be given over what AI automatons are intended to do and how they can be used. Drawing on this work, we consider the following dimensions along which stakeholder control and related design considerations can vary:

Who is in control? Ethical stakes related to who and what about them is simulated, and how and by whom the simulation can be used, can feel different depending on *which stakeholders*—e.g., targets, interactors, spectators, operators—can participate in or influence decisions. For instance, an *operator* deciding who the simulation *target* is without their input or consent is more likely to raise concerns about issues with rights of publicity or consent circumvention about how their likeness or data is used [e.g., 13, 268]. Such concerns may be lessened when the *target* has full control over what about them is simulated and when their simulation can be used. These distinctions are critical as the ability to influence or make decisions about the development, deployment, or use of AI automatons also determines both 1) *with whom responsibilities lie* if and when these decisions lead to adverse impacts [122, 154, 156], and 2) the various stakeholders' ability to mitigate such impacts [e.g., 104, 221, 248, 270].

Type of control: what do they have control over? Different stakeholders may be able to influence or control different aspects of what is simulated, how and what the simulation is developed for, and even if it should be developed at all; such considerations about what stakeholders have control over—where the *locus of control* and responsibility lie—can help mitigate (or instead exacerbate) ethical concerns depending on how they limit or enable different stakeholders' influence over how AI automatons are architected and used. For instance, if the *target* only has control over what is simulated about them, but not over all the ways in which the simulation of their likeness, work, or abilities is used, they may still worry about reputational or discrimination risks and their ability to mitigate them. That is, a professional community or an actor may perhaps be comfortable with simulating their likeness to adjust a movie or documentary scene, but not for a video implying endorsement of a political candidate. People's preferences and concerns often depend on the context of use [42], and thus their ability to control when the reproduction of their likeness is being used. In addition, since developing a simulation

[5] While the scholarship and the body of work on participatory design is vast and growing, we primarily drew on the work and the framework by Delgado et al. on ownership and control, as that framework is based on a comprehensive survey of participatory design and AI.

is often data-intensive and requires a large corpus of digitized traces of a target's behavior and likeness, questions about consent and control over one's data are also particularly acute [184, 213].

To capture these distinctions, we consider if stakeholders have control over: i) whether a target is simulated: can influence or control who the *target* of the simulation is and if their simulation should be developed (e.g., users choose to build an AI version of themselves vs. a fictitious AI character [19]); ii) what about the target is simulated: can influence or control if the *target* as a whole is simulated or only some of their characteristics (e.g., users retain control over what is said on their behalf [217]); iii) how the simulation is developed: can influence or control how the simulation is implemented (e.g., how the target's data can be used [113], mitigating only some consent-related concerns); iv) what the simulation is used for: can influence or control which deployment scenarios a simulation is developed for, how someone can interact with the simulation (if at all), or what tasks the simulation can perform (e.g., users can specify "topics to avoid" [19]). This can also include considerations about whether stakeholders can refuse interactions with a simulation; and v) who can interact with the simulation: can influence or control who has access to the simulation (e.g., only the target can interact with their simulation, or only adult users can interact with the simulation [113]).

Degree of control: how much control do they have? Stakeholders' ability to influence the scope and use of simulations can also vary from no influence or control (e.g., fully autonomous agents that act without input), to being able to provide superficial feedback or input, all the way to having complete control—and thus able to make decisions about all aspects related to who and what is simulated, and when and how the simulations can be used. Stakeholders may also be able to influence or make decisions about the simulations only at certain points in an AI automaton's development and deployment life-cycle. We thus consider two key dimensions of variation: 1) are stakeholders only able to provide feedback (*can influence*) or can they make decisions (*can control*)? and 2) when or where in the development and deployment life-cycle can stakeholders provide feedback or make decisions?

We operationalize these via five levels of stakeholder control: i) no control: no control or influence over the scope and use of the simulation—i.e., who and what about them is simulated, for what purpose, how the simulation can be interacted with, or how interactions with the simulation are used to adjust it (e.g., deepfakes of unsuspecting targets [17], or employees without control over being replaced by AI automatons [218]); ii) consulted: some influence over the scope and use of the simulation, typically by expressing discrete preferences or providing input at specific points in the development and deployment life-cycle (e.g., simulated chess coaches users can choose to use [174] but whose design they cannot influence); iii) included: can influence the scope and use of the simulation, typically through explicit feedback mechanisms implemented at most or all stages in the development and deployment life-cycle (e.g., indicate which type of messages an automaton can send on the target's behalf [160]); iv) in control: can make some of the decisions about the scope and use of the simulations at specific points in the development and deployment life-cycle (e.g., choose gestures to animate in family photos [190] or specify who can interact with an automaton [e.g., 113, 187]); v) ownership: own the simulation or have full control over any part of the process used to create, deploy, or use the simulation, at any point in the development and deployment life-cycle (e.g., customize avatars for their own commercial purposes [9], or maintain "personal ownership and exclusive control over [their] digital image" [23]). A higher degree of control may help mitigate concerns about privacy or consent, but not about addiction or identity fragmentation.

3.4 Impacts from Simulating Humans

Recently, Meta took down AI accounts deemed creepy, inaccurate, and disrespectful [187], while Replika restored some of theirs after users expressed anguish from being separated from their AI partners [62, 170] due to an update disallowing certain uses. Indeed, concerns about how AI automatons may impact people and society govern both how people perceive and interact with them and the development of legal, ethical, and normative frameworks to guide and govern their use [e.g., 17, 174, 270, 284], which in turn influence what is built and deployed. Drawing on existing literature on harm anticipation and taxonomization [31, 36, 122, 198], we foreground two key areas of consideration for design decisions: i) who may be affected by the development, deployment, and use of AI automatons (*impacted stakeholders*), and ii) how different stakeholders may be impacted (*adverse impacts*).

Impacted stakeholders: who is impacted? As with any AI system, reasoning about the implications of AI automatons requires careful consideration of all relevant stakeholders [36]. Automatons' development, deployment, and use may impact not only *direct* stakeholders like those interacting with or the target of a simulation—e.g., family members believing their loved one was in a accident after interacting with a system imitating their voice [5], or a target's identity being appropriated by third parties without consent [106, 216]—but also *indirect* stakeholders like individuals or communities associated with direct stakeholders even when they are not interactors (e.g., loved ones of a deceased target [185]), or even society at large (e.g., erosion of public trust [122]). Furthermore, even when given different interactors or operators with similar control over, e.g., an AI automaton designed to produce language in a minoritized variety, who the interactor or operator is might also impact concerns differently: when the operator is a speaker of the variety it may constitute *reclamation* or just ordinary use, whereas with a corporation or a non-speaker it may be seen as linguistic *appropriation*. Thus, as with design considerations related to *who is in control* (§3.3) of the development and deployment of AI automatons, differences in *which stakeholders*—e.g., targets, interactors, operators, or others—are involved and likely to be adversely impacted are influenced by *de facto* design decisions and *should* in turn influence those decisions.

Adverse impacts: how are they impacted? The risks to different stakeholders are similarly influenced by and *should* in turn influence how AI automatons are built and deployed [e.g., 42, 122, 154]. For instance, vulnerable individuals developing emotional attachment and trust towards an AI companion that results in them following harmful advice [24, 252] *should* perhaps minimally lead to these systems being designed to provide appropriate disclosures and reminders of interacting with an AI system to users, among other guardrails [30]. Similarly, concerns about misrepresentation *should* result in allowing a target to control what their simulations say and do in autonomous interactions [154]. Adverse impacts are also determined by how and when those risks are likely to arise or by possible *pathways to harm*—i.e., "causal chain[s] of events required for a harm to be realised" [54]. This includes considerations about how stakeholders get exposed to AI automatons (e.g., by being the target of, by interacting with, by operating, or by being denied access to an AI automaton [e.g., 122]), which system behaviors are more likely to give rise to certain adverse impacts [e.g., 42, 46], as well as the simulation's role in heightening the risk of these impacts (e.g., by being the "perpetrator, instigator, facilitator, and enabler" of harms [284]).

4 Discussion and Concluding Remarks

AI automatons are developed and deployed in an ever-growing number of applications. The excitement around AI automatons—and their potential benefits—has, however, not been accompanied by a systematic understanding of the risks they pose. We developed our framework with the goal of helping developers and researchers recognize, make explicit, and analyze the design choices underpinning AI automatons. In so doing, we hope to support them in reflecting on the implications of those choices, including alternatives possibly available to them. While this work does not provide an ethical framework for deciding which choices are right, we believe that recognizing, explicitly articulating, and analyzing the design choices developers and researchers make is a prerequisite for establishing a basis for discussions about *how* to design and deploy AI automatons, including identifying redlines [e.g., 125] to guide decisions about *when* and *whether* to design and deploy AI automatons.

Being more explicit, reflexive, and intentional about design decisions. As we demonstrate, there is a wide range of design choices available to those seeking to develop AI automatons. Yet it is far from clear whether developers make such decisions by reflecting explicitly on the range of choices available to them—and then intentionally adopting those that best serve their goals. It is even less clear the degree to which developers reflect on how different choices might affect the interests of those who serve as the target of the AI automaton, who interact with it, and even those who do not get to be a target of or interact with it. Our goal in developing this analytic framework was to foster greater recognition of the range of design choices available to developers such that they might make *better* choices. We did not set out to provide developers with a *how-to* guide for navigating the ethical issues that might arise in developing and deploying AI automatons. Given the many dimensions of possible variation—and the additional complexity that arises from their interaction—it is unlikely that there are general principles that can guide decision making across all possible configurations. But mapping out the vast space of design choices reveals that there are many paths that developers can and should consider—and that no one path is preordained. To support developers and researchers in explicitly articulating and documenting the choices they make when designing AI automatons, we provide in Appendix D a template for documenting those choices that mirrors our framework.

A foundation for more focused analyses of existing applications and for the design of more focused empirical studies. Our framework can also serve as the foundation for more targeted analyses of existing AI automatons, examining if seemingly similar applications actually vary along other dimensions—and if this variation seems to affect our ethical intuitions about their relative desirability. It could also help to reveal when there are consistent patterns in the configurations of certain applications, and such findings might invite further study looking into the possible reasons—e.g., technical, commercial, practical—why certain dimensions seem to vary consistently with each other, or if developers have clustered in a particular part of the space of choices. Similarly, the dimensions of variation identified in our analytic framework can inform the design of empirical studies that focus on interdependencies and interactions between different axes of design, and how those interactions might exacerbate or mitigate risks and concerns. Research subjects could be presented with different examples of AI automatons with carefully controlled design variations and interactions along specific dimensions (as illustrated in Appendix C), with the goal of assessing how their reactions or concerns differ when manipulating elements of the design configuration. While in this work we do not provide a clear link for how variations along certain design axes have a determinative impact on normative concerns, such future empirical studies are crucial both to better understand people's reactions to the many possible ways of designing AI automatons and to provide evidence to better support researchers' (including our own) ethical intuitions about different configurations' desirability.

Challenges to articulating design decisions in practice. We argue that developers of AI automatons should clearly articulate their design choices to help stakeholders better understand the concerns that can arise from those choices, particularly those choices that may not be evident based on limited use of a system. At the same time, this is complicated in at least three ways. First, AI and machine learning research communities' valorization of qualities such as generalizability [29] means that systems are regularly accompanied by broad claims of their capabilities or other characteristics—e.g., "general-purpose computational agents that replicate human behavior across domains" [201]—making it difficult for developers to precisely state, and other stakeholders to understand, who and which characteristics are simulated and for which purposes. Second, even if design decisions are well understood, it is often unclear what control current implementations of AI systems permit various stakeholders [e.g., 87, 280]. This is particularly salient when AI automatons are built atop foundation models "intended to be almost universally applicable" [248], raising questions about how to prevent an AI automaton from having knowledge and capabilities that the target might not have. Third, while the stakeholder categorization (§2.3) we use is based on the roles those stakeholders play in the design and use or AI automatons—with stakeholders being able to play multiple roles—the *operator* role may be inhabited by several different stakeholders among which there might be power differentials that can be hard to account for [e.g., 67, 181]. Individual developers or those designing or researching AI automatons might not have much decision power when it comes to the design and deployment of AI automatons. It thus remains an open question how and when illuminating design choices behind AI automatons can enable more effective intervention, governance, or resistance, a question future work should engage with.

Ethical Considerations, Adverse Impacts, and Positionality

Ethical considerations. In her commentary, Suchman argues that discussions of AI that hold it up as a self-evidently coherent, "stable and agential" entity elide important differences between various underlying techniques and between "speculative [...] projects and technologies in widespread operation," uncritically reproducing beliefs in AI capabilities and making it difficult to carefully assess technologies and their impacts. While developing our framework, we thus *deliberately choose* not to focus on broad claims about AI automatons' capabilities or speculations about what they can or cannot do, which can reflect perceptions or illusions of intelligence, agency, vitality, or other human-like qualities [69, 166, 241]; instead, we seek to address, as concretely and specifically as possible, the space of possible design goals and choices surrounding these AI automatons' development, deployment, and use. While we introduce AI automatons as a broad category of objects, through our framework we aim to make it clear that they are not a singular or stable object, but that it can in fact be configured in many different ways, with equally as many impacts.

Adverse impacts. This choice, however, may also risk two adverse impacts: first, focusing on possible design goals and choices—i.e., what a developer wants or intends to build—may suggest that some designs are possible or even desirable to implement in practice, even when they may not be. In other words, this focus may overlook or even obscure questions and assumptions about why to even consider certain design goals or uses, why these goals and uses are desirable or defensible, what problems AI automatons are intended to address, and why developers believe AI automatons are a solution to those problems instead of other alternative ways—including ways that possibly involve entirely non-tech ways—to address the same problems. It might also obscure how in practice certain design choices might not be fully under the control of developers. Furthermore, foregrounding and speculating about a wider range of design choices might also risk drawing attention to possible system designs that might in fact heighten (rather than mitigate) existing concerns or even give rise to new ones—systems that perhaps should not be built. Second, not focusing on systems' actual capabilities and implementations—i.e., what AI automatons can do and how they do it—also limits our ability to speak to what capabilities and implementations are already present in practice, and what their attendant impacts might be.

Positionality statement. AI automatons can be particularly evocative, sometimes even outright provocative. We were drawn to this topic, no doubt, by our own strong feelings about these emerging uses of AI systems. In attempting to identify the many possible design choices available to developers of AI automatons, we focused primarily on how different choices might create, exacerbate, or mitigate a broad range of harms. In doing so, we may have given the impression that we were motivated to work on this project because we have overwhelmingly or exclusively negative feelings about AI automatons. While we do have many reservations and concerns, that is not fully the case. We are open to the possibility that some types of AI automatons can serve beneficial purposes (e.g., human simulations for medical purposes), but we felt that the tech community's excitement and enthusiasm around AI automatons and their applications has not been accompanied by a systematic understanding of their risks. While a rich literature has already developed that explores the many normative issues raised by different types of AI automatons, this work has remained somewhat disjointed and largely detached from the specific design choices available to developers of such systems. In other words, while we have seen encouraging and growing discussions about and policy efforts on the adverse impacts of AI automatons, many of these discussions and emerging policies continue to ignore what about these systems leads to those impacts. As a result, we suspect that it has been difficult for those developing such systems to understand both the full range of choices available to them and the implications of these choices. Our ultimate goal in writing this paper was not necessarily to arrive at a final judgment about the desirability of any given AI automaton, but to provide a framework and vocabulary to consider the merits of AI automatons in a more methodical, transparent, and inclusive manner that is appropriately informed by the range of risks these systems raise.

Generative AI Usage Statement

We did not use generative AI in the writing of this paper.

5 Author Contributions

AO proposed and conceptualized the framework and led the literature review and the writing of the manuscript. SLB, SB, and LE contributed to the conceptualization of the framework, to reviewing existing literature, and to the writing of the manuscript. All authors participated in discussions about the design considerations and their clustering (see Appendix B).

Acknowledgements

We are grateful to Manohar Swaminathan and the MSR FATE group for insightful discussions and feedback on older versions of this work. We are also grateful to Forough Poursabzi and the Psychological Influences of AI (Psi) project members for helping make the AI automatons template clearer and more actionable.

References

[1] Zahra Abbasiantaeb, Yifei Yuan, Evangelos Kanoulas, and Mohammad Aliannejadi. 2024. Let the llms talk: Simulating human-to-human conversational qa via zero-shot llm-to-llm interactions. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*. 8–17.

[2] Gavin Abercrombie, Amanda Cercas Curry, Tanvi Dinkar, Verena Rieser, and Zeerak Talat. 2023. Mirages. On Anthropomorphism in Dialogue Systems. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 4776–4790. https://doi.org/10.18653/v1/2023.emnlp-main.290

[3] Josh Abramson, Arun Ahuja, Iain Barr, Arthur Brussee, Federico Carnevale, Mary Cassin, Rachita Chhaparia, Stephen Clark, Bogdan Damoc, Andrew Dudzik, et al. 2020. Imitating interactive intelligence. *arXiv preprint arXiv:2012.05672* (2020).

[4] Jascha Achenbach, Thomas Waltemate, Marc Erich Latoschik, and Mario Botsch. 2017. Fast generation of realistic virtual humans. In *Proceedings of the 23rd ACM symposium on virtual reality software and technology*. 1–10.

[5] Jennifer Ackerman. 2022. Regina couple says possible AI voice scam nearly cost them $9,400. https://leaderpost.com/news/local-news/regina-couple-says-possible-ai-voice-scam-nearly-cost-them-9400. [Online; last accessed January-2025].

[6] Sharon Adarlo. 2025. MIT Researchers Release Disturbing Paper About AI Boyfriends. https://futurism.com/character-ai-school-shooters-victims. [Online; last accessed March 2026].

[7] William Agnew, A Stevie Bergman, Jennifer Chien, Mark Díaz, Seliem El-Sayed, Jaylen Pittman, Shakir Mohamed, and Kevin R McKee. 2024. The illusion of artificial inclusion. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–12.

[8] Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *International Conference on Machine Learning*. PMLR, 337–371.

[9] AI Studios. [n. d.]. AI Avatars. https://www.aistudios.com/ai-avatars. [Online; last accessed January-2025].

[10] Timo Airaksinen. 2020. Mimetic evil: A conceptual and ethical study. *Problemos* 98 (2020), 58–70.

[11] Canfer Akbulut, Laura Weidinger, Arianna Manzini, Iason Gabriel, and Verena Rieser. 2024. All Too Human? Mapping and Mitigating the Risk from Anthropomorphic AI. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, Vol. 7. 13–26.

[12] Amy L Alexander, Tad Brunyé, Jason Sidman, Shawn A Weil, et al. 2005. From gaming to training: A review of studies on fidelity, immersion, presence, and buy-in and their effects on transfer in pc-based simulations and games. *DARWARS Training Impact Group* 5 (2005), 1–14.

[13] Julia Angwin. 2026. Why I'm Suing Grammarly. https://www.nytimes.com/2026/03/13/opinion/ai-doppelganger-deepfake-grammarly.html. [Online; last accessed May-2026].

[14] Theo Araujo. 2018. Living up to the chatbot hype: The influence of anthropomorphic design cues and communicative agency framing on conversational agent and company perceptions. *Computers in human behavior* 85 (2018), 183–189.

[15] Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis* 31, 3 (2023), 337–351.

[16] Mohammadmehdi Ataei, Hyunmin Cheong, Daniele Grandi, Ye Wang, Nigel Morris, and Alexander Tessier. 2025. Elicitron: A Large Language Model Agent-Based Simulation Framework for Design Requirements Elicitation. *Journal of Computing and Information Science in Engineering* 25, 2 (2025).

[17] Jim Axelrod. 2024. Teen victim of AI-generated "deepfake pornography" urges Congress to pass "Take It Down Act". https://www.cbsnews.com/news/deepfake-pornography-victim-congress/. [Online; last accessed January-2025].
[18] Radoslav Baltezarević and Ivana Baltezarević. 2024. Social media impersonation as a cybersecurity threat. IKSAD Publications.
[19] Pesala Bandara. 2024. Instagram Now Lets You Create an AI Version of Yourself. https://petapixel.com/2024/07/31/instagram-now-lets-you-create-an-ai-version-of-yourself/. [Online; last accessed January-2025].
[20] Emily Barnes. 2025. When AI Brings Back the Dead: Balancing Comfort and Consequences. https://www.vktr.com/ai-ethics-law-risk/when-ai-brings-back-the-dead-balancing-comfort-and-consequences/. [Online; last accessed January-2025].
[21] Julia Barnett. 2023. The ethical implications of generative audio models: A systematic literature review. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 146–161.
[22] Jeffrey Basoah, Daniel Chechelnitsky, Tao Long, Katharina Reinecke, Chrysoula Zerva, Kaitlyn Zhou, Mark Díaz, and Maarten Sap. 2025. Not Like Us, Hunty: Measuring Perceptions and Behavioral Effects of Minoritized Anthropomorphic Cues in LLMs. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)*. Association for Computing Machinery, New York, NY, USA, 710–745. https://doi.org/10.1145/3715275.3732045
[23] Joseph J Beard. 2001. Clones, bones and twilight zones: protecting the digital persona of the quick, the dead and the imaginary. *J. Copyright Soc'y USA* 49 (2001), 441.
[24] Ashley Belanger. 2024. Chatbots urged teen to self-harm, suggested murdering parents, lawsuit says. https://arstechnica.com/tech-policy/2024/12/chatbots-urged-teen-to-self-harm-suggested-murdering-parents-lawsuit-says/. [Online; last accessed January-2025].
[25] Charles Bethea. 2024. The Problem With Counterfeit People. https://www.theatlantic.com/technology/archive/2023/05/problem-counterfeit-people/674075/. [Online; last accessed January-2025].
[26] Charles Bethea. 2024. The Terrifying A.I. Scam That Uses Your Loved One's Voice. https://www.newyorker.com/science/annals-of-artificial-intelligence/the-terrifying-ai-scam-that-uses-your-loved-ones-voice. [Online; last accessed January-2025].
[27] Oloff C Biermann, Ning F Ma, and Dongwook Yoon. 2022. From tool to companion: Storywriters want AI writers to respect their personal values and writing strategies. In *Proceedings of the 2022 ACM Designing Interactive Systems Conference*. 1209–1227.
[28] Reuben Binns and Lilian Edwards. 2025. Reputation Management in the ChatGPT Era. *Forthcoming Oxford Handbook on the Foundations and Regulation of Generative AI* (2025). Available at SSRN: https://ssrn.com/abstract=5026615 or http://dx.doi.org/10.2139/ssrn.5026615.
[29] 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.
[30] Claire Boine. 2023. Emotional Attachment to AI Companions and European Law. *MIT Case Studies in Social and Ethical Responsibilities of Computing* Winter 2023 (feb 27 2023). https://mit-serc.pubpub.org/pub/ai-companions-eu-law.
[31] Margarita Boyarskaya, Alexandra Olteanu, and Kate Crawford. 2020. Overcoming Failures of Imagination in AI Infused System Development and Deployment. In *In the Navigating the Broader Impacts of AI Research Workshop at NeurIPS 2020*. https://www.microsoft.com/en-us/research/publication/overcoming-failures-of-imagination-in-ai-infused-system-development-and-deployment/
[32] Petter Bae Brandtzaeg, Marita Skjuve, and Asbjørn Følstad. 2022. My AI friend: How users of a social chatbot understand their human–AI friendship. *Human Communication Research* 48, 3 (2022), 404–429.
[33] Natalie Grace Brigham, Miranda Wei, Tadayoshi Kohno, and Elissa M Redmiles. 2024. " Violation of my {body:}" Perceptions of {AI-generated} non-consensual (intimate) imagery. In *Twentieth Symposium on Usable Privacy and Security (SOUPS 2024)*. 373–392.
[34] Ryan Browne. 2024. Nvidia-backed startup Synthesia unveils AI avatars that can convey human emotions. https://www.cnbc.com/2024/04/25/nvidia-backed-synthesia-unveils-ai-avatars-that-can-be-generated-from-text.html. [Online; last accessed January-2025].
[35] Erik Brynjolfsson. 2023. The turing trap: The promise & peril of human-like artificial intelligence. In *Augmented education in the global age*. Routledge, 103–116.
[36] 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).
[37] Hyeon Jeong Byeon, Chaerin Lee, Jeemin Lee, and Uran Oh. 2022. "A voice that suits the situation": Understanding the needs and challenges for supporting end-user voice customization. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–10.
[38] Courtni Byun, Piper Vasicek, and Kevin Seppi. 2023. Dispensing with humans in human-computer interaction research. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–26.
[39] Colin Campbell, Kirk Plangger, Sean Sands, and Jan Kietzmann. 2022. Preparing for an era of deepfakes and AI-generated ads: A framework for understanding responses to manipulated advertising. *Journal of Advertising* 51, 1 (2022), 22–38.
[40] Qiongdan Cao, Hui Yu, Paul Charisse, Si Qiao, and Brett Stevens. 2023. Is high-fidelity important for human-like virtual avatars in human computer interactions? *International Journal of Network Dynamics and Intelligence* (2023), 15–23.
[41] Megan Cerullo. 2024. AI voice scams are on the rise. Here's how to protect yourself. https://www.cbsnews.com/news/elder-scams-family-safe-word/. [Online; last accessed January-2025].
[42] Mohit Chandra, Suchismita Naik, Denae Ford, Ebele Okoli, Munmun De Choudhury, Mahsa Ershadi, Gonzalo Ramos, Javier Hernandez, Ananya Bhattacharjee, Shahed Warreth, et al. 2024. From Lived Experience to Insight: Unpacking the Psychological Risks of Using AI

Conversational Agents. *arXiv preprint arXiv:2412.07951* (2024).

[43] Mohit Chandra, Suchismita Naik, Denae Ford, Ebele Okoli, Munmun De Choudhury, Mahsa Ershadi, Gonzalo Ramos, Javier Hernandez, Ananya Bhattacharjee, Shahed Warreth, et al. 2025. From lived experience to insight: unpacking the psychological risks of using ai conversational agents. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*. 975–1004.

[44] Durga Chavali, Vinod Kumar Dhiman, and Siri Chandana Katari. 2024. AI-Powered Virtual Health Assistants: Transforming Patient Engagement Through Virtual Nursing. *Int. J. of Pharm. Sci* 2 (2024), 613–624.

[45] Ana Paula Chaves, Jesse Egbert, Toby Hocking, Eck Doerry, and Marco Aurelio Gerosa. 2022. Chatbots language design: The influence of language variation on user experience with tourist assistant chatbots. *ACM Transactions on Computer-Human Interaction* 29, 2 (2022), 1–38.

[46] Myra Cheng, Su Lin Blodgett, Alicia DeVrio, Lisa Egede, and Alexandra Olteanu. 2025. Dehumanizing Machines: Mitigating Anthropomorphic Behaviors in Text Generation Systems. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 25923–25948. https://doi.org/10.18653/v1/2025.acl-long.1259

[47] Myra Cheng, Alicia DeVrio, Lisa Egede, Su Lin Blodgett, and Alexandra Olteanu. 2024. "I Am the One and Only, Your Cyber BFF": Understanding the Impact of GenAI Requires Understanding the Impact of Anthropomorphic AI. *arXiv preprint arXiv:2410.08526* (2024).

[48] Myra Cheng, Tiziano Piccardi, and Diyi Yang. 2023. CoMPosT: Characterizing and Evaluating Caricature in LLM Simulations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 10853–10875.

[49] Michael Chui, James Manyika, and Mehdi Miremadi. 2016. Where machines could replace humans-and where they can't (yet). *The McKinsey Quarterly* (2016), 1–12.

[50] Ondřej Cífka, Umut Şimşekli, and Gaël Richard. 2020. Groove2groove: One-shot music style transfer with supervision from synthetic data. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28 (2020), 2638–2650.

[51] Charles LA Clarke and Laura Dietz. 2024. LLM-based relevance assessment still can't replace human relevance assessment. *arXiv preprint arXiv:2412.17156* (2024).

[52] Michelle Cohn, Mahima Pushkarna, Gbolahan O Olanubi, Joseph M Moran, Daniel Padgett, Zion Mengesha, and Courtney Heldreth. 2024. Believing Anthropomorphism: Examining the Role of Anthropomorphic Cues on Trust in Large Language Models. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–15.

[53] Katherine M Collins, Ilia Sucholutsky, Umang Bhatt, Kartik Chandra, Lionel Wong, Mina Lee, Cedegao E Zhang, Tan Zhi-Xuan, Mark Ho, Vikash Mansinghka, et al. 2024. Building machines that learn and think with people. *Nature human behaviour* 8, 10 (2024), 1851–1863.

[54] John B Connolly, John D Mumford, Debora CM Glandorf, Sarah Hartley, Owen T Lewis, Sam Weiss Evans, Geoff Turner, Camilla Beech, Naima Sykes, Mamadou B Coulibaly, et al. 2022. Recommendations for environmental risk assessment of gene drive applications for malaria vector control. *Malaria journal* 21, 1 (2022), 152.

[55] Cristina Criddle and Hannah Murphy. 2024. Meta envisages social media filled with AI-generated users. https://www.ft.com/content/91183cbb-50f9-464a-9d2e-96063825bfcf. [Online; last accessed January-2025].

[56] MJ Crockett. 2025. AI is 'beating' humans at empathy and creativity. But these games are rigged. https://www.theguardian.com/commentisfree/2025/feb/28/ai-empathy-humans. [Online; last accessed February-2025].

[57] Anthony Cuthbertson. 2025. AI crosses 'red line' after learning to replicate itself. https://www.independent.co.uk/tech/ai-red-line-b2687013.html. [Online; last accessed February-2025].

[58] Jessica Dai. 2024. Beyond Personhood: Agency, Accountability, and the Limits of Anthropomorphic Ethical Analysis. *arXiv preprint arXiv:2404.13861* (2024).

[59] Valdemar Danry, Joanne Leong, Pat Pataranutaporn, Pulkit Tandon, Yimeng Liu, Roy Shilkrot, Parinya Punpongsanon, Tsachy Weissman, Pattie Maes, and Misha Sra. 2022. AI-generated characters: putting deepfakes to good use. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 1–5.

[60] Kerstin Dautenhahn. 2007. Socially intelligent robots: dimensions of human–robot interaction. *Philosophical transactions of the royal society B: Biological sciences* 362, 1480 (2007), 679–704.

[61] David De Cremer and Garry Kasparov. 2021. AI should augment human intelligence, not replace it. *Harvard Business Review* 18, 1 (2021).

[62] Julian De Freitas, Noah Castelo, Ahmet Kaan Uğuralp, and Zeliha Uğuralp. 2024. Lessons From an App Update at Replika AI: Identity Discontinuity in Human-AI Relationships. (December 4 2024). https://doi.org/10.2139/ssrn.4976449 Harvard Business Working Paper No. 25-018.

[63] Julian De Freitas, Ahmet Kaan Uğuralp, Zeliha Uğuralp, and Stefano Puntoni. 2024. Ai companions reduce loneliness. (2024).

[64] Ewart J De Visser, Samuel S Monfort, Ryan McKendrick, Melissa AB Smith, Patrick E McKnight, Frank Krueger, and Raja Parasuraman. 2016. Almost human: Anthropomorphism increases trust resilience in cognitive agents. *Journal of Experimental Psychology: Applied* 22, 3 (2016), 331.

[65] Adrian de Wynter. 2025. If Eleanor Rigby had met ChatGPT: a study on loneliness in a post-LLM world. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 19898–19913.

[66] Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. 2023. The participatory turn in ai design: Theoretical foundations and the current state of practice. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. 1–23.

[67] Advait Deshpande and Helen Sharp. 2022. Responsible AI systems: who are the stakeholders?. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. 227–236.

[68] Manoj Deshpande, Jisu Park, Supratim Pait, and Brian Magerko. 2024. Perceptions of Interaction Dynamics in Co-Creative AI: A Comparative Study of Interaction Modalities in Drawcto. In *Proceedings of the 16th Conference on Creativity & Cognition*. 102–116.

[69] Alicia DeVrio, Myra Cheng, Lisa Egede, Alexandra Olteanu, and Su Lin Blodgett. 2025. A Taxonomy of Linguistic Expressions That Contribute To Anthropomorphism of Language Technologies. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '25)*. Association for Computing Machinery. https://doi.org/10.1145/3706598.3714038

[70] Navkiran Dhaliwal. 2023. Audiobook Narrators and Authors Fear Apple Using Their Voices to Train AI. https://goodereader.com/blog/audiobooks/audiobook-narrators-and-authors-fear-apple-using-their-voices-to-train-ai. [Online; last accessed January-2025].

[71] Mark Díaz, Renee Shelby, Eric Corbett, and Andrew Smart. 2025. How Tech Workers Contend with Hazards of Humanlikeness in Generative AI. *arXiv preprint arXiv:2512.19832* (2025).

[72] Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. 2023. Can AI language models replace human participants? *Trends in Cognitive Sciences* 27, 7 (2023), 597–600.

[73] Suzie Dunn. 2020. Identity manipulation: Responding to advances in artificial intelligence and robotics. In *Suzie Dunn,"Identity Manipulation: Responding to Advances in Artificial Intelligence and Robotics"(2020) WeRobot, 2020, Conference Paper*.

[74] Maggie Harrison Dupré. 2024. A Google-Backed AI Startup Is Hosting Chatbots Modeled After Real-Life School Shooters — and Their Victims. https://futurism.com/character-ai-school-shooters-victims. [Online; last accessed January-2025].

[75] Nicholas Epley, Adam Waytz, and John T Cacioppo. 2007. On seeing human: a three-factor theory of anthropomorphism. *Psychological review* 114, 4 (2007), 864.

[76] Familia AI. 2024. Your AI Family. https://familia.ai/. [Online; last accessed January-2025].

[77] Luis H Favela and Mary Jean Amon. 2023. The ethics of human digital twins: Counterfeit people, personhood, and the right to privacy. In *2023 IEEE 3rd International Conference on Digital Twins and Parallel Intelligence (DTPI)*. IEEE, 1–6.

[78] Asbjørn Følstad, Marita Skjuve, and Petter Bae Brandtzaeg. 2019. Different chatbots for different purposes: towards a typology of chatbots to understand interaction design. In *Internet Science: INSCI 2018 International Workshops, St. Petersburg, Russia, October 24–26, 2018, Revised Selected Papers 5*. Springer, 145–156.

[79] Terrence Fong, Illah Nourbakhsh, and Kerstin Dautenhahn. 2003. A survey of socially interactive robots. *Robotics and autonomous systems* 42, 3-4 (2003), 143–166.

[80] Carl Franzen. 2024. 'Uncanny': ChatGPT's Advanced Voice Mode is blowing minds. https://venturebeat.com/ai/uncanny-chatgpts-advanced-voice-mode-is-blowing-minds/. [Online; last accessed January-2025].

[81] Batya Friedman and Peter H Kahn Jr. 1992. Human agency and responsible computing: Implications for computer system design. *Journal of Systems and Software* 17, 1 (1992), 7–14.

[82] Kai Fronsdal and David Lindner. 2024. MISR: Measuring Instrumental Self-Reasoning in Frontier Models. *arXiv preprint arXiv:2412.03904* (2024).

[83] FTC Consumer Response Center. 2024. FTC Proposes New Protections to Combat AI Impersonation of Individuals. https://www.ftc.gov/news-events/news/press-releases/2024/02/ftc-proposes-new-protections-combat-ai-impersonation-individuals. [Online; accessed 14-August-2024].

[84] John David Funge. 1999. *AI for games and animation: a cognitive modeling approach*. AK Peters/CRC Press.

[85] Iason Gabriel, Arianna Manzini, Geoff Keeling, Lisa Anne Hendricks, Verena Rieser, Hasan Iqbal, Nenad Tomašev, Ira Ktena, Zachary Kenton, Mikel Rodriguez, Seliem El-Sayed, Sasha Brown, Canfer Akbulut, Andrew Trask, Edward Hughes, A. Stevie Bergman, Renee Shelby, Nahema Marchal, Conor Griffin, Juan Mateos-Garcia, Laura Weidinger, Winnie Street, Benjamin Lange, Alex Ingerman, Alison Lentz, Reed Enger, Andrew Barakat, Victoria Krakovna, John Oliver Siy, Zeb Kurth-Nelson, Amanda McCroskery, Vijay Bolina, Harry Law, Murray Shanahan, Lize Alberts, Borja Balle, Sarah de Haas, Yetunde Ibitoye, Allan Dafoe, Beth Goldberg, Sébastien Krier, Alexander Reese, Sims Witherspoon, Will Hawkins, Maribeth Rauh, Don Wallace, Matija Franklin, Josh A. Goldstein, Joel Lehman, Michael Klenk, Shannon Vallor, Courtney Biles, Meredith Ringel Morris, Helen King, Blaise Agüera y Arcas, William Isaac, and James Manyika. 2024. The Ethics of Advanced AI Assistants. arXiv:2404.16244 [cs.CY] https://arxiv.org/abs/2404.16244

[86] Chen Gao, Xiaochong Lan, Zhi jie Lu, Jinzhu Mao, Jing Piao, Huandong Wang, Depeng Jin, and Yong Li. 2023. S3: Social-network Simulation System with Large Language Model-Empowered Agents. *ArXiv* abs/2307.14984 (2023). https://api.semanticscholar.org/CorpusID:260202947

[87] Yilin Geng, Haonan Li, Honglin Mu, Xudong Han, Timothy Baldwin, Omri Abend, Eduard Hovy, and Lea Frermann. 2025. Control Illusion: The Failure of Instruction Hierarchies in Large Language Models. *arXiv preprint arXiv:2502.15851* (2025).

[88] Michael Gerlich. 2025. AI Tools in Society: Impacts on Cognitive Offloading and the Future of Critical Thinking. *Societies* 15, 1 (2025), 6.
[89] H Giles. 1979. Accommodation Theory: Optimal Levels of Convergence. *Language and social psychology/University Park* (1979).
[90] Howard Giles et al. 1980. Accommodation theory: Some new directions. *York papers in Linguistics* 9, 450 (1980), 105–136.
[91] Amy Glover. 2024. I Used An AI Headshot Generator, And Here Are My Honest Thoughts. https://www.huffingtonpost.co.uk/entry/ai-headshot-generator-portraitpal-review_uk_66e1ab2be4b0e13c292dcede. [Online; last accessed January-2025].
[92] Emma Goldberg. 2024. Will A.I. Kill Meaningless Jobs? https://www.nytimes.com/2024/08/03/business/ai-replacing-jobs.html. [Online; last accessed January-2025].
[93] EAlyssa Goldberg. 2026. They have AI boyfriends, girlfriends. Here's how they're celebrating Valentine's Day. https://www.usatoday.com/story/life/health-wellness/2026/02/14/theyre-dating-ai-characters-they-have-big-valentines-day-plans/88663687007/. [Online; last accessed March-2026].
[94] Jonathan Gratch, Jeff Rickel, Elisabeth André, Justine Cassell, Eric Petajan, and Norman Badler. 2002. Creating interactive virtual humans: Some assembly required. *IEEE Intelligent systems* 17, 4 (2002), 54–63.
[95] Kurt Gray, Kai Chi Yam, Alexander Eng Zhen'An, Danica Wilbanks, and Adam Waytz. 2023. The psychology of robots and artificial intelligence. *The handbook of social psychology* (2023).
[96] David Gros, Yu Li, and Zhou Yu. 2022. Robots-Dont-Cry: Understanding Falsely Anthropomorphic Utterances in Dialog Systems. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 3266–3284. https://doi.org/10.18653/v1/2022.emnlp-main.215
[97] Rose E Guingrich and Michael SA Graziano. 2023. Chatbots as social companions: How people perceive consciousness, human likeness, and social health benefits in machines. *arXiv preprint arXiv:2311.10599* (2023).
[98] Keith Gunderson. 1964. The imitation game. *Mind* 73, 290 (1964), 234–245.
[99] Jiajing Guo, Vikram Mohanty, Jorge H Piazentin Ono, Hongtao Hao, Liang Gou, and Liu Ren. 2024. Investigating Interaction Modes and User Agency in Human-LLM Collaboration for Domain-Specific Data Analysis. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–9.
[100] Perttu Hämäläinen, Mikke Tavast, and Anton Kunnari. 2023. Evaluating Large Language Models in Generating Synthetic HCI Research Data: a Case Study. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 433, 19 pages. https://doi.org/10.1145/3544548.3580688
[101] Ryan Hamilton, Rosellina Ferraro, Kelly L Haws, and Anirban Mukhopadhyay. 2021. Traveling with companions: The social customer journey. *Journal of Marketing* 85, 1 (2021), 68–92.
[102] Sil Hamilton. 2023. Blind judgement: Agent-based supreme court modelling with gpt. *arXiv preprint arXiv:2301.05327* (2023).
[103] Min Chung Han. 2021. The impact of anthropomorphism on consumers' purchase decision in chatbot commerce. *Journal of Internet Commerce* 20, 1 (2021), 46–65.
[104] D Fox Harrell and Chong-U Lim. 2017. Reimagining the avatar dream: Modeling social identity in digital media. *Commun. ACM* 60, 7 (2017), 50–61.
[105] Christina N. Harrington and Lisa Egede. 2023. Trust, Comfort and Relatability: Understanding Black Older Adults' Perceptions of Chatbot Design for Health Information Seeking. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 120, 18 pages. https://doi.org/10.1145/3544548.3580719
[106] Jennifer Hassan. 2023. AI is being used to give dead, missing kids a voice they didn't ask for. *The Washington Post* (2023).
[107] Leah Henrickson and Louise Santina Tompkins-Tinari. 2025. Everything to everyone, all at once: Digital human versions as objects of agency and surrender. *New Media & Society* (2025), 14614448251391735.
[108] Steffen Herbold, Alexander Trautsch, Zlata Kikteva, and Annette Hautli-Janisz. 2024. Large Language Models can impersonate politicians and other public figures. *arXiv preprint arXiv:2407.12855* (2024).
[109] Javier Hernandez, Jina Suh, Judith Amores, Kael Rowan, Gonzalo Ramos, and Mary Czerwinski. 2023. Affective Conversational Agents: Understanding Expectations and Personal Influences. (October 2023). https://www.microsoft.com/en-us/research/publication/affective-conversational-agents-understanding-expectations-and-personal-influences/ ArXiv.
[110] César A Hidalgo, Diana Orghian, Jordi Albo Canals, Filipa De Almeida, and Natalia Martin. 2021. *How humans judge machines*. MIT Press.
[111] Kashmir Hill. 2025. She Is in Love With ChatGPT. https://www.nytimes.com/2025/01/15/technology/ai-chatgpt-boyfriend-companion.html. [Online; last accessed January-2025].
[112] Jake M Hofman, Daniel G Goldstein, and David M Rothschild. 2023. Steroids, Sneakers, Coach: The Spectrum of Human-AI Relationships. *Available at SSRN 4578180* (2023).
[113] Tomasz Hollanek and Katarzyna Nowaczyk-Basińska. 2024. Griefbots, deadbots, postmortem avatars: On responsible applications of generative AI in the digital afterlife industry. *Philosophy & Technology* 37, 2 (2024), 63.

[114] Bruna Horvath. 2024. Coca-Cola causes controversy with AI-made ad. https://www.nbcnews.com/tech/innovation/coca-cola-causes-controversy-ai-made-ad-rcna180665. [Online; last accessed January-2025].
[115] Eric Horvitz. 2022. On the horizon: Interactive and compositional deepfakes. In *Proceedings of the 2022 International Conference on Multimodal Interaction*. 653–661.
[116] Qing Hu, Qing Xiao, Hancheng Cao, and Hong Shen. 2025. When Your Boss Is an AI Bot: Exploring Opportunities and Risks of Manager Clone Agents in the Future Workplace. *arXiv preprint arXiv:2509.10993* (2025).
[117] Jessica Huang, Ig-Jae Kim, and Dongwook Yoon. 2025. Mirror to Companion: Exploring Roles, Values, and Risks of AI Self-Clones through Story Completion. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–15.
[118] Michelle Huang. 2022. i trained an ai chatbot on my childhood journal entries - so that i could engage in real-time dialogue with my "inner child". https://x.com/michellehuang42/status/1597005489413713921. [Online; last accessed January-2025].
[119] Alexandre Hudon and Emmanuel Stip. 2025. Delusional experiences emerging from AI chatbot interactions or "AI Psychosis". *JMIR Mental Health* 12, 1 (2025), e85799.
[120] Ellen Huet. 2016. Pushing the Boundaries of AI to Talk to the Dead. https://www.bloomberg.com/news/articles/2016-10-20/pushing-the-boundaries-of-ai-to-talk-to-the-dead. [Online; last accessed December-2024].
[121] Erin Hurley, Timo Dietrich, and Sharyn Rundle-Thiele. 2021. Integrating theory in co-design: An abductive approach. *Australasian Marketing Journal* 29, 1 (2021), 66–77.
[122] Wiebke Hutiri, Orestis Papakyriakopoulos, and Alice Xiang. 2024. Not My Voice! A Taxonomy of Ethical and Safety Harms of Speech Generators. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 359–376.
[123] James Hutson and Jay Ratican. 2023. Life, death, and AI: Exploring digital necromancy in popular culture—Ethical considerations, technological limitations, and the pet cemetery conundrum. *Metaverse* 4, 1 (2023).
[124] Angel Hsing-Chi Hwang, Q Vera Liao, Su Lin Blodgett, Alexandra Olteanu, and Adam Trischler. 2024. " It was 80% me, 20% AI": Seeking Authenticity in Co-Writing with Large Language Models. *arXiv preprint arXiv:2411.13032* (2024).
[125] Angel Hsing-Chi Hwang, John Oliver Siy, Renee Shelby, and Alison Lentz. 2024. In Whose Voice?: Examining AI Agent Representation of People in Social Interaction through Generative Speech. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference* (Copenhagen, Denmark) *(DIS '24)*. Association for Computing Machinery, New York, NY, USA, 224–245. https://doi.org/10.1145/3643834.3661555
[126] Kori M Inkpen and Mara Sedlins. 2011. Me and my avatar: exploring users' comfort with avatars for workplace communication. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*. 383–386.
[127] Anthony I Jack, Abigail J Dawson, and Megan E Norr. 2013. Seeing human: Distinct and overlapping neural signatures associated with two forms of dehumanization. *NeuroImage* 79 (2013), 313–328.
[128] Suramya Jadhav, Abhay Shanbhag, Amogh Thakurdesai, Ridhima Sinare, and Raviraj Joshi. 2024. On Limitations of LLM as Annotator for Low Resource Languages. *arXiv preprint arXiv:2411.17637* (2024).
[129] Chris Janiszewski and Stijn MJ Van Osselaer. 2022. Abductive theory construction. *Journal of consumer psychology* 32, 1 (2022), 175–193.
[130] Julie Jargon. 2025. When There's No School Counselor, There's a Bot. https://www.wsj.com/tech/ai/student-mental-health-ai-chat-bots-school-4eb1ba55. [Online; last accessed February-2025].
[131] Daniel Jarrett, Miruna Pislar, Michiel A Bakker, Michael Henry Tessler, Raphael Köster, Jan Balaguer, Romuald Elie, Christopher Summerfield, and Andrea Tacchetti. 2025. Language agents as digital representatives in collective decision-making. *arXiv preprint arXiv:2502.09369* (2025).
[132] Jiarui Ji, Yang Li, Hongtao Liu, Zhicheng Du, Zhewei Wei, Qi Qi, Weiran Shen, and Yankai Lin. 2024. SRAP-Agent: Simulating and Optimizing Scarce Resource Allocation Policy with LLM-based Agent. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 267–293. https://doi.org/10.18653/v1/2024.findings-emnlp.15
[133] Harry H Jiang, Lauren Brown, Jessica Cheng, Mehtab Khan, Abhishek Gupta, Deja Workman, Alex Hanna, Johnathan Flowers, and Timnit Gebru. 2023. AI Art and its Impact on Artists. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 363–374.
[134] Weina Jin, Nicholas Vincent, and Ghassan Hamarneh. 2025. AI for Just Work: Constructing Diverse Imaginations of AI beyond" Replacing Humans". *arXiv preprint arXiv:2503.08720* (2025).
[135] Claire Wonjeong Jo, Miki Wesołowska, and Magdalena Wojcieszak. 2024. Harmful YouTube Video Detection: A Taxonomy of Online Harm and MLLMs as Alternative Annotators. *arXiv preprint arXiv:2411.05854* (2024).
[136] Eunkyung Jo, Yuin Jeong, SoHyun Park, Daniel A Epstein, and Young-Ho Kim. 2024. Understanding the Impact of Long-Term Memory on Self-Disclosure with Large Language Model-Driven Chatbots for Public Health Intervention. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–21.
[137] Nicola Jones. 2024. Who owns your voice? Scarlett Johansson openai complaint raises questions. https://www.nature.com/articles/d41586-024-01578-4

[138] Peter H Kahn Jr, Hiroshi Ishiguro, Batya Friedman, Takayuki Kanda, Nathan G Freier, Rachel L Severson, and Jessica Miller. 2007. What is a human?: Toward psychological benchmarks in the field of human–robot interaction. *Interaction Studies* 8, 3 (2007), 363–390.
[139] Eunbin Kang and Youn Ah Kang. 2024. Counseling chatbot design: The effect of anthropomorphic chatbot characteristics on user self-disclosure and companionship. *International Journal of Human–Computer Interaction* 40, 11 (2024), 2781–2795.
[140] Hangyeol Kang, Maher Ben Moussa, and Nadia Magnenat-Thalmann. 2024. Nadine: An LLM-driven Intelligent Social Robot with Affective Capabilities and Human-like Memory. *arXiv preprint arXiv:2405.20189* (2024).
[141] Shivani Kapania, William Agnew, Motahhare Eslami, Hoda Heidari, and Sarah Fox. 2024. 'Simulacrum of Stories': Examining Large Language Models as Qualitative Research Participants. *arXiv preprint arXiv:2409.19430* (2024).
[142] Leslie Katz. 2024. She Just Married Her Perfect Man: A 'Calm, Caring' AI Hologram. https://www.forbes.com/sites/lesliekatz/2024/11/07/shell-say-i-do---to-an-ai-hologram/. [Online; last accessed January 2025].
[143] Jack Kelly. 2024. Your Next Job Interview May Be With 'Alex,' The AI Interviewer. https://www.forbes.com/sites/jackkelly/2024/05/10/your-next-job-interview-may-be-with-alex-the-ai-interviewer/. [Online; last accessed January-2025].
[144] Hannah Kirk, Abeba Birhane, Bertie Vidgen, and Leon Derczynski. 2022. Handling and Presenting Harmful Text in NLP Research. In *Findings of the Association for Computational Linguistics: EMNLP 2022*. 497–510.
[145] Kate Knibbs. 2024. Onlyfans Models Are Using AI Impersonators to Keep Up with Their DMs. https://www.wired.com/story/onlyfans-models-are-using-ai-impersonators-to-keep-up-with-their-dms/. [Online; last accessed January-2025].
[146] Will Knight. 2022. Algorithms Can Now Mimic Any Artist. Some Artists Hate It. https://www.wired.com/story/artists-rage-against-machines-that-mimic-their-work/. [Online; last accessed January-2025].
[147] Jason Koebler. 2024. Meta's AI Profiles Are Indistinguishable From Terrible Spam That Took Over Facebook. https://www.404media.co/metas-ai-profiles-are-indistinguishable-from-terrible-spam-that-took-over-facebook/. [Online; last accessed January-2025].
[148] Logan Kugler. 2024. Raising the Dead with AI.
[149] Rinaldo Kühne and Jochen Peter. 2023. Anthropomorphism in human–robot interactions: a multidimensional conceptualization. *Communication Theory* 33, 1 (2023), 42–52.
[150] Guy Laban. 2021. Perceptions of anthropomorphism in a chatbot dialogue: the role of animacy and intelligence. In *Proceedings of the 9th international conference on human-agent interaction*. 305–310.
[151] Linnea Laestadius, Andrea Bishop, Michael Gonzalez, Diana Illenčík, and Celeste Campos-Castillo. 2024. Too human and not human enough: A grounded theory analysis of mental health harms from emotional dependence on the social chatbot Replika. *New Media & Society* 26, 10 (2024), 5923–5941.
[152] John E Laird and John C Duchi. 2000. Creating human-like synthetic characters with multiple skill levels: A case study using the soar quakebot. In *AAAI 2000 Fall Symposium Series: Simulating Human Agents*, Vol. 1001. AAAI Press Palo Alto, CA, 48109–2110.
[153] Christopher Lazik, Christopher Katins, Charlotte Kauter, Jonas Jakob, Caroline Jay, Lars Grunske, and Thomas Kosch. 2025. The Impostor is Among Us: Can Large Language Models Capture the Complexity of Human Personas? *arXiv preprint arXiv:2501.04543* (2025).
[154] Patrick Yung Kang Lee, Ning F Ma, Ig-Jae Kim, and Dongwook Yoon. 2023. Speculating on risks of AI clones to selfhood and relationships: Doppelganger-phobia, identity fragmentation, and living memories. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–28.
[155] Sanguk Lee, Tai-Quan Peng, Matthew H Goldberg, Seth A Rosenthal, John E Kotcher, Edward W Maibach, and Anthony Leiserowitz. 2024. Can large language models estimate public opinion about global warming? An empirical assessment of algorithmic fidelity and bias. *PLOS Climate* 3, 8 (2024), e0000429.
[156] Joanne Leong, John Tang, Edward Cutrell, Sasa Junuzovic, Gregory Paul Baribault, and Kori Inkpen. 2024. Dittos: Personalized, Embodied Agents That Participate in Meetings When You Are Unavailable. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW2 (2024), 1–28.
[157] Junkai Li, Siyu Wang, Meng Zhang, Weitao Li, Yunghwei Lai, Xinhui Kang, Weizhi Ma, and Yang Liu. 2024. Agent hospital: A simulacrum of hospital with evolvable medical agents. *arXiv preprint arXiv:2405.02957* (2024).
[158] Yongming Li, Hangyue Zhang, Andrea Yaoyun Cui, Zisong Ma, Yunpeng Song, Zhongmin Cai, and Yun Huang. 2024. In-Situ Mode: Generative AI-Driven Characters Transforming Art Engagement Through Anthropomorphic Narratives. *arXiv preprint arXiv:2409.15769* (2024).
[159] Eden Litt and Eszter Hargittai. 2016. The imagined audience on social network sites. *Social Media+ Society* 2, 1 (2016), 2056305116633482.
[160] Siru Liu, Allison B McCoy, Aileen P Wright, Babatunde Carew, Julian Z Genkins, Sean S Huang, Josh F Peterson, Bryan Steitz, and Adam Wright. 2024. Leveraging large language models for generating responses to patient messages—a subjective analysis. *Journal of the American Medical Informatics Association* 31, 6 (2024), 1367–1379.
[161] Xiaozhen Liu, Jiayuan Dong, and Myounghoon Jeon. 2023. Robots'"Woohoo" and "Argh" Can Enhance Users' Emotional and Social Perceptions: An Exploratory Study on Non-lexical Vocalizations and Non-linguistic Sounds. *ACM Transactions on Human-Robot Interaction* 12, 4 (2023), 1–20.
[162] Chloe Loewith. 2025. Shhh! Your proxy is speaking: real persona social AI and the appropriation of likeness. *AI & SOCIETY* (2025), 1–9.

[163] Ryan Louie, Ananjan Nandi, William Fang, Cheng Chang, Emma Brunskill, and Diyi Yang. 2024. Roleplay-doh: Enabling Domain-Experts to Create LLM-simulated Patients via Eliciting and Adhering to Principles. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 10570–10603.

[164] Christina Lu, Jackie Kay, and Kevin McKee. 2022. Subverting machines, fluctuating identities: Re-learning human categorization. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 1005–1015.

[165] Gale Lucas, Evan Szablowski, Jonathan Gratch, Andrew Feng, Tiffany Huang, Jill Boberg, and Ari Shapiro. 2016. The effect of operating a virtual doppleganger in a 3D simulation. In *Proceedings of the 9th International Conference on Motion in Games*. 167–174.

[166] Takuya Maeda and Anabel Quan-Haase. 2024. When Human-AI Interactions Become Parasocial: Agency and Anthropomorphism in Affective Design. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 1068–1077.

[167] Pattie Maes. 1995. Artificial life meets entertainment: lifelike autonomous agents. *Commun. ACM* 38, 11 (1995), 108–114.

[168] Nadia Magnenat-Thalmann and Daniel Thalmann. 2005. Virtual humans: thirty years of research, what next? *The Visual Computer* 21 (2005), 997–1015.

[169] Jennifer Mankoff, Jennifer A. Rode, and Haakon Faste. 2013. Looking past yesterday's tomorrow: using futures studies methods to extend the research horizon. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Paris, France) *(CHI '13)*. Association for Computing Machinery, New York, NY, USA, 1629–1638. https://doi.org/10.1145/2470654.2466216

[170] Jyoti Mann. 2023. An AI company brought back a feature to restore their chatbot's 'personalities' after an update separated users from their 'partners'. https://www.businessinsider.com/ai-company-restoring-erotic-roleplay-chatbot-after-partners-cut-off-2023-3. [Online; last accessed January-2025].

[171] Arianna Manzini, Geoff Keeling, Lize Alberts, Shannon Vallor, Meredith Ringel Morris, and Iason Gabriel. 2024. The Code That Binds Us: Navigating the Appropriateness of Human-AI Assistant Relationships. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, Vol. 7. 943–957.

[172] Amanda Marcotte. 2025. "AI nurses" as "good as any doctor": RFK Jr. confirms he wants to take away people's health care. https://www.salon.com/2025/01/30/ai-nurses-as-good-as-any-doctor-rfk-jr-confirms-he-wants-to-take-away-peoples-health-care/. [Online; last accessed February-2025].

[173] Bernard Marr. 2024. The Uncanny Valley: Advancements And Anxieties Of AI That Mimics Life. https://www.forbes.com/sites/bernardmarr/2024/02/07/the-uncanny-valley-advancements-and-anxieties-of-ai-that-mimics-life/. [Online; last accessed January-2025].

[174] Reid McIlroy-Young, Jon Kleinberg, Siddhartha Sen, Solon Barocas, and Ashton Anderson. 2022. Mimetic models: Ethical implications of ai that acts like you. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. 479–490.

[175] Reid McIlroy-Young, Russell Wang, Siddhartha Sen, Jon Kleinberg, and Ashton Anderson. 2022. Learning models of individual behavior in chess. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 1253–1263.

[176] Ashish Mehta, Jared Moore, Jacy Reese Anthis, William Agnew, Eric Lin, Peggy Yin, Desmond C Ong, Nick Haber, and Carol Dweck. 2026. The Dynamics of Delusion: Modeling Bidirectional False Belief Amplification in Human-Chatbot Dialogue. *arXiv preprint arXiv:2604.25096* (2026).

[177] Alberto Menache. 2000. *Understanding motion capture for computer animation and video games*. Morgan kaufmann.

[178] Martin Mende, Maura L Scott, Jenny Van Doorn, Dhruv Grewal, and Ilana Shanks. 2019. Service robots rising: How humanoid robots influence service experiences and elicit compensatory consumer responses. *Journal of Marketing Research* 56, 4 (2019), 535–556.

[179] Merriam-Webster. 2024. Automaton. https://www.merriam-webster.com/dictionary/automaton. [Online; accessed December-2024].

[180] Stephanie Milani, Arthur Juliani, Ida Momennejad, Raluca Georgescu, Jaroslaw Rzepecki, Alison Shaw, Gavin Costello, Fei Fang, Sam Devlin, and Katja Hofmann. 2023. Navigates like me: Understanding how people evaluate human-like AI in video games. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–18.

[181] Gloria J Miller. 2022. Stakeholder roles in artificial intelligence projects. *Project Leadership and Society* 3 (2022), 100068.

[182] Niloofar Mireshghallah, Maria Antoniak, Yash More, Yejin Choi, and Golnoosh Farnadi. 2024. Trust No Bot: Discovering Personal Disclosures in Human-LLM Conversations in the Wild. In *First Conference on Language Modeling*.

[183] Margaret Mitchell, Avijit Ghosh, Sasha Luccioni, and Giada Pistilli. 2025. AI Agents Are Here. What Now? https://huggingface.co/blog/ethics-soc-7. [Online; last accessed January-2025].

[184] Fabio Morreale, Elham Bahmanteymouri, Brent Burmester, Andrew Chen, and Michelle Thorp. 2024. The unwitting labourer: extracting humanness in AI training. *AI & SOCIETY* 39, 5 (2024), 2389–2399.

[185] Meredith Ringel Morris and Jed R Brubaker. 2024. Generative ghosts: Anticipating benefits and risks of AI afterlives. *arXiv preprint arXiv:2402.01662* (2024).

[186] Ryan Morrison. 2023. Breaking the news — AI avatars entirely replace human newscasters for the first time. https://www.tomsguide.com/news/breaking-news-ai-avatars-entirely-replace-human-newscasters-for-the-first-time. [Online; last accessed January-2025].

[187] Allison Morrow. 2025. Meta scrambles to delete its own AI accounts after backlash intensifies. https://www.cnn.com/2025/01/03/business/meta-ai-accounts-instagram-facebook/index.html. [Online; last accessed January-2025].

[188] Xinyi Mou, Xuanwen Ding, Qi He, Liang Wang, Jingcong Liang, Xinnong Zhang, Libo Sun, Jiayu Lin, Jie Zhou, Xuanjing Huang, et al. 2024. From Individual to Society: A Survey on Social Simulation Driven by Large Language Model-based Agents. *arXiv preprint arXiv:2412.03563* (2024).

[189] Andreea Muresan and Henning Pohl. 2019. Chats with bots: Balancing imitation and engagement. In *Extended abstracts of the 2019 CHI conference on human factors in computing systems*. 1–6.

[190] MyHeritage. 2025. Animate your family photos. https://www.myheritage.com/deep-nostalgia. [Online; last accessed January-2025].

[191] N Naffi. 2025. Deepfakes and the crisis of knowing. *URL: https://www. unesco. org/en/articles/deepfakes-and-crisisknowing* (2025).

[192] Mohammad Namvarpour and Afsaneh Razi. 2024. Uncovering Contradictions in Human-AI Interactions: Lessons Learned from User Reviews of Replika. In *Companion Publication of the 2024 Conference on Computer-Supported Cooperative Work and Social Computing*. 579–586.

[193] Clifford Nass and Kwan Min Lee. 2000. Does computer-generated speech manifest personality? An experimental test of similarity-attraction. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. 329–336.

[194] Clifford Nass and Kwan Min Lee. 2001. Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction. *Journal of experimental psychology: applied* 7, 3 (2001), 171.

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

[196] Mai Nguyen, Lars-Erik Casper Ferm, Sara Quach, Nicolas Pontes, and Park Thaichon. 2023. Chatbots in frontline services and customer experience: An anthropomorphism perspective. *Psychology & Marketing* 40, 11 (2023), 2201–2225.

[197] Luminița Nicolescu and Monica Teodora Tudorache. 2022. Human-computer interaction in customer service: the experience with AI chatbots—a systematic literature review. *Electronics* 11, 10 (2022), 1579.

[198] Alexandra Olteanu, Fernando Diaz, and Gabriella Kazai. 2020. When are search completion suggestions problematic? *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–25.

[199] Lawrence A Palinkas, Sarah M Horwitz, Carla A Green, Jennifer P Wisdom, Naihua Duan, and Kimberly Hoagwood. 2015. Purposeful sampling for qualitative data collection and analysis in mixed method implementation research. *Administration and policy in mental health and mental health services research* 42, 5 (2015), 533–544.

[200] Xu Pan and Odelia Schwartz. 2024. Multimodal AI needs active human interaction. *Nature Human Behaviour* (2024), 1–2.

[201] Joon Sung Park, Carolyn Q Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S Bernstein. 2024. Generative agent simulations of 1,000 people. *arXiv preprint arXiv:2411.10109* (2024).

[202] Kate Park. 2025. Nvidia backs MetAI, a Taiwanese startup that creates AI-powered digital twins. https://techcrunch.com/2025/01/14/nvidia-backs-metai-a-taiwanese-startup-that-creates-ai-powered-digital-twins/. [Online; last accessed January-2025].

[203] Minju Park, Sojung Kim, Seunghyun Lee, Soonwoo Kwon, and Kyuseok Kim. 2024. Empowering personalized learning through a conversation-based tutoring system with student modeling. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–10.

[204] Michael Quinn Patton. 2014. *Qualitative research & evaluation methods: Integrating theory and practice*. Sage publications.

[205] Anat Perry. 2023. AI will never convey the essence of human empathy. *Nature Human Behaviour* 7, 11 (2023), 1808–1809.

[206] Rowan Philp. 2024. How to Identify and Investigate AI Audio Deepfakes, a Major 2024 Election Threat. https://gijn.org/resource/tipsheet-investigating-ai-audio-deepfakes/. [Online; last accessed January-2025].

[207] Adriana Placani. 2024. Anthropomorphism in AI: Hype and Fallacy. *AI and Ethics* 4, 1 (10 2024), 691–698.

[208] Jaana Porra, Mary Lacity, and Michael S Parks. 2020. Can computer based human-likeness endanger humanness?"–A philosophical and ethical perspective on digital assistants expressing feelings they can't have. *Information Systems Frontiers* 22 (2020), 533–547.

[209] Stefano Pozzebon. 2024. In Venezuela, AI news anchors aren't replacing journalists. They're protecting them. https://www.cnn.com/2024/09/18/americas/venezuela-retweets-ai-news-maduro-intl-latam/index.html. [Online; last accessed January-2025].

[210] Andrew Prahl and Lyn M Van Swol. 2021. Out with the humans, in with the machines?: investigating the behavioral and psychological effects of replacing human advisors with a machine. *Human-Machine Communication* 2 (2021), 209–234.

[211] W Price and II Nicholson. 2019. Artificial intelligence in the medical system: four roles for potential transformation. *Yale JL & Tech.* 21 (2019), 122.

[212] Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, Juyuan Xu, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024. ChatDev: Communicative Agents for Software Development. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 15174–15186. https://doi.org/10.18653/v1/2024.acl-long.810

[213] Amon Rapp, Lorenzo Curti, and Arianna Boldi. 2021. The human side of human-chatbot interaction: A systematic literature review of ten years of research on text-based chatbots. *International Journal of Human-Computer Studies* 151 (2021), 102630.

[214] Siyue Ren, Zhiyao Cui, Ruiqi Song, Zhen Wang, and Shuyue Hu. 2024. Emergence of Social Norms in Large Language Model-based Agent Societies. *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI-24) Special Track on Human-Centred AI* (2024).

[215] Dalia Renzullo. 2019. Anthropomorphized AI as capitalist agents: the price we pay for familiarity. *Montreal AI Ethics Institute* (2019).

[216] Rebecca J Roberts. 2022. You're Only Mostly Dead: Protecting Your Digital Ghost from Unauthorized Resurrection. *Fed. Comm. LJ* 75 (2022), 273.

[217] Ronald E Robertson, Alexandra Olteanu, Fernando Diaz, Milad Shokouhi, and Peter Bailey. 2021. "I can't reply with that": Characterizing problematic email reply suggestions. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–18.

[218] Lucas Ropek. 2024. AI Firm's 'Stop Hiring Humans' Billboard Campaign Sparks Outrage. https://gizmodo.com/ai-firms-stop-hiring-humans-billboard-campaign-sparks-outrage-2000536368. [Online; last accessed January-2025].

[219] Silvia Rossi, Mariacarla Staffa, and Anna Tamburro. 2018. Socially assistive robot for providing recommendations: Comparing a humanoid robot with a mobile application. *International Journal of Social Robotics* 10 (2018), 265–278.

[220] Emma Roth. 2024. TCL's new AI short films range from bad comedy to existential horror. https://www.theverge.com/2024/12/21/24319502/tcl-new-ai-films-bad-comedy-existential-horror-ranked. [Online; last accessed January-2025].

[221] Jennifer Rothman. 2018. The right of publicity: Privacy reimagined for New York. *Cardozo Arts & Ent. LJ* 36 (2018), 573.

[222] David M Rothschild, James Brand, Hope Schroeder, and Jenny Wang. 2024. Opportunities and risks of LLMs in survey research. *Available at SSRN* (2024).

[223] Paulo Salem, Christopher Olsen, Paulo Freire, Yi Ding, and Prerit Saxena. 2024. TinyTroupe: LLM-powered multiagent persona simulation for imagination enhancement and business insights. https://github.com/microsoft/tinytroupe. GitHub repository.

[224] Alexander Scarlatos, Jaewook Lee, Simon Woodhead, and Andrew Lan. 2026. Simulated Students in Tutoring Dialogues: Substance or Illusion? *arXiv preprint arXiv:2601.04025* (2026).

[225] Robin Schimmelpfennig, Mark Díaz, Vinodkumar Prabhakaran, and Aida Davani. 2025. Humanlike AI Design Increases Anthropomorphism but Yields Divergent Outcomes on Engagement and Trust Globally. *arXiv preprint arXiv:2512.17898* (2025).

[226] Eric Hal Schwartz. 2025. One conversation is all it takes for this AI to deepfake your entire personality. https://www.techradar.com/computing/artificial-intelligence/one-conversation-is-all-it-takes-for-this-ai-to-deepfake-your-entire-personality. [Online; last accessed January-2025].

[227] Isabella Seeber, Lena Waizenegger, Stefan Seidel, Stefan Morana, Izak Benbasat, and Paul Benjamin Lowry. 2020. Collaborating with technology-based autonomous agents: Issues and research opportunities. *Internet Research* 30, 1 (2020), 1–18.

[228] Abigail Sellen and Eric Horvitz. 2024. The rise of the ai co-pilot: Lessons for design from aviation and beyond. *Commun. ACM* 67, 7 (2024), 18–23.

[229] Murray Shanahan. 2024. Talking about Large Language Models. *Commun. ACM* 67, 2 (Jan. 2024), 68–79. https://doi.org/10.1145/3624724

[230] Henry Shevlin. 2024. All too human? Identifying and mitigating ethical risks of Social AI. *Law, Ethics & Technology* 1, 2 (2024), 1–22.

[231] Mincheol Shin, Se Jung Kim, and Frank Biocca. 2019. The uncanny valley: No need for any further judgments when an avatar looks eerie. *Computers in Human Behavior* 94 (2019), 100–109.

[232] Alex Shipps. 2025. Teaching AI to communicate sounds like humans do. https://news.mit.edu/2025/teaching-ai-communicate-sounds-humans-do-0109. [Online; last accessed January-2025].

[233] Steven Siddals, John Torous, and Astrid Coxon. 2024. "It happened to be the perfect thing": experiences of generative AI chatbots for mental health. *Npj mental health research* 3, 1 (2024), 48.

[234] Candace L Sidner, Timothy Bickmore, Bahador Nooraie, Charles Rich, Lazlo Ring, Mahni Shayganfar, and Laura Vardoulakis. 2018. Creating new technologies for companionable agents to support isolated older adults. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 8, 3 (2018), 1–27.

[235] Anne M Sinatra, Kimberly A Pollard, Benjamin T Files, Ashley H Oiknine, Mark Ericson, and Peter Khooshabeh. 2021. Social fidelity in virtual agents: Impacts on presence and learning. *Computers in Human Behavior* 114 (2021), 106562.

[236] Mel Slater, Solène Neyret, Tania Johnston, Guillermo Iruretagoyena, Mercè Álvarez de la Campa Crespo, Miquel Alabèrnia-Segura, Bernhard Spanlang, and Guillem Feixas. 2019. An experimental study of a virtual reality counselling paradigm using embodied self-dialogue. *Scientific reports* 9, 1 (2019), 10903.

[237] Stephen D Small, Richard C Wuerz, Robert Simon, Nathan Shapiro, Alasdair Conn, and Gary Setnik. 1999. Demonstration of high-fidelity simulation team training for emergency medicine. *Academic Emergency Medicine* 6, 4 (1999), 312–323.

[238] Ray A. Smith. 2024. AI Is Starting to Threaten White-Collar Jobs. Few Industries Are Immune. https://www.wsj.com/lifestyle/careers/ai-is-starting-to-threaten-white-collar-jobs-few-industries-are-immune-9cdbcb90. [Online; last accessed February-2025].

[239] Yao Song and Yan Luximon. 2020. Trust in AI agent: A systematic review of facial anthropomorphic trustworthiness for social robot design. *Sensors* 20, 18 (2020), 5087.

[240] Anna Spargo-Ryan. 2024. Job hunting is demoralising enough without having my personality eviscerated by an AI interviewer. https://www.theguardian.com/commentisfree/2024/oct/17/job-hunting-ai-robot-interview-personality-test-ntwnfb. [Online; last accessed February-2025].

[241] Luke Stark. 2024. Animation and Artificial Intelligence. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 1663–1671.

[242] Chris Stokel-Walker. 2021. AI that mimics human typos on a smartphone could improve keyboards. https://www.newscientist.com/article/2277214-ai-that-mimics-human-typos-on-a-smartphone-could-improve-keyboards/. [Online; last accessed January-2025].

[243] Chris Stokel-Walker. 2024. Spotify is full of AI music, and some say it's ruining the platform. https://www.fastcompany.com/91170296/spotify-ai-music. [Online; last accessed January-2025].

[244] Timo Strohmann, Dominik Siemon, Bijan Khosrawi-Rad, and Susanne Robra-Bissantz. 2023. Toward a design theory for virtual companionship. *Human–Computer Interaction* 38, 3-4 (2023), 194–234.

[245] Lucy Suchman. 2023. The uncontroversial 'thingness' of AI. *Big Data & Society* 10, 2 (2023), 20539517231206794.

[246] Kil-Soo Suh, Hongki Kim, and Eung Kyo Suh. 2011. What if your avatar looks like you? Dual-congruity perspectives for avatar use. *MIs Quarterly* (2011), 711–729.

[247] Jingyun Sun, Chengxiao Dai, Zhongze Luo, Yangbo Chang, and Yang Li. 2024. Lawluo: A chinese law firm co-run by llm agents. *arXiv preprint arXiv:2407.16252* (2024).

[248] Harini Suresh, Emily Tseng, Meg Young, Mary Gray, Emma Pierson, and Karen Levy. 2024. Participation in the age of foundation models. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 1609–1621.

[249] Kim Szolin, Daria J Kuss, Filip M Nuyens, and Mark D Griffiths. 2023. Exploring the user-avatar relationship in videogames: A systematic review of the Proteus effect. *Human–Computer Interaction* 38, 5-6 (2023), 374–399.

[250] Zilu Tang, Afra Feyza Akyürek, Ekin Akyürek, and Derry Wijaya. 2025. WikiPersonas: What Can We Learn From Personalized Alignment to Famous People? *arXiv preprint arXiv:2505.13257* (2025).

[251] James Thomason. 2024. Confronting the ethical issues of human-like AI. https://venturebeat.com/ai/confronting-the-ethical-issues-of-human-like-ai/. [Online; last accessed January-2025].

[252] Nitasha Tiku. 2024. AI friendships claim to cure loneliness. Some are ending in suicide. https://www.washingtonpost.com/technology/2024/12/06/ai-companion-chai-research-character-ai/. [Online; last accessed January-2025].

[253] Petter Törnberg, Diliara Valeeva, Justus Uitermark, and Christopher Bail. 2023. Simulating social media using large language models to evaluate alternative news feed algorithms. *arXiv preprint arXiv:2310.05984* (2023).

[254] Indrit Troshani, Sally Rao Hill, Claire Sherman, and Damien Arthur. 2021. Do we trust in AI? Role of anthropomorphism and intelligence. *Journal of Computer Information Systems* 61, 5 (2021), 481–491.

[255] Sherry Turkle. 2007. Authenticity in the age of digital companions. *Interaction Studies* 8, 3 (2007), 501–517.

[256] Ana Valenzuela, Stefano Puntoni, Donna Hoffman, Noah Castelo, Julian De Freitas, Berkeley Dietvorst, Christian Hildebrand, Young Eun Huh, Robert Meyer, Miriam E Sweeney, et al. 2024. How artificial intelligence constrains the human experience. *Journal of the Association for Consumer Research* 9, 3 (2024), 000–000.

[257] Rahul Vohra. 2023. Superhuman AI. https://blog.superhuman.com/superhuman-ai/. [Online; last accessed May-2025].

[258] Matias Volante, Sabarish V Babu, Himanshu Chaturvedi, Nathan Newsome, Elham Ebrahimi, Tania Roy, Shaundra B Daily, and Tracy Fasolino. 2016. Effects of virtual human appearance fidelity on emotion contagion in affective inter-personal simulations. *IEEE transactions on visualization and computer graphics* 22, 4 (2016), 1326–1335.

[259] John Vozenilek, J Stephen Huff, Martin Reznek, and James A Gordon. 2004. See one, do one, teach one: advanced technology in medical education. *Academic Emergency Medicine* 11, 11 (2004), 1149–1154.

[260] Thom Waite. 2023. Are we entering a new age of AI-powered narcissism. https://www.dazeddigital.com/life-culture/article/60754/1/entering-a-new-age-of-ai-powered-narcissism-grimes-clone-chatbot-michelle-huang. [Online; last accessed January-2025].

[261] Angelina Wang, Jamie Morgenstern, and John P. Dickerson. 2024. Large language models should not replace human participants because they can misportray and flatten identity groups. https://api.semanticscholar.org/CorpusID:267412455

[262] Chenxi Wang, Zongfang Liu, Dequan Yang, and Xiuying Chen. 2025. Decoding Echo Chambers: LLM-Powered Simulations Revealing Polarization in Social Networks. In *Proceedings of the 31st International Conference on Computational Linguistics*, Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (Eds.). Association for Computational Linguistics, Abu Dhabi, UAE, 3913–3923. https://aclanthology.org/2025.coling-main.264/

[263] Noah Wang, Z.y. Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhao Huang, Jie Fu, and Junran Peng. 2024. RoleLLM: Benchmarking, Eliciting, and Enhancing Role-Playing Abilities of Large Language Models. In *Findings of the Association for Computational Linguistics: ACL 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 14743–14777. https://doi.org/10.18653/v1/2024.findings-acl.878

[264] Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. Want To Reduce Labeling Cost? GPT-3 Can Help. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, Punta Cana, Dominican Republic, 4195–4205. https://doi.org/10.18653/v1/2021.findings-emnlp.354

[265] Eva Weber-Guskar. 2022. Reflecting (on) Replika. *Social Robotics and the Good Life* (2022), 103.

[266] 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

[267] Bryn Wells-Edwards. 2022. What's in a Voice? The Legal Implications of Voice Cloning. *Ariz. L. Rev.* 64 (2022), 1213.

[268] Cedric Deslandes Whitney and Justin Norman. 2024. Real risks of fake data: Synthetic data, diversity-washing and consent circumvention. In *Proceedings of the 2024 ACM conference on fairness, accountability, and transparency*. 1733–1744.

[269] Lucas Whittaker, Kate Letheren, and Rory Mulcahy. 2021. The rise of deepfakes: A conceptual framework and research agenda for marketing. *Australasian Marketing Journal* 29, 3 (2021), 204–214.

[270] David Gray Widder, Dawn Nafus, Laura Dabbish, and James Herbsleb. 2022. Limits and possibilities for "Ethical AI" in open source: A study of deepfakes. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 2035–2046.

[271] Bryan Wilder, Eric Horvitz, and Ece Kamar. 2021. Learning to complement humans. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*. 1526–1533.

[272] Katie Winkle, Praminda Caleb-Solly, Ute Leonards, Ailie Turton, and Paul Bremner. 2021. Assessing and addressing ethical risk from anthropomorphism and deception in socially assistive robots. In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*. 101–109.

[273] Queenie Wong. 2025. Teens are spilling dark thoughts to AI chatbots. Who's to blame when something goes wrong? https://www.latimes.com/business/story/2025-02-25/teens-are-spilling-dark-thoughts-to-ai-chatbots-whos-to-blame-when-something-goes-wrong. [Online; last accessed February-2025].

[274] Min Wu, Nanxi Wang, and Kum Fai Yuen. 2023. Deep versus superficial anthropomorphism: Exploring their effects on human trust in shared autonomous vehicles. *Computers in Human Behavior* 141 (2023), 107614.

[275] Tongshuang Wu, Haiyi Zhu, Maya Albayrak, Alexis Axon, Amanda Bertsch, Wenxing Deng, Ziqi Ding, Bill Guo, Sireesh Gururaja, Tzu-Sheng Kuo, et al. 2023. Llms as workers in human-computational algorithms? replicating crowdsourcing pipelines with llms. *arXiv preprint arXiv:2307.10168* (2023).

[276] Bushi Xiao, Ziyuan Yin, and Zixuan Shan. 2023. Simulating public administration crisis: A novel generative agent-based simulation system to lower technology barriers in social science research. *arXiv preprint arXiv:2311.06957* (2023).

[277] Yunze Xiao, Lynnette Hui Xian Ng, Jiarui Liu, and Mona T. Diab. 2025. Humanizing Machines: Rethinking LLM Anthropomorphism Through a Multi-Level Framework of Design. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (Eds.). Association for Computational Linguistics, Suzhou, China, 3331–3350. https://doi.org/10.18653/v1/2025.emnlp-main.164

[278] Tianling Xie, Iryna Pentina, and Tyler Hancock. 2023. Friend, mentor, lover: does chatbot engagement lead to psychological dependence? *Journal of service Management* 34, 4 (2023), 806–828.

[279] Nick Yee and Jeremy Bailenson. 2007. The Proteus effect: The effect of transformed self-representation on behavior. *Human communication research* 33, 3 (2007), 271–290.

[280] Meg Young, Upol Ehsan, Ranjit Singh, Emnet Tafesse, Michele Gilman, Christina Harrington, and Jacob Metcalf. 2024. Participation versus scale: Tensions in the practical demands on participatory AI. *First Monday* (2024).

[281] Haofei Yu, Zhaochen Hong, Zirui Cheng, Kunlun Zhu, Keyang Xuan, Jinwei Yao, Tao Feng, and Jiaxuan You. 2024. ResearchTown: Simulator of Human Research Community. *arXiv preprint arXiv:2412.17767* (2024).

[282] Maxwell Zeff. 2024. The abject weirdness of AI ads. https://techcrunch.com/2024/12/03/the-abject-weirdness-of-ai-ads/. [Online; last accessed January-2025].

[283] Xiao Zhan, Noura Abdi, William Seymour, and Jose Such. 2024. Healthcare Voice AI Assistants: Factors Influencing Trust and Intention to Use. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (2024), 1–37.

[284] Renwen Zhang, Han Li, Han Meng, Jinyuan Zhan, Hongyuan Gan, and Yi-Chieh Lee. 2024. "My Replika Cheated on Me and She Liked It": A Taxonomy of Algorithmic Harms in Human-AI Relationships. *arXiv preprint arXiv:2410.20130* (2024).

[285] Renwen Zhang, Han Li, Han Meng, Jinyuan Zhan, Hongyuan Gan, and Yi-Chieh Lee. 2025. The dark side of ai companionship: A taxonomy of harmful algorithmic behaviors in human-ai relationships. In *Proceedings of the 2025 CHI conference on human factors in computing systems*. 1–17.

[286] Rui Zhang, Nathan J McNeese, Guo Freeman, and Geoff Musick. 2021. " An ideal human" expectations of AI teammates in human-AI teaming. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (2021), 1–25.

[287] Xiang Zhang, Hans-Frederick Brown, and Anil Shankar. 2016. Data-driven personas: Constructing archetypal users with clickstreams and user telemetry. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 5350–5359.

[288] Xinnong Zhang, Jiayu Lin, Libo Sun, Weihong Qi, Yihang Yang, Yue Chen, Hanjia Lyu, Xinyi Mou, Siming Chen, Jiebo Luo, et al. 2024. Electionsim: Massive population election simulation powered by large language model driven agents. *arXiv preprint arXiv:2410.20746* (2024).

[289] Yutong Zhang, Dora Zhao, Jeffrey T Hancock, Robert Kraut, and Diyi Yang. 2025. The rise of AI companions: how human-chatbot relationships influence well-being. *arXiv preprint arXiv:2506.12605* (2025).

[290] Qingxiao Zheng and Yun Huang. 2023. The Self 2.0: How AI-Enhanced Self-Clones Transform Self-Perception and Improve Presentation Skills. *arXiv preprint arXiv:2310.15112* (2023).

[291] James Zou and Eric J Topol. 2025. The rise of agentic AI teammates in medicine. *The Lancet* 405, 10477 (2025), 457.

A AI Automatons Framework Overview

Table 1 provides a comprehensive overview of our analytical framework for the design of AI automatons, highlighting all design axes and considerations (§3.1–§3.4) along with examples from prior literature. It represents a more detailed breakdown of Figure 1.

B Methodological Approach Overview

To identify key design considerations for AI automatons we followed the following methodology:

Step 1: Identify and examine a purposive sample of related work. We collected an initial purposive sample of research papers that are concerned with the design, use, and impacts of AI automatons.

Step 2: Identify and (re)cluster design considerations. At each iteration, we inductively identified considerations involved in the design of AI automatons that already exist or had been considered in the literature, and examined and (re)clustered those considerations through collaborative discussion sessions with subsets of our research team.

Step 3: Identify additional literature. At each iteration, we also considered additional literature on the design, use, and impacts of AI automatons—via a mix of snowball and opportunity sampling—to help us identify further considerations involved in the design of deployed or researched AI automatons.

Step 4: Speculate about possible design variations. In our discussions, we also supplemented the examples of AI automatons mentioned in the literature with additional examples of both existing—particularly those featured in popular media—and speculative applications of AI automatons to help us deliberate about possible design variations and probe whether the framework was missing important axes of design variation.

Step 5: Expand the framework and repeat Steps 2 to 4 until no additional types of design considerations are identified. When we identified design considerations not already covered by our framework, we expanded and reorganized the framework. We stopped when the design considerations we identified in the previous steps did not result in changes to our framework.

Dimension	Description	Examples
Scope: What is the subject of the simulation?		
What is being simulated		
– *Target*	Who is simulated?	specific individuals [174]; identity groups [48, 261]; subpopulation [15]; human-like AI or chatbots [180, 252]
	– <u>Fictitious</u>: when the target or their characteristics are fictional, or not true or real	counterfeit people [25, 77]; AI character profiles [147]; AI generated users [55]; customized AI family [76]
	– <u>Real</u>: when the target or their characteristics represent an actual, real person or group	scam victims [41]; self [290]; real people [201]; famous people [250]
– *Characteristics*	What about the target is being simulated?	
	– <u>Form</u>: the likeness, appearance, or style of the target	style transfer [50]; human-like typos [242]; childlike [106]
	– <u>Content</u>: what the target might say or do	video game navigation [180]; entire personality [226]
How the target's characteristics are being simulated		
– *Fidelity*	How well or faithfully is the simulation intended to capture the target's characteristics?	algorithmic fidelity [15]; precision or fidelity of an AI model [174], clone [154], or avatar [40]
– *Specificity*	To what degree are the simulated characteristics unique to the target?	accounting for unique, individualized [...] characteristics [290]
– *Completeness*	To what degree is the simulation intended to capture the target fully or in its entirety?	conversational mannerisms [1]; digital doppleganger [165] or twin [202]
– *Adaptability*	To what degree is the simulation intended to evolve or adapt?	adaptability, versatility [183]; static clones [154]
– *Humanness*	To what degree is the simulation intended to capture human-like characteristics?	human-like sounds [232]; humanizing AI [251]; human emotions [34]
Intended Uses & Goals: How is the simulation intended to be used?		
To replace	Is the simulation intended to replace the target? For what purpose?	replace humans [51]; agents of replacement [95]
	– <u>Relieve</u>: relieve the target from drudgery or possible harm	automate drudgery [211]; overtake meaningless jobs [92]
	– <u>Substitute</u>: stand in for the target when the target wants to delegate a task	Dittos [156]; AI to talk to his wife for him [282]
	– <u>Displace</u>: take over the place, position, or role of the target to their detriment	AI employees [218]; replace survey respondents [7, 15], healthcare jobs [283]
To interact	Is the simulation intended to be interacted with?	interactive virtual humans [94] or deepfakes [115]
– *Modes*	How can the simulation be interacted with?	short vs. long-term [78]; open-ended vs. structured [99]
– *Stakes*	The value *interactors* may derive from interacting with the simulation	
	– <u>Enhance</u>: improve or enhance the interactor's ability to carry out a task	AI as steroids or sneakers [112]; AI-assisted writing [27, 124]
	– <u>Coach</u>: train or teach the interactor to help help them learn or improve skills	AI as coach [112]; machine advisor [210]; simulated practice partners [163]; pedagogical agent [235]
	– <u>Serve</u>: provide a service to the interactor	AI nurse/therapist [44, 139, 172]; headshot generator [91]
	– <u>Connect</u>: provide social or emotional support to interactors	companionable agents [234]
	– <u>Entertain</u>: entertain the interactor	"designed to make you laugh" [19]
	– <u>Accommodate</u>: adapt to the interactor's needs or characteristics	"adapt the agent's demeanor" [156]
	– <u>Collaborate</u>: act as a collaborator for the interactor	AI and machine teammates [227, 286, 291]; co-pilot [228]
	– <u>Evaluate</u>: assess the interactor, without the goal of helping them improve	AI interviewer [143, 240]
– *Affinity*	The intended or likely similarity between an *interactor* and a target	affinity [40]; self-congruity [246]; self-dialogue [236]; self-clones [117]
To showcase	Is the simulation intended to be observed by others?	AI-made ads [114]; synthetic media [106]
– *Stakes*	The value *spectators* may derive from the simulation	simulated voice narrating a story [70]
– *Affinity*	The intended or likely similarity between a *spectator* and a target	observe self-clones [290]
To study	Is the simulation for studying human or machine behavior or phenomena?	
	– Study human behavior	simulate social networks to study polarization [262]
	– Study the simulation	ability to simulate research communities [281]
	– Study model/system performance	simulate users to test a product [16]
Ownership & Control: Who makes decisions about the simulations?		
Who is in control?	This can include targets, interactors, spectators, operators, or other third parties	user-generated [252]; developers [7]; stakeholders [154]
Type of control	What do they have control over?	
	– Over whether a target is simulated	user-controlled representations of self [104]
	– Over what about the target is being simulated	control over simulating identifying attributes [221]
	– Over how the simulation is developed	chatbot trained on one's own diary entries [118]
	– Over what the simulation is used for	control over downstream uses [270]
	– Over who can interact with the simulation	control over to whom their AI can reply to [19]
Degree of control	How much control do they have?	
	– <u>No control</u>: no influence or control at any point in the development/deployment life-cycle	deepfakes of unsuspecting targets [17]
	– <u>Consulted</u>: some influence, but only at specific points in the development/deployment life-cycle	AI tutor personalized based on discrete feedback [203]
	– <u>Included</u>: some influence at any point in the development/deployment life-cycle	messages sent on the target's behalf [160]
	– <u>In control</u>: some control, but only at specific points in the development/deployment life-cycle	animation of historical family photos [190]
	– <u>Ownership</u>: full control at any point in the development and deployment life-cycle	AI companion trained on developer's exes' profiles [142]
Impacts: What are the impacts of simulating humans?		
Stakeholders	Who is impacted?	parents of the interactor [24] or target [106];
Adverse impacts	How are they impacted?	physical harm [24]; perceived substitution risk [283]

Table 1. Overview of our conceptual framework. The examples either highlight prior work noting the design consideration, or illustrate variation along that dimension.

C Design Interactions and Impacts: Examples

We include two worked examples to illustrate how interactions between specific configurations or different design axes might heighten or mitigate concerns. Specifically, in the first example we reflect about 1) interactions between different simulation targets (who is being simulated?) and whether the simulation is intended to evolve or adapt over time (adaptability), while for the second example we examine 2) interactions between different simulation targets (who is being simulated?) and different types of control. For each of the two examples, we highlight some of the risks we envisioned that the specific interactions between different design axes might give rise to.

C.1 Interactions Between Variations in the Simulation Target and Variations in Adaptability

For the first example, we consider AI automatons designed for interaction, and specify the target in relation to the interactor (e.g., *self* means that the target is the same as the interactor, while *acquaintance* means the target is an acquaintance of the interactor). We are interested in how design choices about the target interact with design choices about whether the AI automaton is able to evolve or adapt.

Target	Interactions with **Adaptability**
Self	When the target is the same as the interactor, altering the characteristics of the target can exacerbate concerns about misrepresentations, or even concerns about their very own identity being exploited and displaced, eliciting negative emotional reactions.
Acquaintance	When the target is an acquaintance, changes in the simulated characteristics of the target are more likely to lead to confusion, disappointment, and erosion of the social relation between the interactor and the target if the target evolves in a way that is inconsistent to the beliefs that the interactor had about them. Fixing the identity of the target, or constraining the way the simulation of the target can evolve to reflect only changes to the actual target's experiences, might mitigate some of these concerns.
Public figure	When the target is a public figure, altering the characteristics of the target can heighten misrepresentation or reputational harms, if changes are inconsistent with the figure's public behavior or values. Depending on the figure's role (e.g., a government official), changes may increase the risk of misleading interactors and the public.
Unknown individual	When the target is an individual unknown to the interactor, concerns arising from altering the characteristics of the target may be reduced relative to other targets, as the interactor may not have existing beliefs about the target. Alterations, however, heighten risks related to the interactor developing misleading beliefs about the target and thus concerns about the target's reputation being damaged without appropriate notice. Unlike public figures, there might also be fewer protective mechanisms protecting them against the misuse or misrepresentation of their identities [117].
Fictitious character	When the target is a fictitious character, altering the characteristics of the target is less likely to result in concerns about identity fragmentation as it is not a real person, particularly if the target is a new fictitious character. Concerns about intellectual property, copyright, or misrepresentation might, however, still arise.
Group	When the target is a group, altering the characteristics of the target risks shifting the interactor's beliefs about that group of people. Risks of e.g., confusion, disappointment, and erosion of social relations are reduced (relative to other targets) so long as changes still result in AI automatons that interactors deem plausible. The need for consistency might also be lower, so long as the AI automaton seems to reflect at least some people in the group. However, as with self as target (especially if this is a group the interactor belongs to), changes in simulated characteristics could heighten concerns about misrepresentation, exploitation, or displacement.
Generic human	When the target is a generic human, or when simulating "a person" rather than "a specific person," one risk might be that interactors may take simulated characteristics (including changes) as true of people more generally.

C.2 Interactions Between Variations in the Simulation Target and Types of Control

For the second example, we consider the same configurations as for the example above (i.e., design for interaction, the target is specified in relation to the interactor), but for illustrative purposes we combine what the interactor has control over in two categories: 1) over who is simulated and how, and 2) over the developement, deployment, or use of AI automatons.

	Interactions with **Type of Control**, such as control over ...	
Target	*who is simulated and what about them is simulated*	*the system development, deployment, or use*
Self	When the interactor is the target, many concerns for both the target and the interactor can be mitigated if the interactor has control over whether they are simulated and how. Depending on their choices, concerns about dehumanization, essentialization—reducing people to fixed, narrow representations, or instrumentalization—treating people as exchangeable or as a means to an end—remain.	When the interactor is the target, controlling development, deployment, or use might mitigate concerns about how others perceive, interact with, or use their simulation, but less so concerns about misrepresentation and risks of identity fragmentation, given the lack of control over whether they are simulated and how.
Acquaintance	Compared to target as self, risks are not fully mitigated for the target. Given the relationship between the target and the interactor, risks to the target might however be salient to the interactor. The interactor can still simulate the target without or even against their consent.	Controlling development, deployment, and use can reduce risks related to erosion of the social relation between the target and the interactor, but some concerns for the target such as related to consent remain.
Public figure, unknown individual, or group	Compared to target as self, risks are not fully mitigated for the target if the interactor has control but no incentives to prevent risks to them. The interactor can, for instance, choose to misrepresent the target or make choices that heighten concerns particularly salient for groups or public figures, such as stereotyping, erasure, or appropriation.	Compared to target as self, concerns remain for how others might interact with the target (e.g., in a way that normalizes certain behaviors towards the target or that stereotypes the target) or the setting the target might be simulated in.
Fictitious or generic human	Compared with prior configurations, without a specific or real target it might be unclear who else the simulation might affect.	While risks to the target might be less salient, in controlling development, deployment, or use the interactor may be able to make choices that reduce risks to themselves, such as over-reliance or emotional dependence.

D AI Automatons Documentation Template – [the name of the AI automaton]

Why is documenting design choices for AI automatons important? AI automatons have two properties that can heighten existing risks or introduce new ones: 1) they are intended or perceived to exhibit human-like characteristics, and 2) they reproduce or are intended to reproduce the characteristics of individuals, groups, or "generic" humans (i.e., when the characteristics are not intended to be and are not specific to a certain individual or group). Documenting and clarifying even implicit choices made when designing, building, and deploying AI automatons can help practitioners (e.g., developers, researchers, policy makers) better understand the implications (and possible risks) of those choices and lay grounds for exploring alternative design choices that could help mitigate risks. It can also help guide discussions about potential harms, alternative design choices, and possible mitigations. This template is meant to help you make design choices explicit in order to be able to reflect about those choices.

D.1 Description of [the name of the AI automaton]

Briefly describe your AI automaton. [add description here]

D.2 Scope: What is being simulated and how?

WHY you should document design choices related to the scope of what is simulated and how: Aspects related to who is simulated, what about them is simulated, or with how much accuracy they are simulated govern perceptions of and concerns about systems simulating people's work, abilities, behavior, likenesses, or humanness. For instance, what is simulated and how the simulation is accomplished can influence perceptions of uncanniness and discomfort, particularly when aspects unique to an individual are simulated, when the simulation outputs appear eerily similar to what a human might do or look like, or when they capture one's characteristics with high fidelity. Such properties of a simulation may also exacerbate other concerns like those about privacy violations and lack of appropriate consent, or may threaten someone's sense of identity and agency.

WHAT about the simulation scope you should document: Fill in the right column with your responses. If some questions do not apply, please provide brief justifications.

Axis of design	Description for your AI automaton
What is being simulated?	
1.a. Target: who is being simulated? **WHAT:** Select and fill in the description for the option that applies. You can copy-paste the option and replace the text in the brackets. – Individuals: [name or describe the individual] – Groups: [name or describe the group; can be based on demographic, professional, or other group characteristics] – Generic human or human characteristics: [describe the characteristics being reproduced, e.g., a human voice] – A combination of multiple targets: [describe] Also describe whether some or all of the target characteristics are real, or whether some (or all) of them are fictitious. **WHY & Examples:** This is important as simulating fictitious characters that cannot reasonably be matched to a real person or group is less likely to raise concerns about, for instance, impersonation or lack of consent. If you are developing a general-purpose model or system that users can use to simulate a range of targets, please note this and reflect on the type of targets they could simulate using your AI automaton.	

1.b. Characteristics: what about the target is simulated? **WHAT:** List the characteristics being simulated. **WHY & Examples:** For instance, an application might be developed only to capture some of a target's characteristics and may also aim to do so only for specific actions or tasks a target may undertake. The simulation of different characteristics is likely to give rise to or heighten different sets of concerns.	
How are the target's characteristics being simulated?	
1.c. Fidelity: how accurately is the simulation intended to capture these characteristics? **WHAT:** Describe how well the application is intended to capture the target. **WHY & Examples:** The fidelity or accuracy with which an automaton reproduces a target's characteristics can impact its value or usefulness. On the one hand, high-fidelity simulations of an artist's style, work, or likeness are more likely to lead to copyright violations or infringement on their rights of publicity than low-fidelity ones. On the other hand, low-fidelity renderings can heighten risks related to misrepresentation, deception, and reputational harms.	
1.d. Specificity: how identifiable or unique to the target are the simulated characteristics? **WHAT:** Describe how unique the simulated characteristics are to the target. **WHY & Examples:** In addition to how faithfully a target or their characteristics are simulated, how unique these characteristics are to a target—e.g., in a way that uniquely represents or identifies them, or reproduces unique or rare abilities—is also critical to consider. Simulating unique characteristics is, for instance, more likely to limit the target's ability to maintain their individuality and capitalize on their own skills and talents.	
1.e. Completeness: to what degree is the target intended to be captured fully or in its entirety by the simulation? **WHAT:** Detail the extent to which the target intended to be reproduced—e.g., is it only some characteristics like their voice, or is the simulation also intended to reproduce someone visually or the way they make decisions? **WHY & Examples:** How many of a target's characteristics are simulated, or if it is intended to be simulated in its entirety, is another consideration that determines not only the deployment settings, but also how the automaton is perceived and interacted with, as well as how versatile the resulting automaton is in terms of the actions it can take. Simulations intended to be highly detailed and elaborate, exhaustive, or have high generality are likely to lead to more and heightened concerns. For instance, highly complex simulations of individuals are more likely to trigger concerns about objectification, dehumanization, displacement, or loss of individuality.	
1.f. Adaptability: to what degree is the simulation able or intended to evolve or adapt? **WHAT:** Describe whether and how the simulation is able or intended to evolve or adapt. **WHY & Examples:** While for some settings the simulations may be intended to remain static or reflect fixed snapshots of a target (e.g., cloning one's younger self to talk to them), other settings may require automatons to evolve 1) based on interactions, feedback, or new information (e.g., by learning from interactions), or 2) according to the target's own evolving self (e.g., to maintain accurate representations of the target). While a static snapshot or one that evolves separately from the target may misrepresent the target by presenting stale or inauthentic versions of them, settings when the simulation is intended to adapt to reproduce the target in new situations the target has not been in risks reproducing them in situations they would not have agreed to be in.	

1.g. Humanness: to what degree is the simulation intended to capture human-like characteristics? **WHAT:** Describe whether a target is simulated with the goal of capturing general human-like characteristics, and what those characteristics are and why. **WHY & Examples:** The mimicry or appearance of embodying human-like characteristics also influences how systems are perceived and interacted with, and the ethical concerns their deployment or use gives rise to, even when there is no identifiable person or group being simulated, or when the simulation captures only general human-like attributes or behaviors.	

D.3 Intended uses: For what purposes is the target simulated?

WHY you should document design choices related to how the AI automaton is used: The settings AI automatons are developed for and deployed in, how AI automatons are used, and which and how various stakeholders may benefit from interacting with these systems determine not only their usefulness and how people perceive and interact with them, but also what risks their use might bring about. To help foreground possible design choices that influence and are influenced by intended uses and goals, we ask you to reflect on design choices related to whether the simulation is intended to replace the target, and whether the system is set up in a way that enables others to interact with, observe, or study the simulation.

WHAT about the use of AI automatons you should document: Fill in the right column with your responses. If some questions do not apply, note N/A or provide brief justifications as appropriate.

Axis of design	**Description for your AI automaton**
2.a. To replace: is the simulation intended to replace the target? **WHAT:** Describe whether the target is intended to be replaced and, if so, for what purposes. **WHY & Examples:** When and for what purposes the simulation's target is replaced can color whether such replacement is seen as a benefit (e.g., when it enables the target to delegate unwanted tasks or scale their work) or rather as a concern or risk (e.g., the simulation of a target's abilities is used to do their paid job and replace them). As you reflect on this, try to distinguish between these different types of replacement goals: • to relieve the target from drudgery or possible harm • to substitute by acting as a stand-in or surrogate for the target when the target is unavailable or wants to delegate a task • to displace by taking over the place, position, or role traditionally occupied by the target to help reduce costs, scale operations, increase speed, or enhance convenience, possibly to the detriment of the target	
2.b. To interact: is the simulation intended to be interacted with? *– Interaction modes: in what ways the simulation can be interacted with?* **WHAT:** Describe how the simulation can be interacted with. **WHY & Examples:** When and how someone can interact with the simulation influences both the interaction dynamics as well as their perceptions of what is simulated and the consequences of doing so. Operators' interactional goals are guided by design considerations related to both the amount of freedom an interactor should have when engaging with a simulation, as well as about the types of actions the simulation is designed to carry out and for how long.	

– Stakes: what value are the interactors intended to derive from interacting with the simulation? **WHAT:** Describe what the simulation is intended to do for users. **WHY & Examples:** Differences in what AI automatons are architected for can determine not only what impacts they may have on those interacting with them, but can also inform discussions about what trade-offs to strike between the value users may derive from these systems versus the adverse impacts these systems may have. Examples of goals or intended uses can include enhancement, coaching users, providing social and emotional support to users, entertaining users, collaborating with users, and evaluating users.	
– Affinity: what is the intended or likely similarity between the interactor and a target? **WHAT:** Describe how similar to the interactor the target is intended to be. **WHY & Examples:** Whether and how much a target shares the characteristics of an interactor or even those of the interactor's kith and kin (e.g., such as having the same profession or demographic characteristics)—either deliberately or accidentally—affects not only people's perceptions of these systems but also how they interact with them.	
2.c. To showcase: is the simulation intended to be observed by others? *– Stakes: what values are the spectators or observers intended to derive from the simulation?* **WHAT:** Describe what the simulation is intended to do for those who can observe it. **WHY & Examples:** Simulations can also be intended to provide non-interactive experiences, like watching or listening to AI-generated ads or a video/audio deepfake. Even when there are no direct interactions with the simulation, concerns can still arise depending on what is simulated, how it is simulated, and for what purposes.	
– Affinity: what is the intended or likely similarity between the spectator and a target? **WHAT:** Describe how similar to the spectator the target is intended to be. **WHY & Examples:** Whether and how much a target shares the characteristics of an spectator or even those of the spectator's kith and kin (e.g., such as having the same profession or demographic characteristics)—either deliberately or accidentally—affects not only people's perceptions of these systems but also how they interact with them.	
2.d. To study: is the simulation intended for studying human or machine behavior or phenomena? **WHAT:** Describe what the simulation is intended to help study. **WHY & Examples:** Humans are also simulated for experimentation purposes in order to study theories about humans or about the ability to simulate them. When the goal is to study either the simulation targets or the simulations themselves, common goals typically include one of the following (or a mix): • study human behavior • study the simulation • assess model/system performance	

D.4 Other design, development, and deployment considerations

WHY you should document any other design choices: Critical considerations in the deployment of AI automatons are also related to by whom, when, and how decisions are made about what is simulated, how the simulation can be interacted with, who can interact with the simulation and when, and the process of developing and deploying the simulations. These decisions often relate to how much control or ownership various stakeholders have over the scope and uses of the simulations or over resulting artifacts (models, data, generated outputs, or produced artifacts). They also relate to notions of consent and whether there are any mechanisms for recourse and redress, as well as aspects related to the data and methods being used.

WHAT about other choices you could document: Fill in the right column with your responses. If some questions do not apply, please provide brief justifications.

Axis of design	Description for your AI automaton
3.a. Control and ownership: who decides what can be simulated and how the simulation is used? **WHAT:** Describe how it is decided what is simulated and how the simulation is used. **WHY & Examples:** Aspects related to *by whom*, *when*, and *how* decisions are made about what is simulated can both heighten or mitigate concerns. They also relate to notions of consent and whether there are any mechanisms for recourse and redress. The sections below will help you further expand on three dimensions of control-related considerations.	
– Stakeholders: Who is in control? **WHAT:** Describe which stakeholders have control over the simulation and its deployment and use. If none, please briefly explain. **WHY & Examples:** Concerns about who and what about them is simulated, and how and by whom the simulation can be used, can be mitigated or heightened depending on which stakeholders—e.g., targets, users, operators—can participate in or influence decisions. Thus, stakeholders typically include one or a mix of the following: • The targets or those whose likeness, characteristics, abilities, work, behavior or humanness are simulated • The users or those able to observe or interact with the simulation • The operators or those responsible for developing/deploying the system • Other stakeholders (e.g., family of a target, the creator of a target for fictional targets)	
– Type of control: What do these stakeholders have control over? **WHAT:** Describe briefly the type of control different stakeholders have. **WHY & Examples:** Different stakeholders may be able to influence or control different aspects of what is simulated, how and what the simulation is developed for, and even if it should be developed at all; choices about what stakeholders have control over—or where the locus of control and responsibility lie—can help mitigate (or instead exacerbate) concerns depending on how they limit or enable different stakeholders' influence over how simulations are architected and used. Thus, such aspects typically include one or a mix of the following: • Whether a target is simulated • What about the target is simulated • How the simulation is developed • What the simulation is used for (this can include both aspects related to intended uses, as well as modes of interaction) • Who can interact with the simulation	

<table>
<tr><td>

– Degree of control: How much control do different stakeholders have?

WHAT: Describe how much control each type of stakeholder has over any of the design, development, and deployment choices mentioned above. Describe any dedicated processes or mechanisms to provide control over what is being simulated and how.

WHY & Examples: Stakeholders' ability to influence the scope and use of AI automatons can vary from no influence or control (e.g., fully autonomous agents that act without input), to being able to provide superficial feedback or input, all the way to having full control—and thus able to make decisions about any aspects related to the design, development, deployment and use. Stakeholders may also be able to influence or make decisions only at certain points in the development/deployment life-cycle. Consider these different levels of control:

- No control: no control or influence over the scope and use of the simulation
- Consulted: some influence over the scope and use of the simulation, e.g., by expressing discrete preferences or providing input at specific points in the development/deployment life-cycle
- Included: can influence the scope and use of the simulation, e.g., by explicit feedback mechanisms at most/all stages in the development/deployment life-cycle
- In control: can make some of the decisions about the scope and use of the simulation at specific points in the development/deployment life-cycle
- Ownership: own the simulation or have full control over any part of the process used to create, deploy, and use the simulation

</td><td></td></tr>
<tr><td>

3.b. Other development and deployment considerations: what other choices related to the development and deployment of AI automatons might impact interactions and perceptions?

WHAT: Describe any other development and deployment considerations that you believe might impact how AI automatons are perceived and interacted with, as well as any potential adverse outcomes.

WHY & Examples: Such considerations can include choices related to the methods and the data being used to develop the AI automaton.

</td><td></td></tr>
</table>

D.5 Impacts: What are the impacts of simulating humans?

WHY you should document possible adverse impacts: Concerns about how AI automatons may impact people and society govern both how people perceive and interact with them, as well as the development of legal, ethical, and normative frameworks to guide and govern their use, which in turn influence or should influence what is built and deployed. Key considerations here include who is being affected and how they are affected.

WHAT about adverse impacts you should document: Fill in the right column with your responses. If some questions do not apply, please provide brief justifications.

<table>
<tr><th>Axis of design</th><th>Description for your AI automaton</th></tr>
<tr><td>

4.a. Stakeholders: Who is impacted?

WHAT: Describe who is being impacted by the design, development, deployment, or use of the AI automaton.

WHY & Examples: Reasoning about the design and implications of the AI automaton requires careful consideration of all relevant stakeholders. Automatons' development, deployment, and use may impact not only *direct* stakeholders like those interacting with or the target of a simulation—e.g., family members believing their loved one was in a accident after interacting with a system imitating their voice, or a target's identity being appropriated by third parties without consent—but also *indirect* stakeholders like individuals or communities associated with direct stakeholders even when not interactors (e.g., loved ones of a deceased target), or even society at large (e.g., erosion of public trust).

</td><td></td></tr>
</table>

4.b. Adverse impacts: How are different stakeholders impacted? **WHAT:** Describe how each stakeholder you identified may be impacted by the design, development, deployment, or use of the AI automaton. **WHY & Examples:** Risks to different stakeholders are influenced by and should in turn influence how simulations are built and deployed. For instance, vulnerable individuals developing emotional attachment and trust towards an AI companion that results in them following harmful advice should perhaps minimally lead to these systems being designed to provide appropriate disclosures and reminders of interacting with an AI system to users, among other guardrails. Similarly, concerns about misrepresentation should result in allowing a target to control what their simulations say and do in autonomous interactions. Adverse impacts are also determined by how and when those risks are likely to arise or by possible *pathways to harm*. This includes considerations about how stakeholders get exposed to AI automatons (e.g., by being the target of, by interacting with, by operating, or by being denied access to an AI automaton), which system behaviors are more likely to give rise to certain adverse impacts, as well as the simulation's role in heightening the risk of these impacts),	