Topoi (2024) 43:757–769
https://doi.org/10.1007/s11245-023-09995-2

The Affective Scaffolding of Grief in the Digital Age: The Case of Deathbots

Regina E. Fabry[1] · **Mark Alfano**[1]

Accepted: 7 December 2023 / Published online: 8 January 2024

Abstract
Contemporary and emerging chatbots can be fine-tuned to imitate the style, tenor, and knowledge of a corpus, including the corpus of a particular individual. This makes it possible to build chatbots that imitate people who are no longer alive — deathbots. Such deathbots can be used in many ways, but one prominent way is to facilitate the process of grieving. In this paper, we present a framework that helps make sense of this process. In particular, we argue that deathbots can serve as affective scaffolds, modulating and shaping the emotions of the bereaved. We contextualize this affective scaffolding by comparing it to earlier technologies that have also been used to scaffold the process of grieving, arguing that deathbots offer some interesting novelties that may transform the continuing bonds of intimacy that the bereaved can have with the dead. We conclude with some ethical reflections on the promises and perils of this new technology.

Keywords Grief · Chatbot · GPT · Situated affectivity · Affective scaffolding

1 Introduction

Digital technologies are often designed to engage users' affects, which we here construe broadly to include emotions, moods, and related states (Stephan 2012). Recently, digital apps designed to help people manage negatively valenced emotions such as anxiety and loneliness have become big business. Replika started with humble beginnings but has raised at least US$ 11 million. The estimated market value of Happify (2022), whose mission statement says "we're here to empower you to take control of your emotional life," is US$ 500 million (Crunchbase 2023).

At the time of writing, digital technologies that are designed or used to deal with bereavement, especially the recent loss of loved ones such as romantic partners, friends, parents, and children, have gained momentum (Stokes 2021). One prominent example is the "digital graveyard," which people can visit to reminisce amidst someone's "digital remains" (Ambrosino 2016; Stokes 2015). More recent technologies are not just peaceful digital environments in which to reminisce alone or with other mourners. They use generative artificial intelligence (AI) models — called large language models (LLMs) — to implement chatbots that flexibly respond, in a path-dependent way given their large context windows, to textual inputs with customised text and can be trained to imitate the deceased. In light of the pace at which LLMs are improved and monetised, a better understanding of the affective roles of so-called *deathbots* is needed.[1]

While recent philosophical research has made progress in exploring the descriptive and normative implications of deathbots (Buben 2015; Krueger and Osler 2019, 2022; Lindemann 2022; Stokes 2021), more work needs to be done to understand what is at stake in this particular kind of human-technology interaction. The purpose of this paper is therefore to offer a new descriptive account of the affective dimension of human-deathbot interactions and to develop

Regina E. Fabry and Mark Alfano contributed equally to the paper.

✉ Regina E. Fabry
regina.fabry@mq.edu.au

Mark Alfano
mark.alfano@gmail.com

[1] Department of Philosophy, Macquarie University, Sydney, NSW 2109, Australia

[1] In this paper, we address contemporary deathbots, as well as ones that can be envisioned in the near future (within the next three years). We do not speculate beyond the next few years, however, as developments in this space are too fast and unpredictable to allow for meaningful speculation.

a systematic assessment of their ethical implications. This will be achieved by bringing together recent research on affective scaffolding, grief, and the ethical implications of AI.

Here is the plan for this paper. In Sect. 2, we review relevant aspects of recent philosophical research on grief and affective scaffolding. The purpose of this section is not to offer a systematic review and discussion of the current state of research, but to summarise and contextualise those aspects that can help us capture the affective roles that can be ascribed to deathbots. In Sect. 3, we turn to deathbots such as Replika. How do these technologies work, and what are their capabilities and limitations? We argue that it is illuminating to construe human-deathbot interactions as a case of affective scaffolding. These interactions may, at least under certain conditions, foster continuing habits of intimacy (Krueger and Osler 2022). Then in Sect. 4, we explore the ethical considerations that should motivate and constrain the development and deployment of deathbots. These resources have the potential to promote flourishing, but they are also potentially dangerous and may harm or violate the moral rights of both the deceased and the bereaved.

2 Grief and Affective Scaffolding

In this section, we introduce and review recent work in philosophy of mind that can help us describe and assess the roles of deathbots for processes of grief. In Sect. 2.1, we summarise and contextualise key aspects of emerging philosophical research on grief with a focus on Matthew Ratcliffe's (2017, 2023) phenomenological account of grief experiences. In Sect. 2.2, we consider accounts of affective scaffolding.

2.1 Theoretical Considerations on Grief

Grief is a ubiquitous human experience that is shared within and across cultures and historical times. We here conceptualise it as an emotional experience of irreversible loss of a significant person in combination with a profound disturbance of life possibilities (Ratcliffe 2017, 2023). This disturbance can take various forms across time, including but not limited to "[l]ocalized and nonlocalized experiences of tension, conflict, negotiation, lack, absence, unreality, and being cut off from a shared world" (Ratcliffe 2023, p. 8). These experiences are associated with the challenge to cope with the *indeterminacy* that follows from the loss of a significant person (see Ratcliffe 2023, Chap. 4; Ratcliffe and Byrne 2022). The deceased person had been implicated in a wide range of projects, commitments, habits, activities, and social interactions. In a situation of profound loss, the bereaved person has to navigate, negotiate, and restructure a lifeworld that has been substantially and irreversibly altered. In addition, the bereaved person has to modify their relationship to the person who has died.

There are different accounts of the relation that a bereaved person usually establishes to the deceased over time. According to the relinquishment account, which is often attributed to Freud (1917), the goal state of grieving is to relinquish one's emotional connection to the deceased (for a discussion, see Higgins 2013). According to the continuing bonds account, which is endorsed by most contemporary philosophers, the bereaved person continues their affective connection to the deceased (Klass et al. 1996). This connection, however, has to be adapted to the changed circumstances and is therefore usually reconfigured and transformed. In the case of loving relationships, this continuation of a bond comes with a theoretical problem. As pointed out by Millar and Lopez-Cantero (2022), loving relationships presuppose reciprocity, which is no longer possible in the aftermath of someone's death. The solution to this problem, they propose, consists in acknowledging that at least some components of the loving relationship, which do not depend on reciprocity, can be continued. Specifically, the love that a bereaved person experiences becomes a personal, unreciprocated form of emotional connectedness that is directed at the characteristics and virtues of the deceased person. Furthermore, the bereaved can come to feel or believe that the deceased continues to shape their interests, concerns, and commitments, for example by engaging in various forms of *imaginal engagement*, as Cholbi calls it — "interactions in which we envision and engage with the deceased dialogically or conversationally" (Cholbi 2021, p. 89; see also Norlock 2017; Ratcliffe 2023, Chap. 6). In this sense, Millar and Lopez-Cantero (2022) argue, salient personal aspects of the deceased can continue to influence the projects, commitments, and concerns of the bereaved person. However, many other aspects of the loving relationship are irrevocably lost. The experience of love associated with this particular person will not be characterised by reciprocity and mutuality ever again. The upshot is that grief is characterised by complex and variable configurations of loss and continuation, of adaptation and retention. On a theoretical level, then, it is reasonable to give up on the idea that a strong dichotomy of relinquishing and continuing a bond with the deceased can be maintained (Millar and Lopez-Cantero 2022; Ratcliffe 2023).

The navigation and negotiation of life possibilities in response to the irreversible loss of a significant person poses practical, social, and emotional challenges (Markovic 2022). One of these challenges concerns the regulation of one's emotions in times of disturbance and upheaval. We here understand emotion regulation as an active process

of influencing the quality, valence, intensity, duration, and expression of one's emotional experiences across time (Gross 1999, 2001). This process crucially depends on the agent's interaction with environmental resources (e.g., other people, artefacts, objects, or technological devices). In cases of grief, bereaved agents often face what Ratcliffe calls a *double disorientation*: "Where the loss of a person is at the same time the loss of a resource for coping with loss, it is *doubly disorienting*. One is not only lost in the middle of a forest without any visible paths to follow; one is lost in the absence of a potential guide" (Ratcliffe 2023, p. 171; italics in original). Accordingly, the regulation of the all-encompassing, temporally extended emotion of grief poses particular challenges. Theoretically, this highlights the relevance of considering the dynamical interaction between bereaved persons and their environment in navigating, negotiating, and restructuring life possibilities in the face of irreversible loss. One way of exploring this interaction, as Ratcliffe (2023) proposes, consists in adopting an affective scaffolding approach. In what follows, we will suggest that research on affective scaffolding can help us better understand the role of environmental resources, including digital technologies such as deathbots, for actively influencing and regulating processes of grief.

2.2 Affective Scaffolding

Originating in attempts to understand the acquisition of cognitive abilities in developmental psychology (Wood et al. 1976), the notion of 'scaffolding' has received considerable attention in philosophy of mind. In contrast to internalist assumptions about the brain-bound realisation of mental phenomena (e.g., Adams and Aizawa 2001; Fodor 1980), proponents of scaffolding accounts hold that mental processes are often causally influenced by the agent's embodied interaction with environmental resources (Varga 2019). This causal influence, it has been argued, can be identified across cognitive and affective domains.[2] To date, philosophical research has proceeded by investigating key aspects of cognitive scaffolding (e.g., Clark 1997; Sterelny 2010) and affective scaffolding (e.g., Colombetti and Krueger 2015; Coninx and Stephan 2021; Krueger 2020; Maiese 2016; Saarinen 2020). While we recognize that it is difficult, if not impossible, to establish a clear-cut distinction between cognitive and affective processes, we proceed under the assumption that it is possible to foreground either cognitive or affective aspects of mental target phenomena for analytic purposes (Fabry 2021). In what follows, we review key aspects of affective scaffolding, which are relevant for capturing the influence of deathbots on the emotional process of grief.

According to Coninx and Stephan's (2021) working definition, "we may speak of affective scaffolding when aspects of the environment are used or structured to enable, support, enhance, or regulate the affective experiences of oneself or others" (p. 44). Across all these different configurations, the notion of 'scaffolding' refers to a causal relation holding between an affective experience (the scaffolded) and an environmental resource (the scaffold). As Coninx and Stephan (2021) point out, the class of environmental resources that can enter scaffolding relations is deliberately wide and includes, but is not limited to, other agents, tokens of representational systems (e.g., written text), and technical devices (e.g., computers, smartphones). Similarly, the notion of 'affective scaffolding' implies that various affective phenomena (e.g., emotions, moods, feelings, sensations) can be causally influenced, under certain conditions, by environmental resources. However, most research has focused on the regulation of emotional experiences (Colombetti and Krueger 2015; Krueger 2020; Krueger and Osler 2019; Saarinen 2020). With the advent of digital technologies, Krueger and Osler (2019) argue, social media, chatbots, and other online resources can scaffold, in various ways, emotional experiences. Across these cases, it is important to introduce distinctions concerning the temporal resolution and causal configuration of scaffolding relations. First, scaffolding relations can be identified on different temporal scales. Following Sutton (2016), we can distinguish phylogenetic, cultural-historical, ontogenetic, and occurrent timescales. Second, the causal relation holding between the scaffold and the scaffolded can be uni-directional or bi-directional (Coninx and Stephan 2021; Krueger 2020; Saarinen 2020). In the former case, an environmental resource influences, but is not influenced by, the agent's affective experience. In the latter case, the environmental resource, or at least some of its components, and the agent's affective experience mutually influence each other.

Affective scaffolding relations can be further specified along several dimensions. Coninx and Stephan (2021), elaborating Saarinen's (2020) systematisation, distinguish

[2] This claim about the realisation base of many mental phenomena has a strong methodological implication. Specifically, proponents of philosophical accounts of scaffolding are committed to the view that many mental phenomena, ranging from emotion regulation (Colombetti and Krueger 2015) to narrative practices (Fabry 2021), can only be explained if the causal dependence relations holding between an agent's embodied processes and environmental resources are taken into account. For this reason, scaffolded mind theorists reject the internalist view that the brain is and should be the only relevant unit of analysis for explanations of mental phenomena (for details, see Varga 2019). Scaffolding accounts are thus inconsistent with weak embodied and embedded accounts (Goldman 2012; Rupert 2004), according to which extra-cranial bodily and environmental states and processes are only causally relevant insofar as they enable physiological, somatosensory, and kinaesthetic states and processes that are represented in the brain.

among trust, robustness, mineness, individualisation, incorporation, awareness, intent, and control. Trust refers to the reliability that an agent ascribes to an environmental resource in terms of its accessibility, the provided informational value, or its systematic effect on the agent's emotional experience (Colombetti and Krueger 2015; see also Krueger and Osler 2019). Robustness captures the degree to which a resource is integrated into one's affective experience. Mineness is a graded notion that describes the integration of an environmental resource into one's practical identity, which is defined as a relational collection of self-defining values, concerns, and commitments (Korsgaard 1996). Individualisation refers to the gradual adaptation of an environmental resource to the agent. Incorporation captures the degree to which an environmental resource becomes a proper part of the agent's affective experience. The notions of 'awareness', 'intent', and 'control' refer to the scaffolding relation as a whole: how aware is the agent of the scaffolding relation? To what extent has the agent formed the intention to enter a certain scaffolding relation and has control over its unfolding? In what follows, we assume that these dimensions can help us specify the causal relationship between deathbots (and other technological resources) and an agent's grief experiences on an occurrent timescale. This assumption is informed by Ratcliffe's (2023, Chap. 7) suggestion that emotion regulation in the context of grief often depends on scaffolding by environmental resources.

Before proceeding, we note that our understanding of affective scaffolding, in contrast to most existing approaches, seeks to avoid the *harmony bias* that is prevalent in the vast majority of research on situated cognition and affectivity: an emphasis on the beneficial effects and a neglect of (potentially) harmful consequences of agent-technology interactions (Aagaard 2021; Bruineberg and Fabry 2022; Timms and Spurrett 2023). What is needed, we argue, is a context-sensitive description and nuanced understanding of the ways in which digital technologies in general and deathbots in particular can influence and regulate emotional experiences of grief, for better and for worse.

3 Deathbots as Affective Scaffolds

3.1 What are Deathbots?

Let us turn to the deathbots of today and perhaps tomorrow (though not the more distant future, which is too difficult to predict). Generative AI systems based on transformer models have recently entered public awareness. Currently, the most popular are LLMs, especially GPT-n (Generative Pre-trained Transformer n), which are able to generate textual content based on input prompts, and related systems such as ChatGPT, which functions as a chatbot, and Bing Chat, which can also use Microsoft Bing's search engine and the DALL-E 3 image generator to enhance its functionality.

These generative systems are based on transformer models: neural networks that learn context and, according to some enthusiasts, even "meaning by tracking relationships in sequential data" (Merritt 2022; see also Vaswani et al. 2017). These systems depend on large data models — akin to their "vocabulary" — which have been trained on collections of words and the relations between them, over many iterations. The end-user merely has to provide a textual prompt as input to the system, which then uses its model to generate and rank candidate corresponding outputs.

GPT-n is the state-of-the-art technology available to retail customers. This resource is easy to implement and can generate text with virtually no cost or restriction for an individual user – though of course it does require vast compute, energy, and water resources (see Bender et al. 2021). To better understand how generative AIs work, we offer a bird's-eye-view of the technology here. This understanding is needed for describing and assessing the role of deathbots, which are based on this technology.

First, an LLM is trained on a large collection of textual corpora scraped from the internet. For GPT-3, a 45 TB dataset of text (approximately 409B tokens) from multiple sources was used. Text was drawn from sources indexed by Microsoft's Bing search engine. To train the model, these corpora are passed through an encoder, which enables the model to treat text as an ordered sequence of words (actually tokens, which can also include punctuation, emoji, and other characters) (Brown et al. 2020; Devlin et al. 2018).[3] Tokens are then processed by a decoder, an auto-regressive model that's designed to predict the next token. This final step is used to predict, iteratively, the next token given the preceding sequence of tokens. The model initially assigns random probabilities, but it is trained up over many iterations of feedback to be increasingly accurate in its predictions.

When the user presents the trained model with a prompt, the encoder works on the text as above. Mirroring the training of the model, the text features of the input are used to predict what textual features are likely to be associated with them in the next token. These predictions are then decoded into a sequence of tokens that are outputted to the user.

The models based on this technology are constructed from a large assemblage of human input. A text generation system learns from a large collection of input corpora to infer syntactic and (arguably) semantic properties, some of which may be human-interpretable. For example, inferred properties may have to do with quantity (singular vs. plural)

[3] In earlier models, only the preceding text was used to make predictions. Allowing the model to look both forward and backward significantly enhances performance.

in order to handle subject-verb agreement, but may also be more conceptual (e.g., associating 'apple' more with 'orange' than with 'cucumber' or 'gun' and thus representing something like *fruitiness*). These connections are represented as a vast series of correlations: each token is encoded as a vector of which each entry measures the extent to which it is likely to co-occur with each other token, taking into consideration its associated linguistic context and distribution within a unit of text (Lavelli et al. 2004).

One of the main innovations driving the current generations of transformer models is their ability to do something like self-supervised training. Transformer models use mere token ordering and algorithmic "attention" in parallel processing both to come up with the parameters and to assign weights, using the vast quantity of textual data available from the web to train and retrain the model instead of relying on bespoke human input.

Chatbots based on transformer models enter a dialogue with a human agent who sends prompts and receives responses in a path-dependent, unfolding process. The model's "personality" can be set by a "system," which establishes a role that the chatbot then adopts until prompted to switch to a different role. In addition, it not only has a generic linguistic representation based on the training described above but also "remembers" both the previous messages sent by the user and its own responses to them. This makes it possible to have full-blown conversations with contemporary chatbots, which often seem capable of resolving anaphora and "remembering" past elements of the conversation. Since human conversation depends on conversational score-keeping (Lewis 1979; Witek 2015), this is crucial to the compellingness of the conversation. In just a few months, the context window for LLMs — the number of tokens from the previous conversation they can take into account when producing a response — has increased rapidly. The context window for GPT-3 was 2048 tokens, roughly the length of a short essay. The context window of GPT-4 is about 32,000 tokens, roughly the length of three full journal articles.

A deathbot is fine-tuned to imitate the vocabulary, style, personality traits, and even memories of a particular person. GPT-3 is trained on text produced by billions of individuals and thus does not have an individual "personality." ChatGPT and related resources can be prompted to adopt an individual style of expression, and when the individual in question has a sufficiently large digital footprint, they are capable of producing impressive results. For most private individuals, ChatGPT would not produce anything resembling a convincing imitation. However, if the bereaved has a sufficiently large collection of correspondence with the deceased, this data can be used to fine-tune the model. The result will be a chatbot that responds, flexibly and dynamically, with utterances that can be uncannily similar to how the deceased would have responded to the same text.

Importantly, a deathbot trained in this way will not imitate the deceased tout court. To do that, it would have to be fine-tuned using correspondence not only with a bereaved person but all of the deceased's past interlocutors. Since we are interested in deathbots designed to assist in the process of grieving, we disregard such multi-track models. If your friend's deathbot responded to you not in the way that the deceased would have responded to *you* but in the way they would have responded to their boss or their mother, that would presumably not be desirable. Thus, we restrict our discussion to bilateral or dyadic conversational deathbots.

Time also matters here: a chatbot fine-tuned only on the last six months of someone's correspondence will behave differently from one fine-tuned on the last ten years of their correspondence. People may use deathbots in different ways. In some cases, they may want to resolve unfinished business with the most recent timeslice of the deceased. For instance, especially in the case of sudden, unexpected death, there may be unresolved conflicts, unexpressed confessions, and other words that the bereaved wishes they had spoken to the deceased. In other cases, they may want to reminisce about years or even decades of shared experience. The curation of the fine-tuning corpus depends, in part, on what the bereaved wants or needs from the conversational exchange.

We also note that, while digitised text is the primary use case, it is already possible to create and deploy multimodal deathbots. In the most straightforward interaction, the bereaved sends text to the digital avatar of the deceased, which responds via text. It is also possible to *speak* rather than *write* to the chatbot, with the audio being automatically transcribed to text before being passed to the model. And, if there is enough training data available, it would also be possible to have the model return not text but audio responses that imitate the pitch, timbre, and other qualities of the deceased's voice.

Technological limitations faced by deathbots are familiar from previous work on natural language generation and processing. Notoriously, models of this size and complexity are "black boxes," in the sense that it is difficult or even impossible to explain in human-intelligible ways why they produced any particular output (Rudin and Radin 2019). While work in explainable AI has been making significant strides, explanations tend to be at the level of general patterns rather than particulars (Xu et al. 2019).

Another technological limitation relates to sufficiently well-documented natural languages in which chatbots can produce plausible responses. Large languages spoken in rich countries, such as English, German, Arabic, Spanish, and Chinese, are the best documented. Small languages such as Lithuanian are not documented nearly as well (Mi

et al. 2022). Purely oral languages and languages spoken by Indigenous people where there exist taboos on who is allowed to know and say which words to whom present even thornier problems. The market for deathbots in these languages would also be much smaller than the market for, say, anglophone deathbots. These technological and social facts combine to make it unlikely that deathbots could be developed and implemented in a way that does not exacerbate global inequalities. Thus, even before we turn to a direct confrontation with the ethics of deathbots, it should be clear that the technological limitations of this emerging technology already make their unregulated, large-scale deployment fraught.

3.2 Human-Deathbot Interaction: A Case of Affective Scaffolding

Against this background, we propose that human-deathbot interactions can be usefully described as a case of affective scaffolding. In the philosophical literature on situated affectivity, Krueger and Osler (2019) were the first to mention that deathbots could be considered as affective scaffolds without providing any details or specifications. They point out that "providing dynamic, ongoing interactions with chatbots offers a novel form of engineering the affective contours of our grief processes" (Krueger and Osler 2019, p. 223). We would like to suggest, however, that human-deathbot interactions are not entirely novel phenomena. Rather, they are best understood as emerging innovations that recombine and reconfigure technologically mediated and socio-culturally enabled patterns (Fabry 2017): conversational chatbots and various forms of imaginal engagement with the dead (Jiménez-Alonso & Brescó de Luna, 2023). First, deathbots resemble psychotherapy chatbots, ranging from ELIZA, a chatbot developed in the 1960s (Weizenbaum 1966), to more recent applications, for example Woebot (Tekin 2021). These psychotherapy chatbots are designed to offer a technologically mediated conversational exchange with distressed agents who are in need of advice, support, or validation. Deathbots also resemble so-called social chatbots such as Replika, which are designed to elicit emotional investments and a sense of trust and belonging in their users (Laestadius et al. 2022; Skjuve et al. 2021; Weber-Guskar 2022). Ever since the emergence of ELIZA, researchers have noted that many human agents show the tendency to attribute human-like characteristics to artificial linguistic behaviour (Mitchell and Krakauer 2023). This has become known as the *ELIZA effect* (Hofstadter 1995). As we will see, the key principles of conversational chatbots, and the effects they can have on their users, matter for our understanding of the affective possibilities and limitations of deathbots.

Second, human-deathbot interactions can be understood as recombinations and reconfigurations of historically earlier technologies, rituals, and practices for connecting and communicating with the deceased (Jiménez-Alonso & Brescó de Luna, 2023; Walter 2015). In the third century BC, for example, the Confucian philosopher Xunzi described a mourning ritual of the impersonation of the deceased (see Elder 2020). As another example, consider the innovation of the telegraph in the United States in the mid-19th century. This innovation inspired attempts to establish new forms of communicating with the dead, leading to new forms of technologically mediated presence of the irrevocably absent (Walter 2015).

Furthermore, as Cholbi (2021) notes, human-deathbot interactions can be understood as technologically mediated forms of imaginal engagements with the deceased (see Sect. 2.1 above). He argues that deathbots "offer genuine opportunities to engage the deceased at an imaginal level and establish continuing bonds between them" (Cholbi 2021, p. 90). The upshot is that the innovation of deathbots can be understood as a recombination and reconfiguration of historically earlier technologies and practices for reconnecting and communicating with the dead. This perspective on the innovativeness of deathbots, rather than a simple appeal to their novelty, can help us better understand their (putative) contribution to grief experiences from the perspective of affective scaffolding.

In Sect. 2.1, we pointed out that grief experiences are often characterised by a double disorientation: the very resources that would usually contribute to emotion regulation are no longer available in the aftermath of the irreversible loss of a significant person. For this reason, alternative environmental resources might play a particularly important role for bereaved agents in their attempts to navigate, negotiate, and restructure their lifeworld. In what follows, we will apply the account of affective scaffolding outlined in Sect. 2.2 to describe how deathbots might contribute to the regulation and negotiation of grief.

Let us start by specifying the temporal scale and the kind of causality at play in human-deathbot relations. First, the interaction of bereaved agents with a deathbot usually unfolds on an occurrent timescale. While deathbots might be recruited on various occasions over an extended period of time, it seems unlikely that they can directly shape the ontogenetic development of emotional experiences. Second, given the current technological specifications of deathbots described in Sect. 3.1, the causal relationship established between the emotional experience of a bereaved human agent and a deathbot can be described as bi-directional. The outputs generated by a deathbot influence the emotional experience of the agent. In turn, the deathbot generates certain conversational patterns that directly depend on the

emotionally shaped prompts it is receiving from the human agent. Accordingly, the human prompts and the chatbot's outputs mutually influence each other (Jiménez-Alonso & Brescó de Luna, 2023).

As mentioned in Sect. 2.2, the role of affective scaffolds can be further specified along several dimensions. First, consider again the dimension of *trust*. This notion is ambiguous in that it can refer to the accessibility, the informational reliability, or the systematic affective effectiveness of the environmental resource. Under the assumption that deathbots can be accessed on all digital computing devices with an interface, bereaved human agents might trust, under certain circumstances, that they are readily available and accessible. The informational value of deathbots, however, is reliable to varying degrees. It is to be expected that deathbots, like other chatbots, can generate responses that are inaccurate, inappropriate, or even offensive (see Laestadius et al. 2022; Mitchell and Krakauer 2023). A deathbot would thus only be deemed trustworthy to the extent that these problems can be avoided or minimised. The extent to which an agent trusts that an environmental resource has a highly predictable and systematic affective effect is a function of the frequency of recruitment (Colombetti and Krueger 2015; see also Krueger and Osler 2019). The more frequently a deathbot is recruited to help navigate and negotiate grief experiences, the more trusted it can become to contribute to the regulation of one's grief experiences. This is closely related to, yet distinct from the second dimension of affective scaffolding: *robustness*. The robustness of an affective scaffold depends on the regularity with which it is integrated into an agent's emotional experience (Coninx and Stephan 2021). It thus becomes possible to distinguish between different degrees of robustness relative to the regularity of deathbot recruitment. Third, the interaction with an environmental resource can, to varying degrees, become an attribute of an agent's practical identity. It is assumed that bereavement is such a disruptive experience in part because the deceased person has been part of the griever's practical identity (Cholbi 2021, Chap. 1; Ratcliffe 2023, Chap. 2). In this sense, certain relational characteristics or attributes of the deceased have been experienced as *mine*. Similarly, certain aspects of a deathbot that are perceived as relational can be captured by the graded notion of *mineness*. Fourth, environmental resources can be more or less *individualised*. That is, they can be adapted, to varying degrees, to the agent's pattern of affective experiences. In the case of human-deathbot interactions, it is reasonable to assume that the outputs generated by the deathbot can lead to the impression that they are tailored or adapted to the affective profile of the agent. Fifth, depending on other factors, including the degree of robustness and mineness, an environmental resource can become gradually incorporated into the agent's overall affective experience. If a deathbot is readily accessible, reliably recruited, and becomes a proper part of the agent's relational practical identity, it becomes an incorporated part of the agent's abilities to navigate and regulate their grief experiences. These dimensions can help characterise the role of a deathbot (the scaffold) for the grief experiences (the scaffolded) of a bereaved agent.

In addition, the occurrent, reciprocal causal relationship between an agent's emotional experience and a deathbot – the scaffolding relation as a whole – can be specified along three dimensions: awareness, intent, and control. It is mostly an empirical question to what extent an agent can be aware of, intending to enter and maintain, and control their interactions with a deathbot. How aware is a grieving agent that their emotional experiences are regulated by a chatbot? Are the emotional effects of the interaction with a deathbot intended, at least to a substantial degree? Is the agent fully in control of the reciprocal causal relationship that unfolds between their emotional experiences and the outputs of the deathbot? As we will argue in Sect. 4 below, the answers to these questions have strong normative implications. For current purposes, however, it should suffice to note that these dimensions are important for specifying human-deathbot interactions.

3.3 Continuing Habits of Intimacy?

In light of our considerations on the guiding principles of deathbots (Sect. 3.1) and the affective scaffolding relationships that can be established between an agent's grief experiences and a deathbot (Sect. 3.2), the question arises what contributions human-deathbot interactions can make to the unfolding of grief experiences. In this subsection, we discuss one important recent proposal for thinking about the effects that deathbots might have on the grief experiences of bereaved agents and relate it to our affective scaffolding account of human-deathbot interactions. Krueger and Osler (2022) have argued that deathbots enable the grieving agent to continue habits of intimacy that they shared with the deceased. On their view, habits of intimacy, for example conversational practices, emotion regulation, and sharing time, establish emotional patterns that connect agents to each other and to a shared experiential world of possibilities. First, they develop the assumption that the interaction with a deathbot can contribute to the continuation of the shared habit of engaging in everyday conversations. This is rendered possible to the extent that the deathbot's outputs have the effect of creating a sense of familiarity for the grieving agent. Second, they develop the idea that deathbots can contribute to emotion regulation. Recall from Sect. 2.1 that emotion regulation in grief is often characterised by a double disorientation (Ratcliffe 2023). Building on this

observation, Krueger and Osler (2022) suggest that deathbots can help continue the habit of emotion regulation that the bereaved agent used to share with the deceased. On their view, the reason is that the outputs of deathbots can help regulate emotional experiences by eliciting laughter or mirth or creating the impression that advice and support is offered in a way that feels familiar and therefore comforting. Finally, they point out, interactions with deathbots offer opportunities to continue or re-establish a habit of *shared time*. Taken together, Krueger and Osler (2022) develop the view that human-deathbot interactions can, at least in certain cases and for a constrained period of time, be helpful for navigating and negotiating grief experiences and for continuing a bond with the deceased. Against the background of our theoretical considerations on grief and our novel affective scaffolding account of human-deathbot interactions, we are now in a position to critically examine this view.

There are at least two problems with Krueger and Osler's (2022) proposal that deathbots can help continue habits of intimacy. First, both grieving and human-deathbot interaction are more complex and variable phenomena than Krueger and Osler (2022) might lead us to assume. Whether and to what extent deathbots can helpfully scaffold the unfolding of emotional grief experiences depends, amongst other things, on the cause of death of the significant person, the shape and scope of the relationship between the bereaved and the person who has been lost, and the wider culturally shaped practices and norms that guide and constrain the grieving process (see Fabry 2023). Without taking the variability of grief experiences into account, it is difficult to arrive at an adequate assessment of the possibilities and limitations of reconstructing habits of intimacy by using a deathbot as an affective scaffold. Similarly, it is reasonable to assume that the extent to which human-deathbot interactions can be conducive to the grieving process depends on the attitudes of the bereaved towards the conversational possibilities and limitations of the deathbot, even under the assumption that the deathbot is based on a sufficiently rich corpus. Krueger and Osler (2022) speculate that bereaved agents are fully aware that deathbots lack mental states and processes and experiential access to a lifeworld. Furthermore, they indicate that bereaved agents might adopt a fictionalist stance towards deathbots and their outputs. However, it is an open empirical question whether this is actually the case. Recent research on people's attitudes towards chatbots in non-bereavement cases and the emotional costs and benefits of human-chatbot interactions paint a complicated picture (e.g., Brandtzaeg et al. 2022; Christoforakos et al. 2021; Laestadius et al. 2022; Skjuve et al. 2022). One implication of this research is that the attitudes of human agents towards the conversational behaviour of deathbots is more variable and context-dependent than suggested by Krueger and Osler (2022). Our affective scaffolding account has the potential to help specify criteria for intra- and inter-individual differences in human-deathbot interactions within and across contexts. Towards the end of the previous subsection, we noted that scaffolding relations as a whole can be specified along three dimensions: awareness, intent, and control. How agents relate to deathbots, we hypothesise, will depend, at least in part, on the circumstances of the agent's bereavement, their relationship to the deceased, and their general attitudes towards deathbots and other digital technologies. These causal factors, in turn, can be identified and systematised by exploring the following questions: Are agents aware, and if so to what extent, that they are interacting with a deathbot (rather than with the deceased person themself)? Do they intend, and if so to what extent, to enter or maintain a communicative relationship with a deathbot? Do agents have control, and if so to what extent, over the unfolding of their relationship with a deathbot? It is a clear advantage of our affective scaffolding account that these questions for future research come into view.

Second, the proposal that deathbots might contribute to a continuation of habits of intimacy seems to be influenced, at least in part, by a harmony bias (Aagaard 2021; Bruineberg and Fabry 2022). This bias consists in a tendency to focus on human-technology interactions that can be characterised as cases of cooperation and collaboration, rather than conflict and interference (Timms and Spurrett 2023). While Krueger and Osler (2022) discuss some of the ethical concerns about bereaved-deathbot interactions, they seem to be committed to a positively biased understanding of digitally mediated grieving. However, what is needed is a nuanced, empirically informed understanding of the impact of deathbots on grief experiences in a variety of cases, which is not susceptible to any biases towards technological optimism – or pessimism, for that matter (see Bruineberg and Fabry 2022). The emerging affective scaffolding account can help develop a context-sensitive description and a nuanced understanding of the advantages and disadvantages of the recruitment of deathbots for the emotional regulation of grief experiences. Each case of human-deathbot interaction can be captured in the dimensional space we have outlined above. The degree of trust, robustness, mineness, individualisation, and incorporation, we hypothesise, would stand in a systematic relationship to the normative evaluation of the influence of deathbots on the emotional experiences of the bereaved agent. Furthermore, the extent to which the scaffolding relationship between an agent's grief experiences and a deathbot can be characterised through the gradual characteristics of awareness, intent, and control will have an impact on the degree of autonomy we can ascribe to the bereaved agent. In sum, we suggest that it remains a theoretical and empirical question for future research whether

and to what extent deathbots can contribute to a continuation of habits of intimacy between the bereaved and the deceased in certain kinds of cases. This question can be specified, we have argued, if an affective scaffolding perspective on human-deathbot interactions is adopted. At this point, descriptions of real-world cases and theoretical considerations indicate that bereaved-chatbot interactions have important normative implications. We describe and discuss these implications in the next section.

4 The Role of Deathbots for Grieving: A Normative Assessment

The normative implications of deathbots are numerous and diverse. While the ethical implications of chatbots in general have received some attention (e.g., Dennis 2022), deathbots have received less attention (but see Lindemann 2022). To bring some order to our assessment, we consider three types of targets of moral concern whose standing seems most relevant in this context: the deceased, the bereaved, and society as a whole. As we will see, the normative implications of affective scaffolding are most relevant to the bereaved, but they also have matter for the deceased and society as a whole. We do not aim to give a comprehensive account of all normative considerations in this section. Instead, our goal is to map the terrain, identify several landmarks that clearly deserve critical reflection, and offer a first attempt to address the issues that arise.

4.1 Moral Claims of the Deceased

When it comes to the deceased, what seems most relevant are their moral claims on anyone who might create a deathbot using a corpus they produced. On the one hand, consider the fact that, for many individuals and cultures, there is an imperative to remember and be remembered (Blustein 2008). Not making use of someone's digital remains – or even deleting them – might amount to a "second death" (Stokes 2015). Especially in cases where the deceased made an express request for a certain type of deathbot to be built and deployed, and provided the resources to follow through on that request, failing to do so could be seen as akin to disrespecting their last will and testament. Of course, we are not morally bound to respect every request that the dying make, but for those among the bereaved who *do* wish to do their best to fulfil the expectations of dead loved ones, this consideration should carry some weight.

On the other hand, training a deathbot on a corpus of intimate conversations without the prior consent of the deceased might be seen as a disrespectful invasion of "mental privacy" (Clowes et al., forthcoming). Maybe some people don't want to be brought back. And not in a way that is creepily adjacent to the way they would actually express themselves. And especially not in a way that is subject to control by the resurrector (through individualization using either system commands or the large context window), such that their chatbot over time could become an overwritten palimpsest of their expressed personality traits mixed with the preferences and projections of the bereaved. In Haitian folklore, a zombie is a resurrected corpse under control of a *bokor* or *caplata* — a sorcerer or witch (Thomas 2010). What deathbots are capable of doing might be better metaphorized not as resurrection but as zombification, taking ownership of the deceased in a way that expresses excessive mineness. Taking seriously the fact that some people may not want to be zombified, perhaps we need to grant people the right to sign the digital equivalent of a "do not resuscitate" order — a "do not resurrect" order, as it were (for a discussion of this possibility, see Lam 2023). Without affirmative permission to build a certain type of chatbot based on someone's corpus, we enter morally grey territory.

Beyond the issue of consent, we might find it disrespectful to train a deathbot of someone that misrepresents them or does not represent them with adequate richness and vibrancy. The process of fine-tuning a language model requires a large corpus. If someone's private digital footprint is too small, then any deathbot trained on it would be a pale imitation of them. Second, and relatedly, the digital footprint of the deceased may include language that we would not want to use to train the model. In an idyllic case where the relationship between the bereaved and the deceased was positive and healthy, almost nothing would need to be excluded. But what about more complicated cases? Suppose someone wanted to use a deathbot of their sometimes-abusive father to arrive at some kind of forgiveness of him after he died. The training data for such a model would have to be carefully curated to exclude the more hostile and abusive language that he directed at his offspring. This would both shrink the digital footprint, making the resulting model less compelling and trustworthy, and risk reinscribing a merely notional father on the actual character of the deceased: not *him*, but a multiple exposure of him and the way the bereaved wishes he had been. Leaving aside whether this would be good for the bereaved, we might wonder whether it would be even worse than *damnatio memoriae*: not to remember *him*, but to overwrite his individuality with a comforting fabrication.

4.2 Moral and Mental Health Needs of the Bereaved

As we explained above, one of the main uses for deathbots would be to affectively scaffold the process of grieving. This process has both moral and mental health dimensions. Morally, there is the possibility that engaging with a deathbot

could undermine the bereaved's agency and autonomy (Clowes et al., forthcoming; Lindemann 2022). Those who use deathbots, especially the recently-bereaved, are likely to be vulnerable and have an impaired sense of agency – the double disorientation that Ratcliffe (2023) refers to. The bereaved are thus at risk of losing their autonomy as they develop a too-intense dependency on the deathbot. We can call this the loss of autonomy problem, as it consists in the possibility that the bereaved might lose (parts of) their autonomy by relying on the interaction with a chatbot in trying to reconfigure their lifeworld in response to their bereavement experience (Lindemann 2022). They may come to trust the deathbot too much, to develop an overly robust reliance on it, or to incorporate it too deeply into their affective lives. Of course, we can also imagine happier outcomes, where the bereaved is slowly, perhaps partially, weaned off dependency on the deathbot, thereby restoring their autonomy. Given that there is both peril and promise here, we suggest that deathbots should generally only be implemented when adequate automated guardrails have been put in place to detect overdependence on the deathbot. And in the case of recent bereavement, especially in the case of unanticipated death, deathbots should only be implemented under the supervision of a therapist or grief counsellor, who could monitor the autonomy-affecting aspect of the scaffolding relationship.

Another consideration, which relates to our discussion in the previous section, might be called the replacement problem. According to the replacement problem, engagement with chatbots and other death technologies could lead to a replacement of the irreversibly lost relationship with the deceased by a digitally-mediated relationship with an artificial system (Buben 2015; Stokes 2021). The replacement problem has both moral and mental health dimensions. Morally, the bereaved is in danger of falling into inauthenticity, self-deception, or even delusion. If the affective scaffolding enabled by a deathbot warps the continuing bonds of the bereaved to such an extent that they are attached not to the deceased but to a fiction, then those affective bonds lose their authenticity, contra Krueger and Osler's (2022) considerations. Furthermore, to the extent that continuous self-deception and delusion lead to patterns of maladaptive behaviour, they could undermine the mental health of the bereaved.

Given the market imperatives in the deathbot sector, these considerations may not receive the attention and weight they deserve, since implementing them may not increase profits. In addition, we note that the business model of a company offering deathbots may create at least two perverse imperatives. First, many of the prominent players in this space, such as Replika, use a freemium subscription model. Basic accounts are free, but the more engaging ones require a monthly payment. This creates an incentive for Replika to fine-tune their language models in such a way that users are motivated to keep interacting, i.e., to make them maximally robust. It is not clear that such endless interaction is consistent with healthy continuing habits of intimacy, as we discussed them above. If, ultimately, users keep chatting with deathbots because they are addictive and not because these interactions are conducive to the flourishing of the bereaved, this would constitute a form of harmful affective scaffolding.

Second, and perhaps more perniciously, companies like Replika that have a freemium business model will face pressure to monetize the users who have free accounts. We already know how this is typically done in other digital sectors: data brokerages and targeted advertising. While Replika's website (https://replika.com/) currently says that they engage in neither of these practices, it is difficult to be sanguine about this given that similar assurances by Facebook turned out to be false (Dance et al. 2018). It could be emotionally devastating to a user to have their intense, private chats with a deathbot sold via a data brokerage. It could also be upsetting if, in the middle of such an intense private chat, the deathbot suddenly made a product recommendation. We are not insinuating that Replika already engages in such unscrupulous practices, but given the clear and present dangers, we suggest that deathbots should only be allowed if their privacy and data security are regularly audited by independent specialists in both ethics and digital technologies.

4.3 Potential Transformative Effects of Deathbots

Finally, consider the potential transformative effects of deathbots at the societal level. Prior to the availability of free or cheap deathbots, our preparations for the inevitable deaths — whether clearly foreseen or not — of loved ones did not need to take into account what would be needed to implement a compelling deathbot of any given individual. Of course, this does not mean that people never prepared for death before, or that they only did so in technologically unmediated ways. But the specific demands of building a convincing deathbot (e.g., a large corpus of recent interactions) were not relevant.

These reflections suggest that the mere prospect of deathbots may transform how we relate to the living, especially the elderly, the infirm, and those whose work (e.g., firefighting, warfighting) or hobbies (e.g., rock climbing) are extremely dangerous. As we mentioned in the introduction, apps like Replika are already big business. Further growth of the deathbot sector is liable to prompt more and more people to prepare for death by more assiduously collecting corpora from their loved ones: pre-building the scaffolding that their loved ones could use to implement a deathbot. Whether this

will foster social flourishing and better relationships is anyone's guess. Optimistically, it may lead to more conscientious and deliberate planning and deeper, more thoughtful relationship-building. Pessimistically, it could poison relationships among the living by leading them to attend to the digital afterlife rather than savouring their precious, finite lives and relationships. Even more pessimistically, by making death seem less profound and permanent, it could lead some people to neglect their relationships with the living, since they can always digitally resurrect a loved one via an all-too-easily-anthropomorphized deathbot. While this prospect remains speculative, the fact that after the COVID pandemic people have less face-to-face interaction because they can always just Zoom or talk on the phone suggests that such transformative effects cannot be dismissed out of hand (Patulny and Bower 2022).

It is unlikely that the directors and employees of companies like Replika have seriously considered these prospects. Given the potentially profound effects of deathbots, this blindspot calls for ethical review. Such review could be implemented in the form of independent, in-house ethicists or audits by professional or governmental specialists in both philosophy and digital technologies.

5 Concluding Remarks

In this paper, we brought resources from philosophy of mind to bear on a recent technological innovation: deathbots. We argued that these deathbots are fruitfully conceptualised in terms of affective scaffolding, and that they have the potential to shape our continuing bonds with the people we have lost. We also pointed to a range of technological and economic issues with contemporary and near-future chatbots. These issues create an imperative to critically assess deathbots, with attention to the claims and needs of the deceased, the bereaved, and society as a whole. We hope that our contribution lays the groundwork for further work in this emerging area of research, as well as the development of policy guidelines and best practices for individuals and corporations that build and deploy deathbots.

Acknowledgements We are grateful to Marc Cheong for helpful and constructive feedback on an earlier version of this paper.

Funding Fabry's work has been funded by a Discovery Early Career Research Award granted by the Australian Research Council (DE210100115).

Declarations

Conflict of Interest None.

References

Aagaard J (2021) 4E cognition and the dogma of harmony. Philosophical Psychol 34(2):165–181. https://doi.org/10.1080/09515089.2020.1845640

Adams F, Aizawa K (2001) The bounds of cognition. Philosophical Psychol 14(1):43–64. https://doi.org/10.1080/09515080120033571

Ambrosino B (2016), March 14 Facebook is a growing and unstoppable digital graveyard. *BBC*. https://www.bbc.com/future/article/20160313-the-unstoppable-rise-of-the-facebook-dead

Bender EM, Gebru T, McMillan-Major A, Shmitchell S (2021) On the dangers of stochastic parrots: Can language models be too big? 列. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Blustein J (2008) The moral demands of memory. Cambridge University Press

Brandtzaeg PB, Skjuve M, Følstad A (2022) My AI friend: how users of a social chatbot understand their human–AI friendship. Hum Commun Res 48(3):404–429. https://doi.org/10.1093/hcr/hqac008

Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler D, Wu J, Winter C, Amodei D (2020) Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in Neural Information Processing Systems* (Vol. 33, pp. 1877–1901). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

Bruineberg J, Fabry R (2022) Extended mind-wandering. Philos Mind Sci 3. https://doi.org/10.33735/phimisci.2022.9190

Buben A (2015) Technology of the dead: objects of loving remembrance or replaceable resources? Philosophical Papers 44(1):15–37. https://doi.org/10.1080/05568641.2015.1014538

Cholbi M (2021) Grief: a philosophical guide. Princeton University Press

Christoforakos L, Feicht N, Hinkofer S, Löscher A, Schlegl SF, Diefenbach S (2021) Connect with me. Exploring influencing factors in a human-technology relationship based on regular chatbot use. Front Digit Health 3. https://www.frontiersin.org/articles/https://doi.org/10.3389/fdgth.2021.689999

Clark A (1997) Being there: putting brain, body, and world together again. MIT Press

Clowes RW, Smart PR, Heersmink R (forthcoming). The ethics of the extended mind: Mental privacy, manipulation and agency. In B. Beck, O. Friedrich, & J. Henrich (Eds.), *Neuroprosthetics: Ethics of applied situated cognition*. Springer

Colombetti G, Krueger J (2015) Scaffoldings of the affective mind. Philosophical Psychol 28(8):1157–1176. https://doi.org/10.1080/09515089.2014.976334

Coninx S, Stephan A (2021) A taxonomy of environmentally scaffolded affectivity. Dan Yearb Philos 1–27. https://doi.org/10.1163/24689300-bja10019

Crunchbase (2023) https://www.crunchbase.com/organization/happify/company_financials

Dance GJX, LaForgia M, Confessore N (2018) December 18). As Facebook raised a privacy wall, it carved an opening for tech giants. N Y Times. https://www.nytimes.com/2018/12/18/technology/facebook-privacy.html

Dennis MJ (2022) Social robots and digital well-being: how to design future artificial agents. Mind Soc 21(1):37–50. https://doi.org/10.1007/s11299-021-00281-5

Devlin J, Chang M-W, Lee K, Toutanova K (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv Preprint arXiv:1810.04805*

Elder A (2020) Conversation from beyond the grave? A neo-confucian ethics of chatbots of the dead. J Appl Philos 37(1):73–88. https://doi.org/10.1111/japp.12369

Fabry RE (2017) Cognitive innovation, cumulative cultural evolution, and enculturation. J Cognition Cult 17(5):375–395. https://doi.org/10.1163/15685373-12340014

Fabry RE (2021) Narrative scaffolding. Rev Philos Psychol. https://doi.org/10.1007/s13164-021-00595-w

Fabry RE (2023) What is the relationship between grief and narrative? Philosophical Explorations 1–7. https://doi.org/10.1080/13869795.2023.2183241

Fodor JA (1980) Methodological solipsism considered as a research strategy in cognitive psychology. Behav Brain Sci 3(01):63–73

Freud S (1917) Mourning and Melancholia. In: Strachey J, Strachey J (eds) The Standard edition of the complete psychological works of Sigmund Freud, vol 14. Hogarth Press, pp 243–258

Goldman AI (2012) A moderate approach to embodied cognitive science. Rev Philos Psychol 3(1):71–88. https://doi.org/10.1007/s13164-012-0089-0

Gross JJ (1999) Emotion regulation: past, present, future. Cogn Emot 13(5):551–573. https://doi.org/10.1080/026999399379186

Gross JJ (2001) Emotion regulation in adulthood: timing is everything. Curr Dir Psychol Sci 10(6):214–219. https://doi.org/10.1111/1467-8721.00152

Happify (2022) Happify. https://www.happify.com/public/about/

Higgins KM (2013) Love and death. In: Deigh J (ed) On emotions: philosophical essays. Oxford University Press, pp 159–178

Hofstadter DR (1995) Fluid concepts and creative analogies: computer models of the fundamental mechanisms of thought. Basic Books

Jiménez-Alonso B, de Brescó I (2023) Griefbots. A new way of communicating with the dead? Integr Psychol Behav Sci 57(2):466–481. https://doi.org/10.1007/s12124-022-09679-3

Klass D, Silverman PR, Nickman SL (eds) (1996) Continuing bonds: New understandings of grief. Routledge

Korsgaard C (1996) The sources of normativity. Cambridge University Press

Krueger J (2020) Schizophrenia and the scaffolded self. Topoi 39(3):597–609. https://doi.org/10.1007/s11245-018-9547-3

Krueger J, Osler L (2019) Engineering affect: emotion regulation, the internet, and the techno-social niche. Philosophical Top 47(2):205–231

Krueger J, Osler L (2022) Communing with the dead online: Chatbots, grief, and continuing bonds. J Conscious Stud 29(9–10):222–252. https://doi.org/10.53765/20512201.29.9.222

Laestadius L, Bishop A, Gonzalez M, Illenčík D, 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. https://doi.org/10.1177/14614448221142007

Lam B (2023) (n.d.). *The digital future of grief* (Season 6, Episode 1). Retrieved August 16, from https://hiphination.org/season-6-episodes/s6-episode-1-the-digital-future-of-grief/

Lavelli A, Sebastiani F, Zanoli R (2004) Distributional term representations: An experimental comparison. *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management*, 615–624. https://doi.org/10.1145/1031171.1031284

Lewis D (1979) Scorekeeping in a language game. J Philosophical Log 8(1):339–359. https://doi.org/10.1007/BF00258436

Lindemann NF (2022) The ethics of 'deathbots'. Sci Eng Ethics 28(6):60. https://doi.org/10.1007/s11948-022-00417-x

Maiese M (2016) Affective scaffolds, expressive arts, and cognition. Front Psychol 7:1–11. https://doi.org/10.3389/fpsyg.2016.00359

Markovic J (2022) Unchosen transformative experiences and the experience of agency. Phenomenology and the Cognitive Sciences 21(3):729–745. https://doi.org/10.1007/s11097-021-09753-y

Merritt R (2022) What is a transformer model. *Nvidia*. https://blogs.nvidia.com/blog/2022/03/25/what-is-a-transformer-model/#:~:text=A%20transformer%20model%20is%20a,25%2 C%202022%20by%20Rick%20Merritt

Mi C, Xie L, Zhang Y (2022) Improving data augmentation for low resource speech-to-text translation with diverse paraphrasing. Neural Netw 148:194–205. https://doi.org/10.1016/j.neunet.2022.01.016

Millar B, Lopez-Cantero P (2022) Grief, continuing bonds, and unreciprocated love. South J Philos 60(3):413–436. https://doi.org/10.1111/sjp.12462

Mitchell M, Krakauer DC (2023) The debate over understanding in AI's large language models. *Proceedings of the National Academy of Sciences*, *120*(13), e2215907120. https://doi.org/10.1073/pnas.2215907120

Norlock KJ (2017) Real (and) imaginal relationships with the dead. J Value Inq 51(2):341–356. https://doi.org/10.1007/s10790-016-9573-6

Patulny R, Bower M (2022) Beware the loneliness gap? Examining emerging inequalities and long-term risks of loneliness and isolation emerging from COVID-19. Australian J Social Issues 57(3):562–583. https://doi.org/10.1002/ajs4.223

Ratcliffe M (2017) Grief and the unity of emotion. Midwest Stud Philos 41(1):154–174. https://doi.org/10.1111/misp.12071

Ratcliffe M (2023) Grief worlds: a study of emotional experience. MIT Press

Ratcliffe M, Byrne EA (2022) Grief, self and narrative. Philosophical Explorations 25(3):319–337. https://doi.org/10.1080/13869795.2022.2070241

Rudin C, Radin J (2019) Why are we using black box models in AI when we don't need to? A lesson from an explainable AI competition. Harv Data Sci Rev 1(2):1–9. https://doi.org/10.1162/99608f92.5a8a3a3d

Rupert RD (2004) Challenges to the hypothesis of extended cognition. J Philos 101(8):389–428

Saarinen JA (2020) What can the concept of affective scaffolding do for us? Philosophical Psychol 33(6):820–839. https://doi.org/10.1080/09515089.2020.1761542

Skjuve M, Følstad A, Fostervold KI, Brandtzaeg PB (2021) My chatbot companion—A study of human-chatbot relationships. Int J Hum Comput Stud 149:102601. https://doi.org/10.1016/j.ijhcs.2021.102601

Skjuve M, Følstad A, Fostervold KI, Brandtzaeg PB (2022) A longitudinal study of human–chatbot relationships. Int J Hum Comput Stud 168:102903. https://doi.org/10.1016/j.ijhcs.2022.102903

Stephan A (2012) Emotions, existential feelings, and their regulation. Emot Rev 4(2):157–162

Sterelny K (2010) Minds: extended or scaffolded? Phenomenology and the Cognitive Sciences 9(4):465–481. https://doi.org/10.1007/s11097-010-9174-y

Stokes P (2015) Deletion as second death: the moral status of digital remains. Ethics Inf Technol 17(4):237–248. https://doi.org/10.1007/s10676-015-9379-4

Stokes P (2021) Digital souls: a philosophy of online death. Bloomsbury

Sutton J (2016) Scaffolding memory: themes, taxonomies, puzzles. In: Bietti L, Stone CB (eds) Contextualizing human memory: an interdisciplinary approach to understanding how individuals and groups remember the past. Routledge, pp 187–205

Tekin Ş (2021) Is Big Data the new stethoscope? Perils of digital phenotyping to address mental Illness. Philos Technol 34(3):447–461. https://doi.org/10.1007/s13347-020-00395-7

Thomas K (2010) Haitian zombie, myth, and modern identity. In *CLCWeb: Comparative Literature and Culture: Vol. 12.2*. https://docs.lib.purdue.edu/clcweb/vol12/iss2/12/

Timms R, Spurrett D (2023) Hostile scaffolding. Philosophical Papers 1–30. https://doi.org/10.1080/05568641.2023.2231652

Varga S (2019) Scaffolded minds: integration and disintegration. MIT Press

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 30). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Walter T (2015) Communication media and the dead: from the Stone Age to Facebook. Mortality 20(3):215–232. https://doi.org/10.1080/13576275.2014.993598

Weber-Guskar E (2022) Reflecting (on) Replika. Can we have a good affective relationship with a social chatbot? In J. Loh & W. Loh (Eds.), *Social Robotics and the Good Life* (pp. 103–126). transcript Verlag. https://doi.org/10.1515/9783839462652

Weizenbaum J (1966) ELIZA—a computer program for the study of natural language communication between man and machine. Commun ACM 9(1):36–45

Witek M (2015) Mechanisms of illocutionary games. Lang Commun 42:11–22. https://doi.org/10.1016/j.langcom.2015.01.007

Wood D, Bruner JS, Ross G (1976) The role of tutoring in problem solving. J Child Psychol Psychiatry 17(2):89–100

Xu F, Uszkoreit H, Du Y, Fan W, Zhao D, Zhu J (2019) Explainable AI: a brief survey on history, research areas, approaches and challenges. In: Tang J, Kan M-Y, Zhao D, Li S, Zan H (eds) Natural Language Processing and Chinese Computing. Springer International Publishing, pp 563–574. https://doi.org/10.1007/978-3-030-32236-6_51