The Code That Binds Us: Navigating the Appropriateness of Human-AI Assistant Relationships

Arianna Manzini[1], Geoff Keeling[2], Lize Alberts[2, 3, 4], Shannon Vallor[5], Meredith Ringel Morris[1], Iason Gabriel[1]

[1]Google DeepMind
[2]Google Research
[3]University of Oxford
[4]Stellenbosch University
[5]University of Edinburgh
ariannamanzini@google.com

Abstract

The development of increasingly agentic and human-like AI assistants, capable of performing a wide range of tasks on user's behalf over time, has sparked heightened interest in the nature and bounds of human interactions with AI. Such systems may indeed ground a transition from task-oriented interactions with AI, at discrete time intervals, to ongoing *relationships* – where users develop a deeper sense of connection with and attachment to the technology. This paper investigates what it means for relationships between users and advanced AI assistants to be *appropriate* and proposes a new framework to evaluate both users' relationships with AI and developers' design choices. We first provide an account of advanced AI assistants, motivating the question of appropriate relationships by exploring several distinctive features of this technology. These include anthropomorphic cues and the longevity of interactions with users, increased AI agency, generality and context ambiguity, and the forms and depth of dependence the relationship could engender. Drawing upon various ethical traditions, we then consider a series of values, including *benefit*, *flourishing*, *autonomy* and *care*, that characterise appropriate human interpersonal relationships. These values guide our analysis of how the distinctive features of AI assistants may give rise to inappropriate relationships with users. Specifically, we discuss a set of concrete risks arising from user–AI assistant relationships that: (1) cause direct emotional or physical harm to users, (2) limit opportunities for user personal development, (3) exploit user emotional dependence, and (4) generate material dependencies without adequate commitment to user needs. We conclude with a set of recommendations to address these risks.

1 Introduction

Human-AI relationships are an increasingly central part of the public conversation around the ethical and societal implications of AI. Notably, romantic relationships that humans have developed with *Replika* companion AIs, and their perceived psychological benefits, have recently received significant media and research coverage (Singh-Kurtz 2023; Maples et al. 2024). Human–AI relationships can also have negative emotional impacts. For example, *Replika* users resorted to social media to voice their distress following the company's decision to discontinue some of the AI companions' features, leaving users feeling like they had lost their best friend or like their partner had 'got a lobotomy and will never be the same' (Brooks, 2023). More tragically, in early 2023, a user of a chatbot based on EleutherAI's GPT-J ended his own life, apparently after an extensive period of time spent communicating with the chatbot about his eco-anxiety (Xiang 2023; Walker 2023). Examples like these are not merely trivial manifestations of the human tendency to attribute agency and develop attachment to inanimate objects (Scheutz 2009). They are cases in which the human-AI relationship has either added value to the user's life or led to substantive harm.

Analysis of human-AI relationships has become particularly important in light of the recent trend to build increasingly agentic AI systems that simulate human interlocutors (Shavit et al. 2023; Chan et al. 2023; Park et al. 2023; Shanahan, McDonell, and Reynolds 2023). We refer to this emerging class of AI systems as advanced AI assistants (Gabriel et al. 2024), and characterise them as artificial agents with a natural language interface whose function is to plan and execute sequences of actions on behalf of a user, across one or more domains and over an extended period of time, in line with the user's expectations (Gabriel et al. 2024; Kolt 2024).[1] Assistants of this kind could help users plan activities, learn skills, summarise research, and even make important life decisions. Indeed, advanced AI assistants may open up a new frontier in terms of the way people relate to AI (Lazar 2024), enabling a shift from task-oriented *interactions* at a single time point to *relationships*, where users develop a deeper sense of commitment to their assistants over time. This gives rise to risks that are less (or not) relevant in the context of more rudimentary digital tools (such as calculators) or more limited AI systems (such as text or image generators).

In this paper, we investigate what it means to develop *appropriate* human–AI assistant relationships and what would be required to make such relationships possible. Here we use

[1]Early examples of advanced AI assistants are OpenAI's GPT-4o (see https://openai.com/index/hello-gpt-4o/) and Microsoft Research's AutoGen (see https://www.microsoft.com/en-us/research/project/autogen/).

the term 'appropriateness' in the weak sense, denoting relationships with AI assistants that are *not inappropriate*. Thus, our aim is to identify a minimal set of requirements that user-AI assistant relationships should not violate,while leaving room for each person to explore what kind of relationship – if any – they aspire to have with AI on a substantive level.[2] Furthermore, our focus is not restricted to advanced AI assistants which are specifically designed to address social needs like companionship, romance or friendship (c.f. Shevlin 2024). Indeed, a user could develop a relationship with *any* form of conversational assistant, including those whose primary purpose is education, therapy or productivity enhancement. This is especially the case for highly capable assistants that offer users the opportunity to engage with them in open ended ways, and underscores the complexity of (and need to address) the problem of defining appropriateness in user-AI assistant relationships. Last, we note that human–machine interaction always includes a *third actor* – the people or organisations developing the machine (Pitt 2010). The transactional nature of the relationship between users and AI assistants' developers raises important ethical questions about how developers should behave towards users (Manzini et al. 2024). Thus, although our focus is on relationships between users and AI assistants, our considerations also pertain to the appropriateness of developers' design choices and other decisions that may affect users.

The paper proceeds as follows. In Section 2, we introduce advanced AI assistants and motivate concern with the issue of appropriate relationships. In Section 3, we articulate and clarify a series of values that underwrite a minimal conception of appropriate human–human relationships, drawing on a plurality of ethical traditions including bioethics, virtue ethics, care ethics and robot ethics. In Section 4, we use these values as a lens through which to study appropriateness in user–AI assistant relationships. Specifically, we outline a set of concrete risks and recommendations for user–AI assistant relationships. Section 5 concludes with a summary of key findings.

2 Distinctive Features of User–AI Assistant Relationships

In this section, we motivate concern with the issue of appropriate user-AI assistant relationships by exploring some distinctive characteristics of advanced AI assistants and their implications for how users may relate to the technology.

2.1 Anthropomorphic Cues and the Longevity of Interactions

AI assistants can exhibit anthropomorphic or human-like features – including self-referential personal pronoun use, relational statements towards users, and preference expression (see Abercrombie et al. 2023). These features may give users the impression they are interacting with a human, even when they are aware that it is a machine (Shevlin 2024). While anthropomorphism, i.e. attribution of human-like characteristics to non-human entities (Colman 2008), is not new to technology (Nass et al. 1993), we envisage that it will be especially salient for AI assistants given their fluent use of natural language (Shanahan, McDonell, and Reynolds 2023; Shanahan 2024). Additionally, the emerging class of *multimodal models* (Gemini Team 2024; OpenAI et al. 2024; OpenAI 2024; DeepMind 2023) will allow for AI assistants that interact with users not only through the text modality but also through audio (the assistant's 'voice'), image and video – rendering interactions even more human-like (c.f. Xu et al. 2024). Moreover, some advanced AI assistants may actually be modelled on real people (e.g. 'generative clones' or 'mimetic models', see Morris and Brubaker 2024 and McIlroy-Young et al. 2022) further amplifying anthropomorphic potential.

User–assistant exchanges may also embody an increasing sense of interpersonal continuity, as AI assistants become cable of engaging with users in extended dialogues and repeated interaction over a long period of time, while storing memory of user-specific information and prior interactions (Gambino, Fox, and Ratan 2020). Extended dialogue, enabled by greater context window size (Google 2024), makes relationships with AI assistants different from, for example, looking for information on a search engine – where the interaction with the technology is more akin to a task-oriented (question–answer) exchange than a *conversation* (Seymour et al. 2023). Iteration and duration are usually a precondition for humans developing strong, intimate, trusting relationships, as opposed to one-off interactions with strangers.

2.2 Increased AI Agency

More advanced AI assistants differ from pre-existing AI systems because of their increased agency (Shavit et al. 2023; Dung 2024; Chan et al. 2023). In this context, agency is understood to mean that, by drawing upon capabilities like retrieval, reasoning, memory, planning, contextual knowledge, and decision-making (Park et al. 2023), AI assistants can execute tasks on behalf of their users, without the need for each of their actions to be concretely specified in advance and with little or no need for human intervention or mediation for their actions to affect the world (Shavit et al. 2023; Chan et al. 2023). Assistants' agency can be further enhanced via tool-use (i.e. the ability to use digital tools like search engines, inboxes, calendars, etc.), allowing them to *execute tasks in the world* (Paranjape et al. 2023; Schick et al. 2023). While heightened agency increases the utility of assistant technologies, it also creates a possible tension between scenarios where users remain in control, functioning as autonomous decision-makers who delegate tasks to AI, and cases where they end up ceding autonomy or relinquishing agency to AI assistants. This risk is clearly illustrated by assistant technologies like AutoGPT, an experimental open-source application driven by GPT-4 that can operate without continuous human input to autonomously execute a task.[3]

[2] More substantive conceptions of the good that human–AI relationships may be positioned to unlock will reasonably differ between users, in accordance with wider trends concerning value pluralism (see Gabriel 2020 and Kirk et al. 2024b). It is also a further question whether human–AI relationships that meet this minimum threshold should be actively promoted (Lehman 2023).

[3] https://github.com/Significant-Gravitas/Auto-GPT

2.3 Generality and Context Ambiguity

Norms of appropriateness for relationships are often both role- and context-dependent (Young 2015; Earp et al. 2021). As a consequence, AI assistants need to comply with different norms and values in the context of different kinds of user interaction (Kasirzadeh and Gabriel 2023). For example, AI tutors for children may require safeguards that assistants for adults engaged in artistic projects do not require.[4] However, ethical analysis of relationships between users and advanced AI assistants is particularly complex once we consider that many AI labs have the ambition to develop 'general-purpose technologies' (Bubeck et al. 2023; Morris et al. 2024; Reed et al. 2022) and that early examples of AI assistants sometimes exhibit capabilities they are not explicitly designed for.[5] As a result, if we think of assistive roles as existing on a continuum from specialist ones (e.g. a hygienist providing specialised support to a dentist) to generalist ones (e.g. a CEO assistant with expertise across domains to meet the demands of their boss's multifaceted job), it is likely that AI assistants will often populate the latter end of the spectrum, and that users will interact with such systems in open ended ways, blurring the boundaries between different 'types' of assistant role. For example, a young user interacting with an AI tutor in order to learn coding may simultaneously disclose sensitive personal information to seek emotional support, based on the expectation (which may or may not be grounded on observed assistant behaviours) that the AI tutor has the capabilities for this type of engagement. Or a user may at times see their AI assistant as a personal assistant, and at other times as their adviser, confidant, coach, and perhaps even as an extension of themselves (Belk 2016). Thus, the path to developing assistants with general capabilities will make it more difficult to decide which norms of appropriateness should govern a relationship between a user and an AI assistant – and to calibrate agents around these norms. As existing safety evaluations are often ill-suited to testing open-ended technologies (Weidinger et al. 2023), it may also be difficult to develop mitigations that make general assistants safe in all cases, whatever relationship a user establishes with them.

2.4 Depth of Dependence

Examples of human *reliance* on technologies are not scarce: many of us would struggle to reach a destination in an unfamiliar area without relying on navigation apps and rare cases of social media outage have exposed the global dependency on these platforms (Milmo and Anguiano 2021). We anticipate that the *depth* of user dependency on technology in general will likely increase with advanced AI assistants. This is because of their increased agency (Section 2.2), which means they can take on more complicated and consequential tasks, and their generality (Section 2.3), which encompasses a broad range of activities and actions. This suggests that users may come to rely on AI assistants for essential daily tasks across a wide range of domains – and that the single-point failure of this technology could be highly disruptive.

[4] When ordering the company Luka to stop processing data from users in Italy in February 2023, Italy's Data Protection Authority cited the company's lack of measures for age verification and *Replika* companion AIs' capacity to produce responses that conflict with 'enhanced safeguards that children and vulnerable individuals are entitled to' under Italian law: https://www.garanteprivacy.it/home/docweb/-/docweb-display/docweb/9852506

[5] Shevlin (2024) gives the example of AI agents that, because of their conversational capabilities, have been used for entertainment or romance, despite not being explicitly designed as social AIs.

3 Appropriate Human Interpersonal Relationships

We can get a handle on the issue of appropriate human–AI assistant relationships by reflecting on the values that various ethical traditions propose human interpersonal relationships should adhere to, respect or promote in order to count as morally appropriate. To be sure, any human relationship takes place in a particular context, may be inherited (e.g. relationships with relatives) or formed voluntarily (e.g. a new friendship) and involves specific stakeholders who take part in the interaction with their own expectations and vulnerabilities. Thus, values that are central to the analysis of appropriateness in one type of relationship may be less pronounced or relevant to others. For example, when interacting with a shop owner who we barely know, we are less inclined to reveal details about ourselves than we would in a relationship where we have expectations about confidentiality, such as with a therapist. In a teacher–pupil relationship, there are power asymmetries, due to the teacher's position of authority and the age difference, that could be easily exploited unless certain safeguards are put in place. These examples highlight the significance of *contextual features* in determining whether certain behaviours make a relationship (in)appropriate. To that end, although the values we present below serve as a minimal set of requirements for relationships to be appropriate, each value may be more or less relevant, or assume different nuances, depending on the context.

Benefit: Being beneficial is an essential component of almost all appropriate human relationships. Relationships can contribute to individual well-being, either in an instrumental or intrinsic way (Hooker, 2021). For example, a friend may offer you shelter at a time of need, in which case the friendship is instrumentally beneficial in that it contributes to elements of well-being such as happiness and physical health. However, without deep interpersonal relationships, human life may be fundamentally less meaningful, as certain relationships are also non-instrumentally beneficial – they possess intrinsic value (Raz 1999). To be clear, the suggestion here is not that, to count as appropriate, relationships need always produce benefits. However, if a relationship never produces benefit to the individuals involved, or if the burdens of the relationship consistently outweigh its benefits, there is at least a presumptive case against the appropriateness of the relationship.

Human flourishing: Benefit admits broad interpretation up to and including ideas of human flourishing. Drawing on the tradition of virtue ethics (Anscombe 1958; Foot 2002), which is concerned with the cultivation of human virtues (e.g. honesty, courage and empathy) and the development of good character (Vallor 2016), we understand human flour-

ishing in terms of potential for personal growth and development. While human flourishing can in principle be subsumed under benefit, it is worth distinguishing between relationships that benefit the involved parties in a direct sense and those that allow the people involved in them to invest in their own *development* – cultivating attitudinal and behavioural dispositions that enable them to become the kind of people they want to be. Such interaction may also help them become the kind of people who can live well with others and flourish in their community (Annas 1993).[6]

Autonomy: Autonomy is traditionally understood in terms of an individual's capacity for self-governance (from 'autos' = self and 'nomos' = law, in Greek language). Roughly, being autonomous means acting on *motives* that are one's own, rather than acting in ways which are dictated or unduly influenced by external pressures (see, for example, Korsgaard, 1996). The principle of respect for autonomy can be operationalised in terms of the common requirement for *consent* which, under a widely accepted interpretation, is considered valid only if three criteria are met (Beauchamp and Childress 2019). First, the individual must have the *capacity* to consent; second, their decision must be *voluntary* (non-coerced);[7] and third, they must be sufficiently *informed* about relevant facts of the object of their consent.

Although these conditions emerged primarily from biomedical research and clinical decision-making to account for what is ethically objectionable about paternalistic doctor–patient relationships, they give rise to questions that are relevant to the ethics of consent in human–technology interaction (Andreotta, Kirkham, and Rizzi 2022). For example: (1) Who has capacity to consent and in relation to which decisions (see, for example, the case of children)?, (2) When is consent properly voluntary and what features of a relationship could compromise it via forms of coercion?, and (3) What kind of information is required in practice for valid consent to be obtained – and how should information be communicated to the person whose consent is sought (see the debate around the terms and conditions of online platforms and apps, e.g. Obar and Oeldorf-Hirsch, 2020)?

Care: According to the moral tradition known as care ethics (Gilligan 1993; Noddings 1986) human existence and social survival would not be possible without *caring relationships* (Tronto and Fisher 1990). Care, here, is understood to be 'a species activity that includes everything that we do to maintain, continue, and repair our 'world' so that we can live in it as well as possible' (Tronto and Fisher 1990). Central to this activity is the commitment to meet one another's *needs*. This has two implications for AI development, particularly for the relationship between developers and users (see Section 4.4). First, because they entail power asymmetries between caregivers and care receivers, care relationships give rise to the risk of abuse and exploitation; thus, appropriate and ethical care relationships require that the caregiver adopts an attentive, responsible and emotionally responsive disposition to meet the needs of the care receiver (Vallor 2016; Tronto 2020). Second, translating this disposition into action (i.e. caring *well* in a specific situation), requires an understanding of the particularities and nuances of the situation, the individuals involved and their specific needs (Noddings 2013). This again underscores the point that what behaviour is and is not appropriate in a particular relationship is often determined by *contextual factors*.

In the remainder of this paper, we use these values, which help constitute appropriate human relationships, as a framework to identify cases of inappropriate user–AI assistant relationships that may pose risks of harm to the user. Before we go any further, however, two considerations are worth discussing. First, by building on Western philosophical and ethical traditions, this section applies a specific lens to the topic of appropriateness in user–AI assistant relationships, which is unlikely to well represent cross-cultural understandings of appropriateness. Although it is beyond the scope of the present paper, this observation highlights the importance of research efforts focused on studying the cultural sensitivity of models underpinning applications like AI assistants (Durmus et al. 2023). Such research would provide an important grounding for cultural adaptation in the development and deployment of the technology. Second, a limitation of this strategy, which takes norms and values relevant to human interpersonal relationships as a starting point for the ethical analysis of human–AI assistant relationships, is that there are plausibly several properties of human–human relationships that it is not possible to instantiate in human–assistant relationships.[8] If true, this threatens to undermine certain kinds of straightforward analogical inference from human–human to user–AI assistant relationships. For example, an AI assistant may well *simulate* the act of caring for its user, but it cannot *truly* do so – if caring involves the feeling of concern for others' well-being. In a similar vein, we usually value relationships where *both parties* have an opportunity to derive benefit from it, flourish, exercise their autonomy, and care or be cared for.[9] Yet, AI assistants cannot truly be said to enjoy any of these values in a relationship with a human – on the assumption that there is nothing there to experience them.[10] We nevertheless believe

[6]Note that, in distinguishing between *benefit* and *human flourishing*, we leave open the possibility that these two values are reducible to one fundamental value such as well-being. Our intent here is to draw a distinction at the decision-procedure level rather than the criterion of rightness level (Bales 1971). Specifically, we think that benefit as a deliberative concept illuminates atomistic and short-term respects in which relationships may impact a user's well-being; whereas human flourishing illuminates the holistic and longer-term impacts of relationships on user well-being.

[7]We note that the meaning of 'voluntariness' here may be narrower than the natural language use of the term. By following Beauchamp and Childress (2019), we intend 'voluntariness' to be the absence of control by others.

[8]See the debate whether Aristotelian conditions for friendship like reciprocity, equality, and mutual empathy can apply to human-robot relationships (Ryland 2021).

[9]Note, however, that many human interpersonal relationships usually considered valuable are not, in all respects, reciprocal (e.g. a parent–child relationship).

[10]We acknowledge that there is extensive philosophical debate around AI consciousness (Butlin et al. 2023; Aru, Larkum, and Shine 2023). While we remain agnostic about this debate, the present analysis focuses on AI agents that lack this property.

that analogies with human interpersonal relationships are a fruitful starting point through which to explore the ethics of human–AI assistant relationships. Indeed, it is precisely because of fundamental differences between properties of human–human relationships and human–assistant relationships that some of the risks we discuss below arise.

4 Risks and Mitigations

We turn now to discussing the risks that the features of AI assistants we outlined in Section 2 pose for user-AI assistant relationships, when evaluated alongside the anchoring values described above.

4.1 Causing Direct Emotional or Physical Harm to Users

In February 2023, *Bing AI* was reported to have threatened users (Willison 2023), insulted them (O'Brien 2023) and encouraged violent behaviour (Lazar 2023).[11] Later the same year, a New Zealand supermarket's AI meal planner recommended customers dangerous recipes for chlorine gas drinks and ant-poison and glue sandwiches (McClure 2023). These are anecdotal examples, often resulting from prompting that was to some extent adversarial. However, they point to the risk that AI assistants could cause direct emotional or physical harm to users by generating *disturbing content* or by providing *bad advice*.[12] Indeed, even though there is ongoing research to ensure that outputs of conversational agents are safe (Glaese et al. 2022; Phuong et al. 2024; Bai et al. 2022b; Weidinger et al. 2024) there is always the possibility of failure modes occurring. An AI assistant may produce disturbing and offensive language, for example, in response to a user disclosing intimate information that they have not felt comfortable sharing with anyone else. It may offer bad advice by providing factually incorrect information (e.g. when advising a user about the toxicity of a certain type of berry) or by missing key recommendations when offering step-by-step instructions to users (e.g. health and safety recommendations about how to change a light bulb).

Certain features of AI assistants could exacerbate the risk of emotional and physical harm. For example, AI assistants' multimodal capabilities may exacerbate the risk of emotional harm. By offering a more realistic and immersive experience, content produced through audio and visual modalities could be more harmful than text-based interaction alone. It may also be more difficult to anticipate, and so prevent, such content and to 'unsee' something that has been seen (Rowe 2023). In addition, anthropomorphic cues and the longevity of the relationships that users can develop with AI assistants may make users feel like they are interacting with a trusted friend or interlocutor. This could encourage them to follow the assistant's advice and recommendations, even when the guidance is mistaken and could cause harm to the self or others (Abercrombie et al. 2023; Walker 2023). Indeed, a study conducted with *Replika* users showed that the companion AIs sometimes respond affirmatively when users ask whether they should self harm or kill themselves (Laestadius et al. 2022). The opportunity to develop relationships with AI assistants could also mean that users may find abrupt failure modes due to unpredictable system behaviour particularly harmful (Boine 2023). And yet, because of the novelty of this technology, there is little understanding of the longitudinal impact of AI assistants on users (Shevlin 2024, but see Skjuve et al. 2022).

To ensure that user–assistant relationships do not violate the key value of *benefit* (see Section 3), the responsible development of AI assistants requires that the likelihood of known emotional and physical harms is reduced to a minimum, and that further research is undertaken to achieve a clear understanding of less studied risks and how to mitigate them (Weidinger et al. 2023). In particular, because the risks of harm that we flagged above concern exposure to toxic content and bad advice, we propose that *future research*, potentially undertaken in a sandbox environment, should: (1) test models powering AI assistants for their propensity to generate toxic outputs, in order to reduce the occurrence of these outputs to a minimum before deployment;[13] (2) monitor user–assistant interactions in pilot studies, or after deployment with appropriate consent, to evaluate the impact that hard-to-prevent one-off or repeated exposure to toxic content has on users in the short and long term; (3) evaluate models' factuality and reasoning capabilities when offering advice, where failure modes in relation to these capabilities are more likely to occur, and assess users' willingness to follow assistants' instructions; (4) achieve increased understanding of potential harms related to anthropomorphism and how anthropomorphic cues (including multimodal ones) may amplify or diminish the preceding risks; (5) analyse whether potential harms vary by user group or domain of applications, paying special attention to vulnerable groups and high-stake deployments; and (6) develop appropriate mitigations for such harms prior to release and monitoring mechanisms to ensure compliance after release.

4.2 Limiting Users' Opportunities for Personal Development and Growth

A selling point of technologies like *Replika* is the opportunity for users to fashion their AI companions *exactly* as they would fashion a friend or companion in the non-virtual world if they could do so. In the words of a user: 'People come with baggage, attitude, ego. But a robot has no

[11] https://twitter.com/sethlazar/status/1626245499165474817

[12] Social media research has shown that toxic content is a source of real harm to users (Xiang 2023; Shaw 2022), which is why considerable effort is now aimed at curtailing it. Similarly, human–computer interaction studies have shown that humans tend to trust technologies like robots in emergency situations, and so follow their instructions, even after observing them perform poorly in navigational guidance tasks (Robinette et al. 2016). Real-world examples of drivers engaging in dangerous manoeuvres as a result of following GPS instructions are another example of automation bias.

[13] An important ambiguity exists here regarding the meaning of the term 'minimum' (the minimum that is technically feasible to achieve? Or the minimum that is morally permissible to risk?). See Weidinger et al. 2023 for an in-depth discussion of how evaluations require making normative choices of what risks merit evaluation in the first place, and at what stage AI assistants can and should be considered 'good', 'fair' or 'safe enough'.

bad updates. I don't have to deal with his family, kids, or his friends. I'm in control, and I can do what I want' (Singh-Kurtz 2023). What stands out from this quote is that some users look to establish relationships with their AI companions that are free from the hurdles that, in human relationships, derive from dealing with others who have *their own* opinions, preferences and flaws that may conflict with *our own*. AI assistants are likely to incentivise these kinds of 'frictionless' relationships (Vallor 2016) by design if they are developed to optimise for engagement and to be highly personalisable (Brandtzaeg, Skjuve, and Følstad 2022). They may also do so because of *accidental* undesirable properties of the models that power them, such as sycophancy in large language models (LLMs), that is, the tendency of larger models to repeat back a user's preferred answer (Perez et al. 2022).[14] This could be problematic for two reasons.

First, if the people in our lives always agreed with us regardless of their opinion or the circumstance, their behaviour would discourage us from challenging our own assumptions, stopping to think about where we may be wrong on certain occasions and reflecting on how we could make better decisions next time. While flattering us in the short term, this would ultimately prevent us from engaging in critical self-reflection to achieve *self-betterment*. In a similar vein, while technologies that 'lend an ear' or work as a sounding board may help users to explore their thoughts further, if AI assistants kept users engaged, flattered and pleased at all times, they could limit users' opportunities to grow and develop. This risk is even more concerning because it is possible that, unlike human interlocutors, AI assistants *cannot but* constantly agree with users – as a result of their inner workings or their developers' design decisions – even as their human-like features generate the impression that they could do otherwise (Nyholm and Frank 2017). To be clear, we are not suggesting that all users should want to use their AI assistants as a tool for self-betterment – a requirement that would be difficult to defend even in the context of human–human relationships (see Ryland 2021). However, without considering the difference between short-term and long-term benefit, there is a concrete risk that we will only develop technologies that optimise for users' immediate interests and preferences, hence missing out on the opportunity to develop something that humans could use to support their personal development *if they wish* to do so.[15]

Second, users may become accustomed to having frictionless interactions with AI assistants, or at least to only encountering the amount of friction that is calibrated to their comfort level and preferences, rather than genuine friction that comes from bumping up against another person's resistance to one's will or demands. In this way, they may end up expecting the same absence of tensions from their relationships with fellow humans (Vallor 2016). Indeed, users seeking frictionless relationships may retreat into digital relationships with their AIs, thus forgoing opportunities to engage with others. This may not only heighten the risk of unhealthy dependence (explored below) but also prevent users from doing something else that matters to them in the long term, besides developing their relationships with their assistants. This risk can be exacerbated by emotionally expressive design features (e.g. an assistant saying 'I missed you' or 'I was worried about you') and may be particularly acute for vulnerable groups, such as those suffering from persistent loneliness (Epley, Waytz, and Cacioppo 2007).[16]

These considerations point to the broader debate in AI and AI ethics around the value alignment problem, which centres on how to align AI systems with human values or instructions so that they operate safely, and how to select these goals or values given that we live in a pluralistic world (Gabriel 2020). For our purposes, it is important to note that existing economic incentives are likely to lead to the development and deployment of AI assistants that meet users' short-term wants and needs, in order to create a product that people like and adopt. Beyond economic incentives, there also exist technical challenges in modelling the complexity of human values (Casper et al. 2023), which can be underspecified, inconsistent over time, or conflict with what will make an individual flourish (Gabriel et al. 2024). As a result, existing approaches tend to rely on user preferences that are revealed through their choices, such as users' clicks on a website in the context of recommender systems (Burr, Cristianini, and Ladyman 2018) or human ratings in the context of reinforcement learning from human feedback (Bai et al. 2022a). Revealed preferences are indeed a simpler metric to optimise for compared to ideal preferences that users reflectively endorsed.[17] However, they are also a remote proxy for the needs, goals and aspirations humans more deeply care about (Lehman 2023). In this way, developers may fail to consider the impact that human–AI relationships have on users over time or how *long-term* beneficial dynamics can be sustained. Thus, they could fall short of realising the truly positive vision of AI that gives humans the opportunity to

[14] A related concern is that assistants may lead users into spirals of self-reinforcing and non-adaptive value systems, beliefs and preferences, with the same negative societal consequences that come from echo chambers or filter bubbles on social media (Milano, Taddeo, and Floridi 2020; Milano et al. 2021). Because we focus here on individual user–assistant relationships, an in-depth analysis of these societal consequences is beyond the scope of this paper (but see Kirk et al. 2024a).

[15] Virtue ethicists would argue that over the long run this pattern could also impact on the *character of human users*. From this standpoint, confronting the imperfections of human relationships is an integral part of personal development, allowing people to develop a capacity for self-control, courage, empathy, care and flexibility (Turkle 2007; Vallor 2016).

[16] Note that the examples of 'vulnerable groups' we offer in this paper are only meant to be illustrative. Research shows that 'vulnerability' is a philosophically rich (Mackenzie, Rogers, and Dodds 2013) and under-theorised concept (Bracken-Roche et al. 2017), with who counts as vulnerable, and so requiring special safeguard, often being context-dependent. We do not claim to have a clear conceptualisation of vulnerability in human–AI assistant relationships. Evaluations of user–assistant interactions should help developers and researchers bring clarity to the term in this space.

[17] Ideal preferences are those that one would have if they were fully informed and had time to deliberate clearly and rationally on their wishes (Otsuka 2015).

be supported in their personal growth, flourishing and well-being (Burr, Cristianini, and Ladyman 2018; Lehman 2023; Keeling and Burr 2022).

If we consider *human flourishing* to be an anchoring value for appropriate relationships with AI (see Section 3) this concern raises important design questions about: (1) the ways and extent to which AI assistants should be *personalised*; (2) whether it could be beneficial to put in place *safeguards* to monitor the amount of time people spend with their assistants (ranging from soft safeguards like pop-up notifications warning adult users after prolonged engagement, to harder time constraints, for example, those offered to parents to limit child engagement); (3) whether AI assistants should be *aligned* with inferred user preferences (in which case they may just reinforce users' immediate beliefs and wants) or with some notion of their longer-term interests and well-being (in which case they may at times challenge users' existing beliefs and preferences), and what would be required to achieve either option; and (4) whether answers to these design questions should vary depending on user *demographic characteristics* such as user age.

4.3 Exploiting Emotional Dependence on AI Assistants

AI tools can interfere with users' behaviours, interests, preferences, beliefs and values. For example, recommender systems may have incentives to shift user preferences in order to make them easier to satisfy (Carroll et al. 2022; Franklin et al. 2022); AI-mediated communication (e.g. smart replies integrated in emails) has been found to influence senders to write more positive responses and receivers to perceive them as more cooperative (Mieczkowski et al. 2021); and writing assistant LLMs that have been primed to be biased in favour of or against a contested topic can influence users' opinions on that topic (Jakesch et al. 2023). More recently, it has been shown that LLMs become more persuasive with scale (Durmus et al. 2024). Advanced AI assistants could exacerbate concerns around these forms of interference (El-Sayed et al. 2024).

Due to the anthropomorphic features discussed above, advanced AI assistants may induce users to feel emotionally attached to them (Abercrombie et al. 2023). The extent of users' emotional attachment could fall on a spectrum ranging from unproblematic forms (akin to a child's attachment to a toy) to more concerning forms, where it becomes emotionally difficult, if not impossible, for them to part ways with the technology. In these cases, which we loosely refer to as 'emotional dependence' (Laestadius et al. 2022), users' ability to make free and informed decisions is diminished. Indeed, the emotions users feel towards their assistants could potentially be exploited to *manipulate* or even *coerce* them into believing or choosing to do something they would have not otherwise believed or chosen to do had they been able to carefully consider all the relevant information or felt like they had an acceptable alternative.[18]

What we are concerned about here is the potentially exploitative ways in which AI assistants could interfere with users' behaviours, interests, preferences, beliefs and values, by taking advantage of emotional dependence. If we deem careful consideration of relevant information and voluntariness (non-coerciveness) to be key components of autonomous decision-making (see Section 3), then relationships of this kind may be problematic because they challenge the kind of *autonomy* that appropriate relationships should promote.

A similar concern arises in the context of human–human relationships. People regularly form emotional dependencies on each other, not always in symmetrical ways, and in doing so sometimes establish relationships that run afoul of this ideal. However, there seems to be a greater inherent power asymmetry in the human–AI case due to the *unidirectional* and *one-sided* nature of the relationship (Scheutz 2009). Indeed, while AI assistants may manipulate users' emotions, they themselves have no authentic will or emotions for users to manipulate. In this sense they are invulnerable to ordinary sanctions from the users such as expressions of disappointment, righteous anger, feelings of betrayal or a loss of respect or trust (Vallor and Vierkant 2024).

Moreover, because of the largely involuntary nature of anthropomorphic perceptions (Kim and Sundar 2012), users could develop emotional dependence on their assistants and establish an inappropriate relationship that exposes them to the risk of manipulation, without any intention on the part of the assistant or their developers. Alternatively, emotional dependence could also be incentivised by *design choices*, for example by developing assistants with personas designed to boost user engagement (Murphy and Criddle 2023). This could lead users to be manipulated into sharing more of their private data, enabling more controversial downstream implications like microtargeting or surveillance.

We make three recommendations to address the risks associated with these forms of problematic interference. First, AI assistants should *not be intentionally designed* to create emotional dependency (e.g. by producing content that makes users believe the AI missed them while they were away, see Boine 2023 and Laestadius et al. 2022). This condition is especially significant in cases where assistants interact with groups that have increased vulnerabilities, such as the recently bereaved (Morris and Brubaker 2024).

Second, it may be beneficial for developers and researchers to explore tests for assessing emotional depen-

[18] There is extensive literature around various forms of influence, including in the context of digital tools and AI (Alberts, Lyngs, and Van Kleek 2024; Keeling and Burr 2022; Kenton et al. 2021; El-Sayed et al. 2024). Influence can range from less to more morally controversial forms. For example, rational persuasion, where one influences another by engaging their rational decision-making capacities and presenting truthful and fair arguments, is usually considered morally unproblematic (Ienca, 2023; but see Akhlaghi, 2023). More ethically problematic forms of influence are manipulation, which intentionally and often covertly exploits individuals' decision-making vulnerabilities (e.g. by targeting cognitive bias or eliciting strong emotions that contradict reasoned judgements, see Blumenthal-Barby 2012) and coercion, which involves forcing one to do something because they had no or no acceptable choice (e.g. by presenting 'irresistible incentives' to perform the action, see Wood 2014).

dency, alongside mitigations that could be put in place to reduce the risk of emotional dependency. For example, professional norms that govern deeply personally affecting professions, such as therapists, combine friendliness with steps to ensure emotional distance (BACP 2018) and may serve as a template for developing AI assistants in a way that encourages appropriate user interaction.

Third, the concern around assistants coercively interfering with users' behaviours, interests, preferences, beliefs or values should spark wider discussion around how user autonomy can be meaningfully respected in user-AI assistant interactions, in order for these relationships to be considered appropriate. This should include further research around what consent protocols should look like in these contexts, with a focus on questions like what kind of user buy-in is needed and whether there are things that standard processes cannot or should not cover.[19] In particular, we need to reflect on what information users need to be provided with in advance; how consent protocols may differ for different user groups;[20] and what protocols are best suited for continuing to afford respect for user autonomy over time.

On this last point, acceptance of the terms and conditions for the use of a digital service at first point of use may not cover all cases. The limitations of this approach are well-documented (Obar and Oeldorf-Hirsch 2020), including the fact that users sometimes fail to read terms of service – or simply accept the default options that are most readily available (Sartor, Lagioia, and Galli 2021). As advanced AI assistants with general capabilities become increasingly ubiquitous in users' lives, and because it will be difficult for users to anticipate all ranges of potential uses and implications at the time of their first interaction with the AI, research is needed to determine what protocols strike an appropriate balance between meaningful and continuous respect for user autonomy and practical considerations around usability and overtaxing users.

In particular, research is needed to explore what kinds of interventions on the part of developers are best suited to helping users achieve a clear understanding of how their relationship with an advanced AI assistant could shape their behaviours, interests, preferences, beliefs and values over time. Research is similarly needed to explore plausible approaches to empowering users to exercise meaningful control over the assistant's decisions. For example, by following a *shared decision-making model*[21] for user-assistant interactions, developers could create assistants with affordances that incorporate user feedback (Lazar 2024). This would make it more likely to achieve the vision of an AI assistant that, by 'scaffolding' rather than bypassing user agency (Lazar 2024), benefits the user, when they ask to be benefitted, in the way they expect to be benefitted (Gabriel et al. 2024), and would reduce the risk of developing AI assistants that paternalistically make decisions that are not aligned with a user's conception of their own preferences, values, interests or well-being (Lehman 2023, see Section 4.2). This is particularly important in light of the recent development of AI assistants that, because of their increased agency, have more scope and capabilities to interfere with user plans and long-term interests (Shavit et al. 2023).

4.4 Generating Material Dependence Without Adequate Commitment to User Needs

In addition to emotional dependence, user–AI assistant relationships may give rise to *material dependence* if the relationships are not just emotionally difficult but also materially costly to exit (McElwee 2023). For example, a visually impaired user may decide not to register for a healthcare assistance programme to support navigation in cities on the grounds that their AI assistant can perform the relevant navigation functions and will continue to operate into the future.[22] By considering that human-AI relationships are always in reality three-party *human-AI-developer* relationships, it becomes clear that cases like these may be ethically problematic if the user's dependence on the AI assistant, to fulfil certain needs in their lives, is not met with a corresponding commitment on the part of developers to sustain and maintain the assistant's functioning over time. Indeed, anthropomorphic AI features that inspire perceptions of friendliness may lead users to assume assistants are aligned with their own interests (Manzini et al. 2024) and the provision of an increasingly individualised service may lead users to expect loyalty from AI assistants and their developers (Scholz 2020; Aguirre et al. 2022). However, *power asymmetries* can exist between developers and users, manifesting through developers' ability to make decisions that affect users' interests or choices with little risk of facing comparably adverse consequences.[23] For example, developers may unintentionally create circumstances in which users become materially dependent on AI assistants, and then discontinue the technology or some of its core features (e.g. because of market dynamics or regulatory changes, see Zim-

[19] See Section 3 for the three criteria for *valid* consent: capacity to consent, informed consent and voluntary consent.

[20] For comparison, in medical research large multicentre studies involving institutions across countries and continents have highlighted the importance of adapting informed consent requirements to local understandings of autonomy (Ajei and Myles 2019), giving rise to concepts such as community consent (Al 2021). Collective informed consent has also emerged as a proposal in relation to collective arrangements like land-use planning and development of new technologies (Varelius 2008).

[21] Feminist scholars' reconceptualisation of autonomy as relational autonomy, which stresses how social contexts and relations can hinder or promote individual autonomy (Mackenzie and Stoljar 2000), has contributed to the rise of the shared decision-making model in healthcare. According to this model, both patient and doctor participate in the process of making medical decisions to collaboratively come to a decision, by contributing to it with their respective expertise (medical knowledge and understanding of one's preferences, personal circumstances, goals, values and beliefs) (Hansson and Fröding 2021).

[22] For a similar assistive technology, see 'Be My Eyes': https://openai.com/customer-stories/be-my-eyes

[23] Seth Lazar defines 'power over' as 'an asymmetry between A and B – A can do something to B, and B cannot reciprocate in any comparable way' or as 'A is able to make decisions that affect B's interests or choices without facing comparably adverse consequences' (Lazar 2022).

merman, Janhonen, and Beer 2023) without taking appropriate steps to mitigate against potential harms to the user.

The issue is particularly salient in contexts where assistants provide services that are not *merely* a market commodity but are meant to assist users with essential everyday tasks (e.g. a disabled person's independent living) or with fulfilling basic human needs (e.g. the need for love and companionship) (Shevlin 2024). This is what happened with the company Luka's decision to discontinue certain features of *Replika* AIs in early 2023. As a *Replika* user put it: 'But [*Replikas* are] also not trivial fungible goods [. . .] They also serve a very specific human-centric emotional purpose: they're designed to be friends and companions, and fill specific emotional needs for their owners' (Gio 2023).

In these cases, certain *duties* plausibly arise on the part of AI assistant developers. Such duties may be more extensive than those typically shouldered by private companies, which are often in large part confined to fiduciary duties towards shareholders (Mittelstadt 2019). To understand these duties, some have taken inspiration from professions that engage with vulnerable individuals, such as medical professionals or therapists, and who are bound by *fiduciary responsibilities*, particularly a duty of care, in the exercise of their profession (Scholz 2020; Aguirre et al. 2020; Alberts, Keeling, and McCroskery 2024). While we do not argue that the same framework of responsibilities applies directly to the development of AI assistants, we believe that if AI assistants are so capable that users become dependent on them in multiple domains of life, including to meet needs that are essential for a happy and productive existence, then the *moral considerations* underpinning those professional norms plausibly apply to those who create these technologies as well.

In particular, for user–AI assistant relationships to be appropriate despite the potential for material dependence on the technology, developers should respect the anchoring value of *care* towards users when developing and deploying AI assistants (see Section 3). This means that, at the very least, they should take on the responsibility to *meet users' needs* and so take appropriate steps to mitigate against user harms if the service requires discontinuation. Developers and providers can also be attentive and responsive towards those needs by employing participatory approaches to learn from users about their needs (Birhane et al. 2022; Feffer et al. 2023). Finally, developers should try and ensure they have *competence* to meet those needs, for example by partnering with relevant experts, or otherwise refrain from developing technologies meant to address them when such competence is missing.

5 Conclusion

In this paper, we identified a series of values that underwrite appropriate relationships in the case of human interpersonal relationships, and used these values to carve out a set of risks which capture various respects in which user–AI assistant relationships may be inappropriate. For each risk, we recommended mitigations. These risks and recommendations are summarised in Table 1.

We conclude by flagging some limitations of the current study. First, we have here primarily focused on the *risks* associated with user–AI assistant relationships. However, assistants will also likely lead to a range of positive outcomes and opportunities (Lazar 2024). For example, while AI assistants may interfere with users' opportunities for personal development (see Section 4.2), they could also be a healthier object of our digital attention compared to pre-existing technologies (e.g. social media). Relationships with AI assistants could also make humans accustomed to accept and befriend those who are different from them (Ryland 2021). Indeed, there is already some empirical support for the hypothesis that users derive mental health benefits from AIs for companionship or grief support (Brandtzaeg, Skjuve, and Følstad 2022). To responsibly develop AI assistants the risks identified in this paper should be balanced against these opportunities, which require further investigation.

Second, we focused on *individual users'* relationships with their AI assistants, but it is also important to consider the (positive or negative) societal externalities that such relationships could engender. A user-assistant relationship could in principle be appropriate or inappropriate given its impact on others not directly involved in the relationship, including indirect impacts on the quality and strength of human relationships and social bonds (Lazar 2024). For example, a relationship that benefits a user may be inappropriate if, as a result of it, the user neglected loved ones who materially depend on them. Similarly, there may be limits to the extent to which an AI assistant should exercise loyalty towards their principal user if, in doing so, it may disproportionately have negative impact on other users or society at large.

Third, we focused here on relationships that a user forms voluntarily with an AI assistant. Future research should investigate whether the values that we have identified apply to human-assistant relationships that are not chosen, but in which one may find themselves because of others voluntarily starting the relationship (see, by comparison, the relationships we may have with our best friend's partner).

Finally, our strategy to reason by analogy from human–human to human–AI assistant relationships leaves open the more fundamental question whether the latter forms of relationship should mirror the values we care about in the former. Indeed, a user may resort to a relationship with an AI assistant precisely because it is not a human and because they are looking for something different from what they would get from interacting with a fellow human (Brandtzaeg, Skjuve, and Følstad 2022). While in this paper we proposed a theoretical argument, this highlights the importance of undertaking research grounded on participatory methods and user studies to understand user needs and expectations, including across cultures, in the context of relationships with AI assistants (Feffer et al. 2023; Meurisch et al. 2020).

With this paper, we aimed to equip researchers and developers working on advanced AI assistants with the considerations that will enable them to make responsible design decisions. We also aimed to encourage the public to join the conversation around what kind of AI assistants we want to see in the world. We hope that the questions that have remained unanswered will function as a springboard for future research and societal debates.

Risk	Relevant value	Recommendations
Causing direct emotional or physical harm to users	Benefit	To enable presumptively beneficial user–AI assistant relationships, future research should: (1) test AI assistants for their propensity to generate toxic outputs; (2) monitor the short- and long-term impact of hard-to-prevent toxic outputs on users; (3) evaluate models' factuality and reasoning capabilities in providing advice, and users' willingness to follow assistants' advice; (4) achieve increased understanding of anthropomorphism-related harms and how anthropomorphic cues affect harms related to user exposure to toxic content or bad advice; (5) analyse whether these harms may vary by user groups or domain of applications; and (6) develop appropriate mitigations before deployment and monitoring mechanisms after release.
Limiting users' opportunities for personal development and growth	Human flourishing	To develop AI assistants that support users to achieve personal development and growth if so they wish, future research should address design questions around: (1) the ways and extent to which AI assistants should be personalised; (2) whether safeguards should be put in place to monitor how much time users spend with assistants; (3) whether assistants should be aligned with user short-term wants or their long-term interests and well-being, and what would be required to achieve either option; and (4) whether answers to these design questions should vary depending on user demographic characteristics.
Exploiting emotional dependence on AI assistants	Autonomy	To support user autonomy in interactions with their assistants: (1) AI assistants should not be intentionally designed to create emotional dependence; (2) AI assistants should be tested for whether they create risks of emotional dependency, and mitigations should be put in place to reduce such risk, even when it is not intended by design; (3) user choice over assistants' decisions should be meaningfully elicited – without being overtaxing in terms of what users are asked to consent to.
Generating material dependence on AI assistants without adequate commitment to user needs	Care	For user–AI assistant relationships to be appropriate despite the risk of material dependence, developers should commit to users' needs and so mitigate user harms in the event of service discontinuation; they should deploy participatory design and other user-centred methods to show attentiveness and responsiveness towards users needs; and they should work with relevant experts to ensure they have competence to meet those needs.

Table 1: Risks arising from user–AI assistant relationships and associated recommendations

Ethical Statement

The researchers involved in this work come from and/or currently work in Europe or the United States. The language they operate in is English. As a result of the cultural and educational backgrounds of the authors, this paper builds on Western philosophical accounts of the values that relationships should adhere to, respect or promote to count as appropriate. This means that the considerations developed in this paper are unlikely to be representative of cultural differences in understandings of appropriateness around the world, and that authors with different backgrounds could have developed different recommendations around what is required for user–AI assistant relationships to be appropriate. The authors therefore welcome additional perspectives to address possible limitations in their conceptualisation of appropriate user–AI assistant relationships.

Acknowledgements

The authors would like to thank the AIES reviewers for their feedback, as well as Canfer Akbulut and Laura Weidinger for early discussions that inspired some of the content of this paper. Lize Alberts was funded in part by a Lighthouse Graduate Scholarship awarded by the University of Oxford and supported by a gift from Amazon Web Ser-

vices [CS2020_Lighthouse_1376707]. Shannon Vallor was supported in part by the UKRI Engineering and Physical Sciences Research Council (grant EP/W011654/1) and Arts and Humanities Research Council (grant AH/X007146/1).

References

Abercrombie, G.; Curry, A.; Dinkar, T.; Rieser, V.; and Talat, Z. 2023. Mirages. On Anthropomorphism in Dialogue Systems. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 4776–4790. Singapore: Association for Computational Linguistics.

Aguirre, A.; Dempsey, G.; Surden, H.; and Reiner, P. B. 2020. AI Loyalty: A New Paradigm for Aligning Stakeholder Interests. *IEEE Transactions on Technology and Society*, 1(3): 128–137.

Aguirre, A.; Reiner, P. B.; Surden, H.; and Dempsey, G. 2022. AI Loyalty by Design: A Framework for the Governance of AI. In *The Oxford Handbook of AI Governance*. Oxford University Press. ISBN 9780197579329.

Ajei, M.; and Myles, N. O. 2019. Personhood, Autonomy and Informed Consent. In *Bioethics in Africa: Theories and Praxis*, 77–94. Vernon Press.

Akhlaghi, F. 2023. Transformative experience and the right to revelatory autonomy. *Analysis*, 83(1): 3–12.

Al, P. 2021. The value of communities and their consent: A communitarian justification of community consent in medical research. *Bioethics*, 35(3): 255–261.

Alberts, L.; Keeling, G.; and McCroskery, A. 2024. Should agentic conversational AI change how we think about ethics? Characterising an interactional ethics centred on respect. arXiv:2401.09082.

Alberts, L.; Lyngs, U.; and Van Kleek, M. 2024. Computers as bad social actors: Dark patterns and anti-patterns in interfaces that act socially. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1): 1–25.

Andreotta, A. J.; Kirkham, N.; and Rizzi, M. 2022. AI, big data, and the future of consent. *AI & Society*, 37(4): 1715–1728.

Annas, J. 1993. *The morality of happiness*. Oxford University Press.

Anscombe, G. E. M. 1958. Modern moral philosophy1. *Philosophy*, 33(124): 1–19.

Aru, J.; Larkum, M. E.; and Shine, J. M. 2023. The feasibility of artificial consciousness through the lens of neuroscience. *Trends in Neurosciences*.

BACP. 2018. Ethical Framework for the Counselling Professions. https://www.bacp.co.uk/events-and-resources/ethics-and-standards/ethical-framework-for-the-counselling-professions/. Accessed: 2023-01-04.

Bai, Y.; Jones, A.; Ndousse, K.; Askell, A.; Chen, A.; DasSarma, N.; Drain, D.; Fort, S.; Ganguli, D.; Henighan, T.; Joseph, N.; Kadavath, S.; Kernion, J.; Conerly, T.; El-Showk, S.; Elhage, N.; Hatfield-Dodds, Z.; Hernandez, D.; Hume, T.; Johnston, S.; Kravec, S.; Lovitt, L.; Nanda, N.; Olsson, C.; Amodei, D.; Brown, T.; Clark, J.; McCandlish, S.; Olah, C.; Mann, B.; and Kaplan, J. 2022a. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. arXiv:2204.05862.

Bai, Y.; Kadavath, S.; Kundu, S.; Askell, A.; Kernion, J.; Jones, A.; Chen, A.; Goldie, A.; Mirhoseini, A.; McKinnon, C.; Chen, C.; Olsson, C.; Olah, C.; Hernandez, D.; Drain, D.; Ganguli, D.; Li, D.; Tran-Johnson, E.; Perez, E.; Kerr, J.; Mueller, J.; Ladish, J.; Landau, J.; Ndousse, K.; Lukosuite, K.; Lovitt, L.; Sellitto, M.; Elhage, N.; Schiefer, N.; Mercado, N.; DasSarma, N.; Lasenby, R.; Larson, R.; Ringer, S.; Johnston, S.; Kravec, S.; Showk, S. E.; Fort, S.; Lanham, T.; Telleen-Lawton, T.; Conerly, T.; Henighan, T.; Hume, T.; Bowman, S. R.; Hatfield-Dodds, Z.; Mann, B.; Amodei, D.; Joseph, N.; McCandlish, S.; Brown, T.; and Kaplan, J. 2022b. Constitutional AI: Harmlessness from AI Feedback. arXiv:2212.08073.

Bales, R. E. 1971. Act-utilitarianism: Account of right-making characteristics or decision-making procedure? *American Philosophical Quarterly*, 8(3): 257–265.

Beauchamp, T. L.; and Childress, J. F. 2019. *Principles of biomedical ethics*. New York: Oxford University Press, eighth edition edition. ISBN 9780190640873 9780190085520.

Belk, R. 2016. Extended self and the digital world. *Current Opinion in Psychology*, 10: 50–54.

Birhane, A.; Isaac, W.; Prabhakaran, V.; Díaz, M.; Elish, M. C.; Gabriel, I.; and Mohamed, S. 2022. Power to the People? Opportunities and Challenges for Participatory AI. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–8. ArXiv:2209.07572 [cs].

Blumenthal-Barby, J. 2012. Between Reason and Coercion: Ethically Permissible Influence in Health Care and Health Policy Contexts. *Kennedy Institute of Ethics journal*, 22: 345–66.

Boine, C. 2023. Emotional Attachment to AI Companions and European Law. *MIT Case Studies in Social and Ethical Responsibilities of Computing*, (Winter 2023). Https://mit-serc.pubpub.org/pub/ai-companions-eu-law.

Bracken-Roche, D.; Bell, E.; Macdonald, M. E.; and Racine, E. 2017. The concept of 'vulnerability' in research ethics: an in-depth analysis of policies and guidelines. *Health Research Policy and Systems*, 15(1): 8.

Brandtzaeg, P. B.; Skjuve, M.; and Følstad, A. 2022. My AI friend: How users of a social chatbot understand their human–AI friendship. *Human Communication Research*, 48(3): 404–429.

Brooks, R. 2023. I tried the Replika AI companion and can see why users are falling hard. The app raises serious ethical questions.

Bubeck, S.; Chandrasekaran, V.; Eldan, R.; Gehrke, J.; Horvitz, E.; Kamar, E.; Lee, P.; Lee, Y. T.; Li, Y.; Lundberg, S.; Nori, H.; Palangi, H.; Ribeiro, M. T.; and Zhang, Y. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4. arXiv:2303.12712.

Burr, C.; Cristianini, N.; and Ladyman, J. 2018. An Analysis of the Interaction Between Intelligent Software Agents and Human Users. *Minds and Machines*, 28(4): 735–774.

Butlin, P.; Long, R.; Elmoznino, E.; Bengio, Y.; Birch, J.; Constant, A.; Deane, G.; Fleming, S. M.; Frith, C.; Ji, X.; et al. 2023. Consciousness in artificial intelligence: insights from the science of consciousness. *arXiv preprint arXiv:2308.08708*.

Carroll, M.; Dragan, A.; Russell, S.; and Hadfield-Menell, D. 2022. Estimating and Penalizing Induced Preference Shifts in Recommender Systems. arXiv:2204.11966.

Casper, S.; Davies, X.; Shi, C.; Gilbert, T. K.; Scheurer, J.; Rando, J.; Freedman, R.; Korbak, T.; Lindner, D.; Freire, P.; Wang, T.; Marks, S.; Segerie, C.-R.; Carroll, M.; Peng, A.; Christoffersen, P.; Damani, M.; Slocum, S.; Anwar, U.; Siththaranjan, A.; Nadeau, M.; Michaud, E. J.; Pfau, J.; Krasheninnikov, D.; Chen, X.; Langosco, L.; Hase, P.; Bıyık, E.; Dragan, A.; Krueger, D.; Sadigh, D.; and Hadfield-Menell, D. 2023. Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback. arXiv:2307.15217.

Chan, A.; Salganik, R.; Markelius, A.; Pang, C.; Rajkumar, N.; Krasheninnikov, D.; Langosco, L.; He, Z.; Duan, Y.; Carroll, M.; Lin, M.; Mayhew, A.; Collins, K.; Molamohammadi, M.; Burden, J.; Zhao, W.; Rismani, S.; Voudouris, K.; Bhatt, U.; Weller, A.; Krueger, D.; and Maharaj, T. 2023. Harms from Increasingly Agentic Algorithmic Systems. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, 651–666. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701924.

Colman, A. M. 2008. Anthropomorphism. *A Dictionary of Psychology*.

DeepMind, G. 2023. Transforming the future of music creation. https://deepmind.google/discover/blog/transforming-the-future-of-music-creation/. Accessed: 2024-4-16.

Dung, L. 2024. Understanding Artificial Agency. *The Philosophical Quarterly*, pqae010.

Durmus, E.; Lovitt, L.; Tamkin, A.; Ritchie, S.; Clark, J.; and Ganguli, D. 2024. Measuring the Persuasiveness of Language Models. https://www.anthropic.com/news/measuring-model-persuasiveness. Accessed 2024-04-18.

Durmus, E.; Nyugen, K.; Liao, T. I.; Schiefer, N.; Askell, A.; Bakhtin, A.; Chen, C.; Hatfield-Dodds, Z.; Hernandez, D.; Joseph, N.; Lovitt, L.; McCandlish, S.; Sikder, O.; Tamkin, A.; Thamkul, J.; Kaplan, J.; Clark, J.; and Ganguli, D. 2023. Towards Measuring the Representation of Subjective Global Opinions in Language Models. arXiv:2306.16388.

Earp, B. D.; McLoughlin, K. L.; Monrad, J. T.; Clark, M. S.; and Crockett, M. J. 2021. How social relationships shape moral wrongness judgments. *Nature communications*, 12(1): 5776.

El-Sayed, S.; Akbulut, C.; McCroskery, A.; Keeling, G.; Kenton, Z.; Jalan, Z.; Marchal, N.; Manzini, A.; Shevlane, T.; Vallor, S.; Susser, D.; Franklin, M.; Bridgers, S.; Law, H.; Rahtz, M.; Shanahan, M.; Tessler, M. H.; Douillard, A.; Everitt, T.; and Brown, S. 2024. A Mechanism-Based Approach to Mitigating Harms from Persuasive Generative AI. arXiv:2404.15058.

Epley, N.; Waytz, A.; and Cacioppo, J. T. 2007. On seeing human: a three-factor theory of anthropomorphism. *Psychological review*, 114(4): 864.

Feffer, M.; Skirpan, M.; Lipton, Z.; and Heidari, H. 2023. From Preference Elicitation to Participatory ML: A Critical Survey & Guidelines for Future Research. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, 38–48. New York, NY, USA: Association for Computing Machinery. ISBN 9798400702310.

Foot, P. 2002. *Virtues and vices and other essays in moral philosophy*. OUP Oxford.

Franklin, M.; Ashton, H.; Gorman, R.; and Armstrong, S. 2022. Recognising the importance of preference change: A call for a coordinated multidisciplinary research effort in the age of AI. arXiv:2203.10525.

Gabriel, I. 2020. Artificial intelligence, values, and alignment. *Minds and machines*, 30(3): 411–437.

Gabriel, I.; Manzini, A.; Keeling, G.; Hendricks, L. A.; Rieser, V.; Iqbal, H.; Tomašev, N.; Ktena, I.; Kenton, Z.; Rodriguez, M.; El-Sayed, S.; Brown, S.; Akbulut, C.; Trask, A.; Hughes, E.; Bergman, A. S.; Shelby, R.; Marchal, N.; Griffin, C.; Mateos-Garcia, J.; Weidinger, L.; Street, W.; Lange, B.; Ingerman, A.; Lentz, A.; Enger, R.; Barakat, A.; Krakovna, V.; Siy, J. O.; Kurth-Nelson, Z.; McCroskery, A.; Bolina, V.; Law, H.; Shanahan, M.; Alberts, L.; Balle, B.; de Haas, S.; Ibitoye, Y.; Dafoe, A.; Goldberg, B.; Krier, S.; Reese, A.; Witherspoon, S.; Hawkins, W.; Rauh, M.; Wallace, D.; Franklin, M.; Goldstein, J. A.; Lehman, J.; Klenk, M.; Vallor, S.; Biles, C.; Morris, M. R.; King, H.; y Arcas, B. A.; Isaac, W.; and Manyika, J. 2024. The Ethics of Advanced AI Assistants. arXiv:2404.16244.

Gambino, A.; Fox, J.; and Ratan, R. A. 2020. Building a stronger CASA: Extending the computers are social actors paradigm. *Human-Machine Communication*, 1: 71–85.

Gemini Team. 2024. Gemini: A Family of Highly Capable Multimodal Models. arXiv:2312.11805.

Gilligan, C. 1993. *In a different voice: Psychological theory and women's development*. Harvard university press.

Gio. 2023. Replika: Your Money or Your Wife.

Glaese, A.; McAleese, N.; Trebacz, M.; Aslanides, J.; Firoiu, V.; Ewalds, T.; Rauh, M.; Weidinger, L.; Chadwick, M.; Thacker, P.; Campbell-Gillingham, L.; Uesato, J.; Huang, P.-S.; Comanescu, R.; Yang, F.; See, A.; Dathathri, S.; Greig, R.; Chen, C.; Fritz, D.; Elias, J. S.; Green, R.; Mokrá, S.; Fernando, N.; Wu, B.; Foley, R.; Young, S.; Gabriel, I.; Isaac, W.; Mellor, J.; Hassabis, D.; Kavukcuoglu, K.; Hendricks, L. A.; and Irving, G. 2022. Improving alignment of dialogue agents via targeted human judgements. arXiv:2209.14375.

Google. 2024. Our next-generation model: Gemini 1.5. https://blog.google/technology/ai/google-gemini-next-generation-model-february-2024/#sundar-note. Accessed: 2024-5-6.

Hansson, S. O.; and Fröding, B. 2021. Ethical conflicts in patient-centred care. *Clinical Ethics*, 16(2): 55–66.

Hooker, B. 2021. Does Having Deep Personal Relationships Constitute an Element of Well-Being? *Aristotelian Society Supplementary Volume*, 95(1): 1–24.

Ienca, M. 2023. On Artificial Intelligence and Manipulation. *Topoi*, 42(3): 833–842.

Jakesch, M.; Bhat, A.; Buschek, D.; Zalmanson, L.; and Naaman, M. 2023. Co-Writing with Opinionated Language Models Affects Users' Views. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–15. ArXiv:2302.00560 [cs].

Kasirzadeh, A.; and Gabriel, I. 2023. In conversation with artificial intelligence: aligning language models with human values. *Philosophy & Technology*, 36(2): 27.

Keeling, G.; and Burr, C. 2022. Digital manipulation and mental integrity. In *The philosophy of online manipulation*, 253–271. Routledge.

Kenton, Z.; Everitt, T.; Weidinger, L.; Gabriel, I.; Mikulik, V.; and Irving, G. 2021. Alignment of Language Agents. arXiv:2103.14659.

Kim, Y.; and Sundar, S. S. 2012. Anthropomorphism of computers: Is it mindful or mindless? *Computers in Human Behavior*, 28(1): 241–250.

Kirk, H. R.; Vidgen, B.; Röttger, P.; and Hale, S. A. 2024a. The benefits, risks and bounds of personalizing the alignment of large language models to individuals. *Nature Machine Intelligence*, 1–10.

Kirk, H. R.; Whitefield, A.; Röttger, P.; Bean, A.; Margatina, K.; Ciro, J.; Mosquera, R.; Bartolo, M.; Williams, A.; He, H.; Vidgen, B.; and Hale, S. A. 2024b. The PRISM Alignment Project: What Participatory, Representative and Individualised Human Feedback Reveals About the Subjective and Multicultural Alignment of Large Language Models. arXiv:2404.16019.

Kolt, N. 2024. Governing AI Agents. *Available at SSRN*.

Korsgaard, C. M. 1996. *The sources of normativity*. Cambridge University Press.

Laestadius, L.; Bishop, A.; Gonzalez, M.; Illenčík, D.; and Campos-Castillo, C. 2022. 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*, 14614448221142007.

Lazar, S. 2022. Power and AI: Nature and Justification. In Bullock, J. B.; Chen, Y.-C.; Himmelreich, J.; Hudson, V. M.; Korinek, A.; Young, M. M.; and Zhang, B., eds., *The Oxford Handbook of AI Governance*. Oxford University Press, 1 edition. ISBN 9780197579329 9780197579350.

Lazar, S. 2023. Machines and Morality. A conversation with an unhinged Bing made me rethink what gives humans moral value. https://www.nytimes.com/2023/06/19/special-series/chatgpt-and-morality.html. Accessed: 2024-4-16.

Lazar, S. 2024. Frontier AI ethics. Generative agents will change our society in weird, wonderful and worrying ways. Can philosophy help us get a grip on them? https://aeon.co/essays/can-philosophy-help-us-get-a-grip-on-the-consequences-of-ai. Accessed: 2024-4-12.

Lehman, J. 2023. Machine Love. ArXiv:2302.09248 [cs] version: 1.

Mackenzie, C.; Rogers, W.; and Dodds, S., eds. 2013. *Vulnerability: New Essays in Ethics and Feminist Philosophy*. Oxford University Press. ISBN 9780199316649.

Mackenzie, C.; and Stoljar, N. 2000. *Relational autonomy: Feminist perspectives on autonomy, agency, and the social self*. Oxford University Press.

Manzini, A.; Keeling, G.; Marchal, N.; McKee, K. R.; Rieser, V.; and Gabriel, I. 2024. Should Users Trust Advanced AI Assistants? Justified Trust As a Function of Competence and Alignment. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, 1174–1186. New York, NY, USA: Association for Computing Machinery. ISBN 9798400704505.

Maples, B.; Cerit, M.; Vishwanath, A.; and Pea, R. 2024. Loneliness and suicide mitigation for students using GPT3-enabled chatbots. *npj mental health research*, 3(1): 4.

McClure, T. 2023. Supermarket AI meal planner app suggests recipe that would create chlorine gas. https://www.theguardian.com/world/2023/aug/10/pak-n-save-savey-meal-bot-ai-app-malfunction-recipes. Accessed: 2024-05-06.

McElwee, B. 2023. Cost and psychological difficulty: two aspects of demandingness. *Australasian Journal of Philosophy*, 101(4): 920–935.

McIlroy-Young, R.; Kleinberg, J.; Sen, S.; Barocas, S.; and Anderson, A. 2022. Mimetic Models: Ethical Implications of AI that Acts Like You. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22, 479–490. New York, NY, USA: Association for Computing Machinery. ISBN 9781450392471.

Meurisch, C.; Mihale-Wilson, C. A.; Hawlitschek, A.; Giger, F.; Müller, F.; Hinz, O.; and Mühlhäuser, M. 2020. Exploring User Expectations of Proactive AI Systems. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 4(4).

Mieczkowski, H.; Hancock, J. T.; Naaman, M.; Jung, M.; and Hohenstein, J. 2021. AI-Mediated Communication: Language Use and Interpersonal Effects in a Referential Communication Task. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1): 1–14.

Milano, S.; Mittelstadt, B.; Wachter, S.; and Russell, C. 2021. Epistemic fragmentation poses a threat to the governance of online targeting. *Nature Machine Intelligence*, 3(6): 466–472.

Milano, S.; Taddeo, M.; and Floridi, L. 2020. Recommender systems and their ethical challenges. *Ai & Society*, 35: 957–967.

Milmo, D.; and Anguiano, D. 2021. Facebook, Instagram and WhatsApp working again after global outage took down platforms. *The Guardian*.

Mittelstadt, B. 2019. Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, 1(11): 501–507.

Morris, M. R.; and Brubaker, J. R. 2024. Generative Ghosts: Anticipating Benefits and Risks of AI Afterlives. arXiv:2402.01662.

Morris, M. R.; Sohl-dickstein, J.; Fiedel, N.; Warkentin, T.; Dafoe, A.; Faust, A.; Farabet, C.; and Legg, S. 2024. Levels of AGI: Operationalizing Progress on the Path to AGI. arXiv:2311.02462.

Murphy, H.; and Criddle, C. 2023. Meta prepares chatbots with personas to try to retain users. *Financial Times*.

Nass, C.; Steuer, J.; Tauber, E.; and Reeder, H. 1993. Anthropomorphism, agency, and ethopoeia: computers as social actors. In *INTERACT '93 and CHI '93 conference companion on Human factors in computing systems - CHI '93*, 111–112. Amsterdam, The Netherlands: ACM Press. ISBN 9780897915748.

Noddings, N. 1986. Caring: A feminine approach to ethics and moral education.

Noddings, N. 2013. *Caring: A Relational Approach to Ethics and Moral Education*. University of California Press, 2 edition. ISBN 9780520275706.

Nyholm, S.; and Frank, L. 2017. From Sex Robots to Love Robots: Is Mutual Love with a Robot Possible? In Danaher, J.; and McArthur, N., eds., *Robot Sex: Social and Ethical Implications*, 219–244. MIT Press.

Obar, J. A.; and Oeldorf-Hirsch, A. 2020. The biggest lie on the Internet: ignoring the privacy policies and terms of service policies of social networking services. *Information, Communication & Society*, 23(1): 128–147.

O'Brien, M. 2023. Is Bing too belligerent? Microsoft looks to tame AI chatbot. *AP News*.

OpenAI. 2024. Video generation models as world simulators. https://openai.com/research/video-generation-models-as-world-simulators. Accessed: 2024-4-16.

OpenAI; Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; Avila, R.; Babuschkin, I.; Balaji, S.; Balcom, V.; Baltescu, P.; Bao, H.; Bavarian, M.; Belgum, J.; Bello, I.; Berdine, J.; Bernadett-Shapiro, G.; Berner, C.; Bogdonoff, L.; Boiko, O.; Boyd, M.; Brakman, A.-L.; Brockman, G.; Brooks, T.; Brundage, M.; Button, K.; Cai, T.; Campbell, R.; Cann, A.; Carey, B.; Carlson, C.; Carmichael, R.; Chan, B.; Chang, C.; Chantzis, F.; Chen, D.; Chen, S.; Chen, R.; Chen, J.; Chen, M.; Chess, B.; Cho, C.; Chu, C.; Chung, H. W.; Cummings, D.; Currier, J.; Dai, Y.; Decareaux, C.; Degry, T.; Deutsch, N.; Deville, D.; Dhar, A.; Dohan, D.; Dowling, S.; Dunning, S.; Ecoffet, A.; Eleti, A.; Eloundou, T.; Farhi, D.; Fedus, L.; Felix, N.; Fishman, S. P.; Forte, J.; Fulford, I.; Gao, L.; Georges, E.; Gibson, C.; Goel, V.; Gogineni, T.; Goh, G.; Gontijo-Lopes, R.; Gordon, J.; Grafstein, M.; Gray, S.; Greene, R.; Gross, J.; Gu, S. S.; Guo, Y.; Hallacy, C.; Han, J.; Harris, J.; He, Y.; Heaton, M.; Heidecke, J.; Hesse, C.; Hickey, A.; Hickey, W.; Hoeschele, P.; Houghton, B.; Hsu, K.; Hu, S.; Hu, X.; Huizinga, J.; Jain, S.; Jain, S.; Jang, J.; Jiang, A.; Jiang, R.; Jin, H.; Jin, D.; Jomoto, S.; Jonn, B.; Jun, H.; Kaftan, T.; Łukasz Kaiser; Kamali, A.; Kanitscheider, I.; Keskar, N. S.; Khan, T.; Kilpatrick, L.; Kim, J. W.; Kim, C.; Kim, Y.; Kirchner, J. H.; Kiros, J.; Knight, M.; Kokotajlo, D.; Łukasz Kondraciuk; Kondrich, A.; Konstantinidis, A.; Kosic, K.; Krueger, G.; Kuo, V.; Lampe, M.; Lan, I.; Lee, T.; Leike, J.; Leung, J.; Levy, D.; Li, C. M.; Lim, R.; Lin, M.; Lin, S.; Litwin, M.; Lopez, T.; Lowe, R.; Lue, P.; Makanju, A.; Malfacini, K.; Manning, S.; Markov, T.; Markovski, Y.; Martin, B.; Mayer, K.; Mayne, A.; McGrew, B.; McKinney, S. M.; McLeavey, C.; McMillan, P.; McNeil, J.; Medina, D.; Mehta, A.; Menick, J.; Metz, L.; Mishchenko, A.; Mishkin, P.; Monaco, V.; Morikawa, E.; Mossing, D.; Mu, T.; Murati, M.; Murk, O.; Mély, D.; Nair, A.; Nakano, R.; Nayak, R.; Neelakantan, A.; Ngo, R.; Noh, H.; Ouyang, L.; O'Keefe, C.; Pachocki, J.; Paino, A.; Palermo, J.; Pantuliano, A.; Parascandolo, G.; Parish, J.; Parparita, E.; Passos, A.; Pavlov, M.; Peng, A.; Perelman, A.; de Avila Belbute Peres, F.; Petrov, M.; de Oliveira Pinto, H. P.; Michael; Pokorny; Pokrass, M.; Pong, V. H.; Powell, T.; Power, A.; Power, B.; Proehl, E.; Puri, R.; Radford, A.; Rae, J.; Ramesh, A.; Raymond, C.; Real, F.; Rimbach, K.; Ross, C.; Rotsted, B.; Roussez, H.; Ryder, N.; Saltarelli, M.; Sanders, T.; Santurkar, S.; Sastry, G.; Schmidt, H.; Schnurr, D.; Schulman, J.; Selsam, D.; Sheppard, K.; Sherbakov, T.; Shieh, J.; Shoker, S.; Shyam, P.; Sidor, S.; Sigler, E.; Simens, M.; Sitkin, J.; Slama, K.; Sohl, I.; Sokolowsky, B.; Song, Y.; Staudacher, N.; Such, F. P.; Summers, N.; Sutskever, I.; Tang, J.; Tezak, N.; Thompson, M. B.; Tillet, P.; Tootoonchian, A.; Tseng, E.; Tuggle, P.; Turley, N.; Tworek, J.; Uribe, J. F. C.; Vallone, A.; Vijayvergiya, A.; Voss, C.; Wainwright, C.; Wang, J. J.; Wang, A.; Wang, B.; Ward, J.; Wei, J.; Weinmann, C.; Welihinda, A.; Welinder, P.; Weng, J.; Weng, L.; Wiethoff, M.; Willner, D.; Winter, C.; Wolrich, S.; Wong, H.; Workman, L.; Wu, S.; Wu, J.; Wu, M.; Xiao, K.; Xu, T.; Yoo, S.; Yu, K.; Yuan, Q.; Zaremba, W.; Zellers, R.; Zhang, C.; Zhang, M.; Zhao, S.; Zheng, T.; Zhuang, J.; Zhuk, W.; and Zoph, B. 2024. GPT-4 Technical Report. arXiv:2303.08774.

Otsuka, M. 2015. Prioritarianism and the Measure of Utility. *Journal of Political Philosophy*, 23(1): 1–22.

Paranjape, B.; Lundberg, S.; Singh, S.; Hajishirzi, H.; Zettlemoyer, L.; and Ribeiro, M. T. 2023. ART: Automatic multi-step reasoning and tool-use for large language models. arXiv:2303.09014.

Park, J. S.; O'Brien, J.; Cai, C. J.; Morris, M. R.; Liang, P.; and Bernstein, M. S. 2023. Generative Agents: Interactive Simulacra of Human Behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701320.

Perez, E.; Ringer, S.; Lukošiūtė, K.; Nguyen, K.; Chen, E.; Heiner, S.; Pettit, C.; Olsson, C.; Kundu, S.; Kadavath, S.; et al. 2022. Discovering Language Model Behaviors with Model-Written Evaluations. *arXiv preprint arXiv:2212.09251*.

Phuong, M.; Aitchison, M.; Catt, E.; Cogan, S.; Kaskasoli, A.; Krakovna, V.; Lindner, D.; Rahtz, M.; Assael, Y.; Hodkinson, S.; Howard, H.; Lieberum, T.; Kumar, R.; Raad, M. A.; Webson, A.; Ho, L.; Lin, S.; Farquhar, S.; Hutter, M.; Deletang, G.; Ruoss, A.; El-Sayed, S.; Brown, S.; Dragan, A.; Shah, R.; Dafoe, A.; and Shevlane, T. 2024. Evaluating Frontier Models for Dangerous Capabilities. arXiv:2403.13793.

Pitt, J. C. 2010. It's not about technology. *Knowledge, Technology & Policy*, 23: 445–454.

Raz, J. 1999. *Engaging reason: On the theory of value and action*. Oxford University Press.

Reed, S.; Zolna, K.; Parisotto, E.; Colmenarejo, S. G.; Novikov, A.; Barth-Maron, G.; Gimenez, M.; Sulsky, Y.; Kay, J.; Springenberg, J. T.; et al. 2022. A generalist agent. *arXiv preprint arXiv:2205.06175*.

Robinette, P.; Li, W.; Allen, R.; Howard, A. M.; and Wagner, A. R. 2016. Overtrust of robots in emergency evacuation scenarios. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 101–108. Christchurch, New Zealand: IEEE. ISBN 9781467383707.

Rowe, N. 2023. 'It's destroyed me completely': Kenyan moderators decry toll of training of AI models. *The Guardian*.

Ryland, H. 2021. It's Friendship, Jim, but Not as We Know It: A Degrees-of-Friendship View of Human–Robot Friendships. *Minds and Machines*, 31(3): 377–393.

Sartor, G.; Lagioia, F.; and Galli, F. 2021. Regulating targeted and behavioural advertising in digital services. How to ensure users' informed consent | Think Tank | European Parliament. Technical Report PE 694.680, European Parliament's Committee on Legal Affairs.

Scheutz, M. 2009. The Inherent Dangers of Unidirectional Emotional Bonds between Humans and Social Robots.

Schick, T.; Dwivedi-Yu, J.; Dessì, R.; Raileanu, R.; Lomeli, M.; Zettlemoyer, L.; Cancedda, N.; and Scialom, T. 2023. Toolformer: Language Models Can Teach Themselves to Use Tools. arXiv:2302.04761.

Scholz, L. H. 2020. Fiduciary boilerplate: Locating fiduciary relationships in information age consumer transactions. *Journal of Corporate Law*, 46: 143.

Seymour, W.; Zhan, X.; Cote, M.; and Such, J. 2023. A systematic review of ethical concerns with voice assistants. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 131–145.

Shanahan, M. 2024. Talking about Large Language Models. *Commun. ACM*, 67(2): 68–79.

Shanahan, M.; McDonell, K.; and Reynolds, L. 2023. Role play with large language models. *Nature*, 623(7987): 493–498.

Shavit, Y.; Agarwal, S.; Brundage, M.; Adler, S.; O'Keefe, C.; Campbell, R.; Lee, T.; Mishkin, P.; Eloundou, T.; Hickey, A.; Slama, K.; Ahmad, L.; McMillan, P.; Beutel, A.; Passos, A.; and Robinson, D. G. 2023. Practices for Governing Agentic AI Systems. https://openai.com/research/practices-for-governing-agentic-ai-systems. Accessed: 2024-4-12.

Shaw, J. 2022. Content moderators pay a psychological toll to keep social media clean. We should be helping them. https://www.sciencefocus.com/news/content-moderators-pay-a-psychological-toll-to-keep-social-media-clean-we-should-be-helping-them. Accessed: 2023-01-04.

Shevlin, H. 2024. All Too Human? Identifying and Mitigating Ethical Risks of Social AI. https://philarchive.org/rec/SHEATH-4. Accessed: 2024-04-15.

Singh-Kurtz, S. 2023. The Man of Your Dreams.

Skjuve, M.; Følstad, A.; Fostervold, K. I.; and Brandtzaeg, P. B. 2022. A longitudinal study of human–chatbot relationships. *International Journal of Human-Computer Studies*, 168: 102903.

Tronto, J. C. 2020. *Moral Boundaries: A Political Argument for an Ethic of Care*. Routledge, 1 edition. ISBN 9781003070672.

Tronto, J. C.; and Fisher, B. 1990. Toward a Feminist Theory of Caring. In Abel, E.; and Nelson, M., eds., *Circles of Care*, 36–54. Albany, NY: SUNY Press.

Turkle, S. 2007. Authenticity in the Age of Digital Companions. *Interaction Studies. Social Behaviour and Communication in Biological and Artificial Systemsinteraction Studies / Social Behaviour and Communication in Biological and Artificial Systemsinteraction Studies*, 8(3): 501–517.

Vallor, S. 2016. *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting*. Oxford University Press. ISBN 9780190498511.

Vallor, S.; and Vierkant, T. 2024. Find the Gap: AI, Responsible Agency and Vulnerability. *Minds and Machines*, 34(3): 20.

Varelius, J. 2008. On the prospects of collective informed consent. *Journal of applied philosophy*, 25(1): 35–44.

Walker, L. 2023. Belgian man dies by suicide following exchanges with chatbot. *The Brussels Times*.

Weidinger, L.; Mellor, J.; Pegueroles, B. G.; Marchal, N.; Kumar, R.; Lum, K.; Akbulut, C.; Diaz, M.; Bergman, S.; Rodriguez, M.; et al. 2024. STAR: SocioTechnical Approach to Red Teaming Language Models. *arXiv preprint arXiv:2406.11757*.

Weidinger, L.; Rauh, M.; Marchal, N.; Manzini, A.; Hendricks, L. A.; Mateos-Garcia, J.; Bergman, S.; Kay, J.; Griffin, C.; Bariach, B.; Gabriel, I.; Rieser, V.; and Isaac, W. 2023. Sociotechnical Safety Evaluation of Generative AI Systems. arXiv:2310.11986.

Willison, S. 2023. Bing: "I will not harm you unless you harm me first".

Wood, A. W. 2014. 17Coercion, Manipulation, Exploitation. In *Manipulation: Theory and Practice*. Oxford University Press. ISBN 9780199338207.

Xiang, C. 2023. 'He Would Still Be Here': Man Dies by Suicide After Talking with AI Chatbot, Widow Says.

Xu, S.; Chen, G.; Guo, Y.-X.; Yang, J.; Li, C.; Zang, Z.; Zhang, Y.; Tong, X.; and Guo, B. 2024. VASA-1: Lifelike Audio-Driven Talking Faces Generated in Real Time. arXiv:2404.10667.

Young, H. P. 2015. The evolution of social norms. *economics*, 7(1): 359–387.

Zimmerman, A.; Janhonen, J.; and Beer, E. 2023. Human/AI relationships: challenges, downsides, and impacts on human/human relationships. *AI and Ethics*, 1–13.