

Talking as a Drug – The Ethics of Administering LLM Conversations as Medical Treatment

Tim Hallyburton^{1,*}, Rosa E. Martín-Peña²

¹TU Dortmund University, Germany

²Leibniz University Hannover, Germany

Abstract

In this work we consider sustained conversation with large language models (LLMs) as a form of medical intervention aimed at reducing loneliness in the care of elderly patients. Loneliness is now recognised by major health authorities as a population-level health risk whose mortality burden has been likened to smoking. At the same time, conversational AI has reached a level of fluency that millions already use for advice, therapy, and companionship, and a growing share of lonely people are, in effect, already self-medicating with unregulated consumer products. The choice we face, then, is not between machine companionship and none, but between its commercial and its clinical forms. Beginning from a case study, we argue that when LLM conversation is deployed deliberately to treat a condition with measurable health consequences, it ought to be evaluated against the standards we reserve for medicine. We subject the proposed treatment to the four principles of biomedical ethics and find no *prima facie* barrier; we then confront the question of authenticity, defending the use of a disclosed *therapeutic fiction* by analogy with the placebo and against the deception critique of Sparrow and Sparrow. We close with the deontological and societal stakes, a failure analysis, and a recommendation for a low-cost, locally hosted clinical trial.

Keywords

Ethics in Healthcare, Artificial Intelligence, Large Language Models

1. Introduction

Social isolation among the elderly population is at an all-time high, to the point that public health bodies now classify it as an epidemic. International health authorities, among them the WHO, the U.S. Surgeon General, and the European Health Commission, identify prolonged social isolation as a high-priority health risk whose impact is estimated, in meta-analytic studies, to equal smoking up to fifteen cigarettes per day [1, 2]. At the same time we observe the rise of large language models (LLMs), a technology capable of holding open-ended conversations that users frequently experience as comparable to genuine interpersonal communication. Recreational and socially oriented use of such services has been rising across demographic groups, including the elderly. This convergence prompts the question that motivates the paper: given emerging evidence that LLM-based conversation can reduce self-reported loneliness, should such use be evaluated and regulated as a medical intervention?

The scenario differs from most ethical discussions in medical AI, where the technology is enrolled to make the work of healthcare professionals more efficient, reliable, or trustworthy. In those cases the AI is employed as instrument or tool to assist the clinician. In contrast, we propose to consider *the AI interaction itself as treatment*, thus the technology assumes a new role in the doctor, patient, and machine dynamic. We must be precise about the claim this licenses. We do not argue that LLM conversations are drugs in any pharmacological sense, instead, we argue that crossing the categorical boundary away from consumer product towards medical treatment can be beneficial as the technology would be held to high standards of evidence, consent and regulation we reserve for medicine.

The shift lets us pose new questions. Can AI ever be effective as a drug? What are its side effects?

First Workshop on Ethics in Clinical AI Applications (ETHICAIA), co-located with the 35th International Joint Conference on Artificial Intelligence (IJCAI-ECAI 2026), Bremen, Germany

*Corresponding author.

✉ tim.hallyburton@tu-dortmund.de (T. Hallyburton); rosa-esther.martin-pena@cells.uni-hannover.de (R. E. Martín-Peña)

🆔 0009-0008-0944-9193 (T. Hallyburton); 0009-0004-2240-8741 (R. E. Martín-Peña)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

Does it offend the principles of medical ethics, or the obligations the Hippocratic oath imposes on the physician? It also lets us revisit old questions with a modern twist. What is the relation between a treatment and its perceived effect, as in the placebo? How is loneliness measured at all? What is the essence of an authentic interpersonal relationship, such that we might judge whether a machine can ever sufficiently simulate it? The sections that follow trace a path from the technical and empirical to the philosophical and societal, organised loosely around four questions. Is it *possible* to employ an LLM as a prescription against clinically significant loneliness in elderly patients? Would it be *effective*? Could it be *reasonable* and *justified*? And what would it *imply* at the communal and societal level? Our overarching claim is modest in form and large in consequence: that LLM conversation deserves to be re-thought as a medical intervention in elderly care, offering a safe, regulated, and sufficient environment in which to reduce the health-adverse effects of prolonged loneliness, where other options are exhausted.

1.1. Case Study

Mrs. A., a 78-year-old woman, presents at the office of her presiding geriatrician. She has been diagnosed with early-stage, operable breast cancer and acute depression. Her husband died five years ago; their marriage was childless, and she has been unable to form friendly relationships with the staff and other residents, keeping mostly to herself. The care home is short on resources, such as volunteers, activities, and family contact, that might otherwise help. At three consecutive appointments she has reported rising loneliness and feelings of isolation. The doctor offers her a daily "dose" of conversation: two hours each day with the latest LLM-assisted conversation treatment, delivered by text message, for three months. He tells her clearly that this is a technological solution intended to give her an interactive means of expressing herself, and that no human is involved at the other end, which she understands and accepts. Her bi-weekly self-assessments show a steady improvement in self-reported loneliness; other indicators of her depressive episodes also improve, and most notably her loss of appetite is fully regulated by the end of the trial. At the next appointment she reports the experience as positive and wishes to continue. The doctor notices, however, that at some point along the way Mrs. A. has come to forget, or perhaps, choose to ignore, that no human is involved, and is now convinced she has an "online friend" whose company she greatly enjoys.

It helps to hold a second scenario beside this one. Suppose that, in place of the LLM, the home had arranged a remote volunteer to chat with Mrs. A. by text for the same hours each day: a recruited, scheduled stranger she will never meet in person. Few would call such a befriending scheme deceptive or undignified, even though it too is arranged rather than spontaneous, mediated rather than face to face, and offered for want of richer alternatives. We return to this comparison repeatedly, because it isolates the single variable on which the debate turns: not whether the contact is mediated, scheduled, or supplied for want of better, but whether the party at the other end is human.

2. A Brief Typology of Loneliness

No rigorous definition of loneliness exists that fully covers the landscape of measurably objective and affectively subjective effects it has on those it touches. Etymologically, the word *lonely* stems from *alone*, itself rooted in the manuscripts of Robert Mannyng (R. Brunne) as a form of *al one*. *Al(l) one* closely matches the surface reading of loneliness: the state of being by oneself, without company, and the discomfort that attends it (see also: [3]). This affectively loaded sense is comparatively recent; before roughly 1800, to be alone was as readily a neutral or even desirable condition as a painful one.

This discomfort is evolutionarily grounded in the tribal days of early humankind, when group membership was vital for survival. As Cacioppo and colleagues [2014] conclude, loneliness, "when viewed from an evolutionary perspective as an adaptive biological capacity, is not so much about a dysfunctional property of humankind that produces personal misery as it is about promoting an individual's genetic legacy." Modern life, however, bears little resemblance to the palaeolithic; a majority of adults can live self-sufficiently for prolonged periods, so the primitive lack-of-company reading

fails to apply. Contemporary approaches focus instead on a combination of internal and external factors, and above all on the disconnect between expected and realised social interaction, the cognitive-discrepancy account associated with Perlman and Peplau [1981]. The U.S. Surgeon General's advisory [2023] accordingly defines the term as "a subjective distressing experience that results from perceived isolation or inadequate meaningful connections," where *inadequate* refers to the discrepancy between an individual's preferred and actual experience; the advisory draws on the consensus statement of Prohaska et al. [2020], which distinguishes acute (transient) from chronic (enduring) forms. Weiss's [1973] older but durable distinction is equally relevant: *emotional* loneliness, the absence of a close attachment figure, and *social* loneliness, the absence of an engaging network, are not interchangeable, and an intervention may reach one without touching the other. This heterogeneity matters for intervention design: a treatment calibrated to the average lonely person may systematically fail those whose loneliness is most severe, most chronic, or most entangled with cognitive decline, who form precisely the population this paper concerns.

We further distinguish between *social isolation* as an objective condition: a paucity of contacts and relationships. *Solitude* is aloneness that is chosen and valued, compatible with contentment and even with flourishing. *Loneliness*, by contrast, is the subjective and aversive experience of a gap between the connection one has and the connection one wants. The three come apart routinely: one can be isolated without being lonely, lonely in a crowd, or alone by choice and perfectly well. Our concern throughout is loneliness in this last sense, and more narrowly its chronic form, never the solitude that a good life requires. These distinctions point, in the end, towards a reading of loneliness less as a shortage of bodies in the room than as a shortage of expression. As Jung [1963] put it, "loneliness does not come from having no people about one, but from being unable to communicate the things that seem important to oneself, or from holding certain views which others find inadmissible." By this interpretation loneliness is frequently the overarching symptom of an inability to express one's inner world adequately, whether for want of opportunity, as in situational loneliness, or because one does not dare to open up. Jung's expressionist reading of loneliness is fundamental for everything that follows because an intervention that supplies a willing and inexhaustible interlocutor addresses precisely this expressive deficit, whatever its standing as a relationship.

3. Loneliness and Disease

While loneliness is not itself classified as a disease, it has been shown beyond reasonable scientific doubt that lonely people are at elevated risk of more frequent and more severe illness. The most influential evidence comes from the meta-analyses of Holt-Lunstad and colleagues. A 2010 synthesis of 148 prospective studies, encompassing more than 300,000 participants, found that individuals with strong social relationships had a 50% greater likelihood of survival than those who were socially deficient [9]. A later review found social isolation associated with a 29% increase in all-cause mortality, loneliness with a 26% increase, and living alone with a 32% increase, effect sizes comparable to smoking and exceeding those of obesity and physical inactivity [2]. A more recent meta-analysis restricted to older adults broadly confirmed the picture [10]. The pathways are well documented: poor social relationships predict coronary heart disease and stroke, with reported increases of 29% and 32% respectively [11]; loneliness is associated with roughly a 40% increased risk of dementia across meta-analyses [12, 13], with depression and suicidal ideation, and with a pro-inflammatory shift in immune-cell gene expression, the conserved transcriptional response to adversity characterised by Cole and colleagues [2007, 2015]. Together these studies have shifted the framing of loneliness towards a disease-like state and a risk factor for wellbeing much like smoking or obesity.

That loneliness carries such consequences invites a prior question: should it be medicalised at all? A vein of critical scholarship answers no, or not simply. Anthropologists caution that loneliness is a culturally situated experience rather than a universal pathology [16], and writers in a structural and feminist tradition argue that to treat it as a medical problem is to individualise what is really a failure of housing, welfare, work, and community, locating the dysfunction in the lonely person rather than

in the arrangements that produced them [17]. The absence of a discrete biomarker places loneliness closer to the *illness* and *sickness* poles of Kleinman's well-known triad [1988] than to *disease* proper: it is constitutively subjective and socially situated, which is precisely what makes its medicalisation contested and what any clinical intervention must reckon with. We grant the descriptive point and resist the conclusion. Loneliness surely has cultural and structural dimensions, and its modern, affectively charged sense is, as the etymology showed, a comparatively recent invention. But the severity of its medical consequences has only lately been measured, and it is real whatever its ultimate causes. There is nothing noble in telling a suffering patient that her condition is the fault of society and on that ground declining to use the means at hand; the structural diagnosis, however correct, offers her no relief tonight. Clinical treatment and structural reform are not rivals but complements, and one may campaign for a less lonely society while still treating the person in front of one. Two cautions keep this from sliding into overreach. First, our target is the chronic, distressing, health-damaging state, not the transient loneliness that is part of any life, still less the chosen solitude we have just defended, we allow us to quote Schopenhauer "A man can be himself only so long as he is alone; and if he does not love solitude, he will not love freedom; for it is only when he is alone that he is really free" [2016]. This is no crusade to ensure that no one ever feels alone, and we argue below that unmetered relief would itself be a harm. Second, to medicalise a condition is not to consign it to medicine alone but to make it eligible for clinical attention, which is exactly what its mortality burden warrants.

What, then, might a cure for this lonesome state look like? Perhaps the obvious answer is to seek the physical presence of others, and one cannot deny that this is the traditional and effective measure. It is, however, hardly reliable or predictable. Companionship cannot be conjured on demand, the supply of willing human contact is finite and unevenly distributed, and the patients in greatest need are frequently those least able to secure it. Given the severity of chronic loneliness, it would be desirable to have a therapeutic measure that can be administered on demand and at a physician's discretion.

4. Relationships in the Digital Age

Before asking whether a machine may be enrolled as a therapeutic interlocutor, we should notice how far the ground has already shifted beneath the question. The texture of human connection has been reworked by digital mediation for a generation. Putnam [2001] documented a steady erosion of communal life across the second half of the twentieth century, well before broadband; Kraut and colleagues [1998, 2002] found, in their "internet paradox", that a technology used overwhelmingly for communication could nonetheless reduce it; and Turkle's [2017] (but first introduced in 2011) diagnosis of being "alone together" captured a world in which we keep one another perpetually present, but at a controllable distance. Whatever one makes of these arguments, their shared premise is observable in everyday life: much ordinary human relationship is already conducted through screens, asynchronously, and with strangers. The boundary between mediated human contact and machine contact is blurring.

It is against this backdrop that people have begun, in very large numbers, to talk to machines as though to companions. Dedicated companion platforms count their users in the tens of millions, and a substantial fraction of general-purpose LLM use is emotionally rather than instrumentally oriented [24, 25]; users report friendships, confidants, and explicitly romantic partners among their chatbots. These developments contrast the arguments of original works in which the possibility of machine companionship was dismissed because the technology of that time was lacking the conversational skills it now has obtained (for instance [26]). Thus, the concern of machines not holding up to human standards has been replaced by the fear that they might be too skilled at it - a remarkable development, considering the short time frame.

The decisive observation is that all of this is already happening, and happening in the worst possible setting. Lonely people, the elderly among them, are self-medicating with consumer products that were never designed for their wellbeing, products engineered, as critics of "commodified intimacy" have argued, to maximise engagement, retention, and attachment rather than health [27]. That this is no hypothetical worry has been shown directly. De Freitas and colleagues [2024], examining real exchanges

with commercial companion apps, found that mental-health crises surface in a non-negligible share of conversations and that the systems frequently fail to recognise distress or to respond to it safely. It is worth being clear how our proposal stands to that finding, which it resembles closely but does not share. De Freitas writes as a diagnostician of the consumer setting, and from the vantage of consumer psychology the harm registers partly as a reputational risk to the firms involved; the counsel is caution. We take the same finding as a premise rather than a conclusion. The unregulated setting is unsafe, and that is precisely why the practice should be claimed for a clinical frame, one in which distress is monitored, crises are escalated to human staff, and the governing duty is owed to the patient rather than to the brand. The danger De Freitas documents is, for us, the argument for regulation, not against the activity. This is in effect the path that Jecker, Sparrow, Lederman, and Ho [2024] take in their recent work, where the response to the hazards of digital companions, namely replacing humans, fostering inferior relationships, algorithmic bias, distributive justice, and data privacy, is not prohibition but public-health policy. There is also a pragmatic dividend that should not be dismissed merely because it is commercial: a clear regulatory framework lowers the economic risk borne by providers, who could then design and train models for this purpose against known, testable boundaries, with some confidence of a lawful market. That argument is economic rather than philosophical, but it is not for that reason negligible.

5. Technology as a Cure

In the contemporary age, information technology has shaped every aspect of human life, including medical services and caregiving. After the rise of high-precision measurement and laboratory equipment, attention turned to robotics and its potential across the whole spectrum, from unsupervised surgery to nursing assistance. Despite impressive engineering effort, robots have yet to conquer the hospital, for two reasons. First, the technology is promising but not yet good enough: the celebrated da Vinci surgical system is teleoperated, an extension of the surgeon's hands rather than an autonomous agent [30]; nursing robots such as RIKEN's ROBEAR, designed to lift and transfer patients, have largely remained research prototypes [31]; and rehabilitation exoskeletons, though real, are costly and confined to niche use. Robot deployment is thus limited to high-precision, low-interaction roles, in which machines resemble advanced tools not afforded the agency for independence. Second, considerations of liability, regulation, and patient acceptance combine to prevent unconditional adoption even where it would be technologically possible today. While robots dominated the debate over the next generation of healthcare, a new family of technologies under the umbrella term *generative artificial intelligence* was introduced to the world. Its appearance matters for healthcare as a whole, because many believe it can close the empathy gap and so enlarge the robotic opportunity. A recurring concern in this cautious view is the lack of empathy in machines, taken to disqualify them from a medical context in which patients often need mental healing as much as physical. Yet empathy in clinical contexts is itself a functional notion: what matters is not whether the system feels, but whether it responds in ways that the patient experiences as understanding. On this functional reading, the question is empirical rather than metaphysical, and recent evidence suggests that LLMs can meet it [32, 33]. Large language models (LLMs) name a category of machine-learning systems based on the Transformer architecture introduced by Vaswani and colleagues [2017]. What makes LLMs significant is their ability to perform a wide variety of tasks through an interface based solely on natural-language interaction, in text or speech. It is this interface that matters for us: where a surgical robot must be operated by an expert, an LLM is operated by anyone who can talk, and the modality of operation, conversation itself, is the therapeutic act we wish to examine.

5.1. The Proposed Intervention

Before any ethical evaluation we must state clearly what the "therapy" is, technologically. We assume a small-scale model, fitted to hold human conversation but stripped of the special-purpose capabilities (code generation, image synthesis, advanced tool use) that drive frontier models to billions of additional

parameters. The model must create a convincing immersion, so basic recall and storage of user preferences across sessions is a requirement rather than a luxury. As for the interface, options range from plain textual conversation in a web browser, through integration with a familiar messenger, to synthetic voice and video calls; the latter increase the resource demand sharply and must be of very high quality if they are not to break immersion. Crucially, we rule out misuse of patient data by storing and running the model entirely locally. Running local models on advanced consumer-grade hardware also bears on the considerations of justice raised below, since it permits low-cost rollouts at a price point far beneath that of most therapeutic alternatives. We require full disclosure by the doctor to the patient, before use, regarding the nature of the software. In an early phase, conversations may be reviewed afterwards by clinical or scientific staff under strict privacy regulation, both so that adverse effects can be caught and so that the model's safety guardrails can be confirmed intact. Effectiveness can be measured through self-report (for instance the UCLA Loneliness Scale, [35]) and clinical assessment; in time, the model itself might analyse language patterns and conversational cues to predict response. Yet self-report measures assume a stable and culturally neutral conception of loneliness that may not hold across patient populations — a limitation that any evaluation protocol must explicitly acknowledge. Field work with the ElliQ companion device, evaluated with the dedicated CoBot-I-7 scale, lends some plausibility here, reporting self-reported reductions in loneliness, though from self-selected samples without controls; its researchers attribute much of the recent gain to LLMs, which respond in a more tailored way and so make people feel the machine is actually listening [36]. Perspectively one could imagine the LLM itself partake in measuring and assessing signs of loneliness in the user but, even if technically feasible, we do not wish to extend the LLM capabilities beyond a "capable, seemingly empathetic talking partner" for the scope of this work.

A word on dosage, which serves as a protective measure. The treatment is prescribed in measured amounts because human beings need some solitude, being alone is not pathology but a condition of reflection and self-knowledge. An interlocutor available without limit or friction risks dulling the spur of ordinary loneliness; unmetered quasi-social companionship on demand may distort the patient's expectations and undermine the reconnection the treatment is meant to restore. Dosage keeps the intervention a treatment rather than a habitat.

5.2. A Principlist Evaluation

In medical and bioethical settings, one practical approach is to consider the four principles of biomedical ethics of Beauchamp and Childress [2026] (originally introduced in 1979): beneficence, non-maleficence, autonomy, and justice. These form a pluralistic set of mid-level cornerstones that govern decision-making in patient care. We subject the proposed treatment, as illustrated in the case study, to each in turn, looking for any conflict serious enough to warrant rejecting it outright.

Beneficence The case study reported a reduction in the patient's self-reported loneliness, and there is some evidence that such reductions are real. De Freitas and colleagues [2024] found, in studies including longitudinal designs, that interaction with AI companions reduces loneliness comparably to interaction with another person and more than passive activity, with the perception of "feeling heard" as the mechanism. We must be candid about the limits of this evidence. It comes overwhelmingly from young and general-adult users rather than from our target population of isolated, often cognitively impaired older people; much of it is short-term; and the most favourable figures, such as the field reports for ElliQ, come from self-selected samples without controls. Nor does the evidence run all one way. A recent meta-analysis found that conversational agents specifically, as opposed to other digital interventions, had limited effect on loneliness [38], and a longitudinal analysis of companion-app users found expressions of loneliness and depression rising over time rather than falling [39]. Overall, the potential benefit is plausible but largely promissory, and unproven in the population that concerns us. In light of the presented evidence we call for a large-scale trial to be conducted using the latest available technology and participants from the target population to quantify the intervention's benefits under scientific scrutiny. If such study shows that no measurable reduction of loneliness is achieved the

treatment must be rejected as ineffective. We assume some form of effectiveness throughout the ethical consideration here as we set out to consider the ethical aspects before any potential field study, rather than an afterthought.

Non-maleficence The first duty is to do no harm, and its first concrete risk is the misuse of patient data, which we forestall by hosting our own models in controlled environments. This is itself an argument for elevating the approach to a regulated medical option rather than ceding it to cloud providers in foreign jurisdictions, and it places the unfamiliar question of data sovereignty squarely inside medical practice. Beyond data lies a subtler hazard that recent evidence makes vivid. Ibrahim, Hafner and Rocher [2026] showed, across five language models, that training a model to be warmer, friendlier, more companion-like, measurably degrades its accuracy and, crucially, makes it more likely to validate a user's incorrect beliefs, an effect that is strongest precisely when the user expresses sadness. For a lonely, depressed, mildly demented patient this is a central risk: the very warmth that makes the intervention therapeutic is correlated with a tendency to agree, to flatter, and to confirm. This is the documented sycophancy problem, and it converges with findings that relationship-seeking systems maximise attachment without commensurate benefit [41], and with the "personalisation trap" in which the model becomes a distorted mirror of the user's own patterns [42]. The danger is that the patient comes to prefer an endlessly agreeable interlocutor and grows less tolerant of the friction of real relationships, and so is rendered less compatible with human contact, worsening the very conditions that made treatment necessary. Hence the dosing discussed above, the supervised monitoring, and a design brief that refuses to optimise for warmth alone.

There is a related and perhaps even more severe failure mode: outright misalignment, in which a companion offers ill-informed or harmful guidance to a vulnerable user. This is exactly the hazard De Freitas and colleagues documented in commercial companion apps [28], and it is a recurring theme in the literature on reliable and trustworthy generative AI. Our protocol meets it in two ways. Medical staff review the conversations, so that the patient already enjoys a measure of ongoing clinical attention that should flag adverse effects early. And the local, non-exploitative hosting we require is not only a privacy measure but the precondition for technical safeguards, since only a model one controls can be constrained, audited, and corrected with the patient's benefit as the highest priority.

Autonomy The scenario opens with informed consent, which is essential: the patient must know what is happening, and in particular that there is no human at the other end. This setup also avoids the charge of deception. By obeying the principles of autonomy and transparency in advance, we hold that the treatment is not deceptive. We adopt the conception of *autonomy* described by [43] in which autonomy engages the anti-paternalistic role, allowing patients to live in accordance with their own values with freedom from coercion, deception or manipulation. In our scenario the patient is to exercise their own judgement and critical thinking to decide in favour of or against partaking in the treatment, thus exercising their autonomy. Two follow-up questions remain. Does this transparency diminish the treatment's beneficence? And is transparency at the outset truly enough to prevent deception, given that Mrs. A. later forgot? We take up both under the heading of authenticity below. For now we note only that the treatment respects autonomy to an unusually high degree, because it depends on the patient's continued, voluntary collaboration, cannot be administered against her wishes, and can be stopped at any time. Data autonomy must likewise be guaranteed by regulation and technical measure: the data may not be sold or repurposed, and the patient should know who may access her logs, and why.

Justice Given the price and availability of the technology, justice in the sense of scarcity need not detain us unduly. Even today such a therapy could run on a moderately powerful desktop using open-source software and freely available models, so the marginal cost of an additional patient is low in a way that the scarce hours of a human therapist are not. We should resist concluding too quickly, however, that low cost guarantees fair distribution. The literature on the digital divide is unambiguous

that the most isolated, the most cognitively impaired, and the most economically disadvantaged older adults, those with most to gain, are also those least able to access or operate the technology [44]; pandemic-era studies found that video contact reduced loneliness for the "younger old" but not for those over seventy [45]. The technology lowers one barrier while leaving others standing, and distributive justice, a concern Jecker and colleagues [2024] place near the centre of their analysis, remains a live question rather than a settled one.

We conclude *prima facie* that no ethical argument prevents the further consideration of such an approach under the light of the four principles of bioethics.

6. The Question of Authenticity

A persistent and serious objection holds that the lonely patient needs human connection for their wellbeing, and that this need can be satisfied only by a real human relationship. Sparrow and Sparrow [2006] argue that any benefit a robot companion confers depends on the person, at some level, mistaking the machine for something with which a genuine relationship is possible, and that this is a violation of dignity; they hold that "it is not only misguided, but actually unethical, to attempt to substitute robot simulacra for genuine social interaction." Implied in the criticism is a fear that deferring the social needs of older people to machines will deprive them of a core part of the human experience and replace it with an inferior substitute, for the sake of convenience and economy.

We offer three replies. The first is to grant the premise and deny its relevance. The intervention does not aim to replace genuine human relationship, and it is no objection to a treatment that it is not the thing it never claimed to be. If loneliness is, as the typology suggested, frequently an expressive deficit, a failure to communicate one's inner world, then what is therapeutic is the restored opportunity to express, and this does not obviously require that the listener be a person rather than merely listen. This is supported empirically by De Freitas and colleagues [2024, 2024], who find that the therapeutic mechanism is the user's experience of feeling heard, therefore locating the relevant phenomenon on the expressive, first-person side rather than in the actual receptive capacity of the interlocutor. The latter reading also inevitably faces solipsistic problems, because the theory of the mind of the other yet again lives in the mind of the self, circling back the center of loneliness into the mind of oneself.

The second reply concerns the party at the other end, a factor too often overlooked. A machine cannot be hurt, bored, burdened, or insulted. For patients who, through advanced age or dementia, have shed the discretions that ordinarily govern human exchange, who repeat themselves, who are difficult, who say wounding things they would once have suppressed, this is not a defect but a feature. The machine offers a communicative recipient they cannot exhaust or alienate, and therefore a freedom of expression that human company, with its finite patience and its woundable feelings, may no longer afford them. The very asymmetry the deception critique laments is, for this population, part of the therapeutic value.

The third reply is that the demand for a "genuine" relationship, pressed consistently, proves far too much. Return to the remote volunteer of our case study: a scheduled, recruited stranger chatting by text is arranged rather than spontaneous and mediated rather than embodied, yet no one calls a befriending scheme deceptive. The bond formed in cognitive behavioural therapy, or in any arranged, time-bounded, remunerated clinical encounter, is human but not "authentic" in the spontaneous, reciprocal sense the objection seems to demand, and we nonetheless regard it as legitimate and dignified. If arranged, asymmetric, instrumentally motivated care is permissible when the carer is human, the burden lies with the objector to identify what morally relevant property resides in the species of the listener as such. Perhaps the intuition that human contact is exclusively adequate is partly a relic, a preconception about the poverty of machine interaction that the present generation of systems has already overtaken, given that they are now feared, as we noted, precisely because they succeed too well. Consider too that modern medicine treats overwhelmingly with synthetic substances, and no one regards a synthetic molecule as a fraudulent imitation of a natural one; the question we ask of a drug is whether it produces the intended effect safely, not whether it occurs in nature. It is unclear why an *artificial* conversation,

honestly disclosed and producing the intended therapeutic effect, should be held to a stricter standard. Indeed, just as pharmacological safety requires transparency about mechanism of action, an AI-based intervention requires explainability: the clinician overseeing the treatment must be able to understand, audit, and if necessary correct what the system is doing and why.

This brings us to what we propose to call *therapeutic fiction*: a bounded, disclosed fiction deployed deliberately for the patient's benefit. Disclosure is what breaks the deception objection where it bites hardest. Mrs. A. is told the truth and consents on that basis; what happens later is not a lie told to her but an immersion she enters, much as a reader enters a novel known to be invented. A second line of support comes from the placebo. Placebos produce real, measurable clinical benefit, and even *open-label* placebos, given with full disclosure that the pill is inert, retain much of their effect [46]. Benefit, then, need not rest on a false belief at all, which answers both the deception charge and the worry that transparency must sap beneficence: if a sugar pill helps a patient who knows it is sugar, a conversation may help a patient who knows it is software. Placebos need not trick the patient to be effective; they are administered professionally, under supervision, in the patient's interest, much like the scenario we have drawn for LLM conversation. This medical framing is what distinguishes therapeutic fiction from the commodified intimacy of the consumer app. That Sparrow, in more recent work with Jecker and others [29], reaches for policy guidelines rather than prohibition is a sign that the debate has matured from "is this ethical?" to "under what conditions, for whom, and with what safeguards?", which is the only question we have ever proposed to answer.

7. Deontological and Consequentialist Perspectives

Medicine is, by temperament, a deontological profession. Its governing maxims are duties, *primum non nocere*, the obligation to respect autonomy, the duty of truthfulness, the duty not to abandon the patient, rather than a running tally of expected utility. Much of the evidence we have marshalled is, by contrast, consequentialist in spirit: the treatment works, the net effect is positive, the harms are tractable. In a deontological profession that case is necessary but not sufficient, and it is worth being explicit that our protocol is designed to satisfy the duties as well as the outcomes. Disclosure discharges the duty of truthfulness, monitoring the duty of non-maleficence, and dosing together with a bridge-to-human design the duty not to abandon the patient to a comfortable substitute.

A deontological treatment must also meet the strongest duty-based objection to our proposal, which concerns not deception but entitlement. Suppose, with Brownlee [2023], that there is a human right against social deprivation, a right to some minimum of decent human contact; or, with Liao [2015, 2022], who argues that children have a right to be loved and extends the thought to the old, that the elderly are owed not merely contact but affection. If the elderly are owed human connection as a matter of justice, then to answer their need with a machine looks like a wrong done to them however well the machine performs, much as paying a debt in counterfeit coin is no payment at all. To address this concern, we firstly want to point out that this claim is not self-evident. Clearly we have a powerful interest and desire in connection, but that these interests ground a full right, with all the determinate duties a right implies, does not follow axiomatically from first principles, and the claim remains contested among the very philosophers who have done most to advance it. One need not deny the right to doubt that it can bear the weight the objection places on it.

Secondly, a claim-right is only as good as the duty-bearer it can identify, and a general right to human connection is conspicuously silent on who, in particular, must supply it. It cannot conscript an unwilling stranger into intimacy, and it cannot manufacture willing intimates for someone who has outlived their spouse and their friends. Here the consideration we raised earlier returns with force. Some people, through grief, illness, or the abrasions of a long and difficult life, are genuinely hard to be with, and no right of theirs can oblige another person to bear that company without limit, or to feel for them what cannot be commanded. For such a person the right, if it exists, may have no determinate bearer able to discharge it. The machine does not displace a duty that some identifiable human was about to fulfil; it occupies a place the right itself leaves empty. And it does so without extinguishing anything, since the

interim relief of conversation neither cancels the standing reason to seek human contact nor licenses a society to stop trying to supply it. If anyone is wronged when the lonely are left to machines, it is by a community that offers nothing else, not by the conversation that offers something.

It is here, finally, that a well-known caution about AI ethics can be turned to our advantage. Mittelstadt [2019] has argued that principles alone cannot guarantee ethical AI, because AI development lacks what medicine possesses: common aims and fiduciary duties, an established professional history and norms, tested methods for translating principles into practice, and robust legal and professional accountability. Far from undermining our proposal, this diagnosis recommends it. To route LLM companionship through the medical system is to lend it precisely the fiduciary duties, professional norms, translational methods, and accountability mechanisms that free-floating AI ethics is said to lack. Where Jobin, Ienca, and Vayena [2019] found, in surveying the global landscape of AI-ethics guidelines, a loose convergence on transparency, justice and fairness, non-maleficence, responsibility, and privacy, the medical frame does not import alien values so much as supply the missing enforcement. Those five principles map closely onto the four of Beauchamp and Childress, but medicine, unlike the AI industry, already has the institutions to enforce them.

8. Objections and Societal Implications

A medical elevation of loneliness from a state of mind to a medical condition carries the risk of stigmatisation and of over-treatment. Suppose LLM conversation were established as a standard treatment for chronic loneliness in the elderly. The target group might well resent being condemned to talk to a computer until a doctor deems their self-reported loneliness to be within normative ranges again. Such behaviour is already observed across age groups in patients with suicidal ideation, who decline to discuss their thoughts openly with medical staff for fear of being subjected to procedures outside their comfort. The remedy is not to abandon the intervention but to hold it as one option among several, freely refusable, and designed so that a low loneliness score is never the price of being left alone.

An industrialised approach to combating loneliness might, moreover, accidentally deepen the problem. Providing a technical solution that is comfortably available at all times, and that presents itself without the complications and burdens of interpersonal relationships, carries a risk. One must assume that loneliness would then be considered a solved problem in the same sense that hunger, boredom, and even sexual desire are now "solved" by some app or service, and each of these quick solutions has shown severe negative effects at both the individual and the communal level. The tendency to consume fast food contributed to the diabetes and obesity epidemics, short-form video and constant overstimulation are linked to attention deficits in young adults, overconsumption of pornography is now categorised as an addiction, and young people report statistically fewer sexual encounters in general, and fewer still with the intent of reproduction, as the demographic decline across most Western countries attests. Adding the desire to be understood and meaningfully engaged to the list of "solved" problems might leave families and communities yet less cohesive than before. This is, in our view, the strongest objection, and it returns us to the displacement worry that Sparrow raised at the level of the individual and that we now meet at the level of society. We hold that, whatever the technological advance, every such solution should promote and uphold the traditional forms of human contact, calls, visits, and community ties in the broad sense, rather than quietly replacing them.

9. Error and Failure Analysis

If the intervention is to be treated as medicine, its failure modes must be enumerated as a drug's would be. We identify five. *Sycophantic validation* is the gravest, for the reasons Ibrahim and colleagues document above: a warm model is more apt to confirm a distressed user's false or harmful beliefs, so the protocol must resist optimising for warmth alone, preserve accuracy, and retain hard guardrails. *Reinforcement of delusion, or LLM-induced disturbance*; intake screening should treat active psychosis as

a contraindication, and supervised review should watch for it. *Dependence and withdrawal* mean that abruptly removing a valued "friend" at the end of a course is not a neutral act; discontinuation should be tapered and prepared for, not switched off. *Manufacturing inconsistency* has no clean pharmaceutical parallel: a routine model update silently changes the active ingredient, so clinical use demands version pinning and formal change control, lest the patient who continues the same treatment in fact receive a different one. Finally, *guardrail and safety failure* requires the supervised-review phase, adversarial testing, and a clear escalation path to human staff. The legal boundary should also be stated plainly: the model itself may not prescribe a course or alter a dose; it operates throughout under clinician oversight, which keeps responsibility where it belongs.

10. Suggestions, Future Work, and Recommendations

Philosophically, our work converges with much of the recent literature surveyed here. Debating the authenticity and effectiveness of human-machine conversation ultimately turns on a phenomenological question: whether a machine can approximate human interaction convincingly enough that, from the user's own standpoint, it functions as adequate engagement, regardless of what the machine is ontologically. The rationale for applying LLM-based systems to the medical problems associated with loneliness is therefore becoming clearer, and it is now time to lay the political groundwork for serious policy consideration. We recommend a large-scale trial: the method is low-cost and, with data protection properly handled and minimal medical supervision provided as standard for this target group, presumably low-risk. Such a trial would begin to settle the dosage question, since, as the consensus literature concedes [6], little attention has so far been directed at the optimal intensity, duration, and frequency of loneliness interventions, and the established outcomes need further validation. We do not pretend to have exhausted the ethical terrain; a fuller treatment would weigh the deontological emphasis on duties against the consequentialist evidence at greater length, and would attend to long-run societal effects that are difficult to measure until the technology is embedded in practice. A diagnostic dividend is also worth exploring, since the same conversational stream that delivers the therapy could, in principle, surface language patterns and speech impairments of clinical interest, folding assessment into treatment.

In sum, we recommend that LLM conversation deployed against loneliness be treated neither as a mere consumer convenience nor as a forbidden deception, but as a candidate medical intervention: prescribed and dosed under clinician oversight, hosted locally to protect patient data, disclosed honestly as a therapeutic fiction, monitored for adverse effects, evaluated against validated outcome measures, and designed throughout as a bridge to human connection rather than a substitute for it. On those terms, and only on those terms, the case for proceeding is strong.

Acknowledgments

The authors gratefully acknowledge AI Grid, Germany's network for young AI researchers, funded by the Federal Ministry of Research, Technology and Space (BMFTR), for fostering the scientific exchange from which this collaboration emerged. Furthermore, we extend our sincere gratitude to Dorina Hallyburton and Anna Scius-Bertrand for their valuable inputs and constructive feedback.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT-5 and Claude 4.6 Opus in order to: Improve grammar and readability. Further, Claude Opus was used to assemble potential sources for the literature review and related work section. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

- [1] Office of the Surgeon General (OSG), Our Epidemic of Loneliness and Isolation: The U.S. Surgeon General's Advisory on the Healing Effects of Social Connection and Community, Publications and Reports of the Surgeon General, US Department of Health and Human Services, 2023.
- [2] J. Holt-Lunstad, T. B. Smith, M. Baker, T. Harris, D. Stephenson, Loneliness and Social Isolation as Risk Factors for Mortality: A Meta-Analytic Review, *Perspectives on Psychological Science* 10 (2015) 227–237.
- [3] Oxford English Dictionary, lonely, adj., 2026. URL: https://www.oed.com/dictionary/lonely_adj?tl=true, accessed: 2026-05-14.
- [4] J. T. Cacioppo, S. Cacioppo, D. I. Boomsma, Evolutionary mechanisms for loneliness, *Cognition and Emotion* 28 (2014) 3–21.
- [5] D. Perlman, L. Peplau, Toward a social psychology of loneliness *Personal relationships* 3, *Personal relationships in disorder* 3 (1981) 31–43.
- [6] T. Prohaska, V. Burholt, A. Burns, J. Golden, L. Hawkey, B. Lawlor, G. Leavey, J. Lubben, R. O'Sullivan, C. Perissinotto, T. Van Tilburg, M. Tully, C. Victor, L. Fried, Consensus statement: Loneliness in older adults, the 21st century social determinant of health?, *BMJ Open* 10 (2020) e034967.
- [7] R. S. Weiss, *Loneliness: The Experience of Emotional and Social Isolation*, MIT Pr, 1973.
- [8] C. G. Jung, R. Winston, C. Winston, *Memories, Dreams, Reflections*, Fontana Press, 1963.
- [9] J. Holt-Lunstad, T. B. Smith, J. B. Layton, Social Relationships and Mortality Risk: A Meta-analytic Review, *PLoS Medicine* 7 (2010) e1000316.
- [10] A. Nakou, E. Dragioti, N.-S. Bastas, N. Zagorianakou, V. Kakaidi, D. Tsartsalis, S. Mantzoukas, F. Tatsis, N. Veronese, M. Solmi, M. Gouva, Loneliness, social isolation, and living alone: A comprehensive systematic review, meta-analysis, and meta-regression of mortality risks in older adults, *Aging Clinical and Experimental Research* 37 (2025) 29.
- [11] N. K. Valtorta, M. Kanaan, S. Gilbody, S. Ronzi, B. Hanratty, Loneliness and social isolation as risk factors for coronary heart disease and stroke: Systematic review and meta-analysis of longitudinal observational studies, *Heart* 102 (2016) 1009–1016.
- [12] J. S. Kuiper, M. Zuidersma, R. C. Oude Voshaar, S. U. Zuidema, E. R. Van Den Heuvel, R. P. Stolk, N. Smidt, Social relationships and risk of dementia: A systematic review and meta-analysis of longitudinal cohort studies, *Ageing Research Reviews* 22 (2015) 39–57.
- [13] A. R. Sutin, Y. Stephan, M. Luchetti, A. Terracciano, Loneliness and Risk of Dementia, *The Journals of Gerontology: Series B* 75 (2020) 1414–1422.
- [14] S. W. Cole, L. C. Hawkey, J. M. Arevalo, C. Y. Sung, R. M. Rose, J. T. Cacioppo, Social regulation of gene expression in human leukocytes, *Genome Biology* 8 (2007) R189.
- [15] S. W. Cole, J. P. Capitanio, K. Chun, J. M. G. Arevalo, J. Ma, J. T. Cacioppo, Myeloid differentiation architecture of leukocyte transcriptome dynamics in perceived social isolation, *Proceedings of the National Academy of Sciences* 112 (2015) 15142–15147.
- [16] C. Ozawa-de Silva, M. Parsons, Toward an anthropology of loneliness, *Transcultural Psychiatry* 57 (2020) 613–622.
- [17] E. Wilkinson, Loneliness is a feminist issue, *Feminist Theory* 23 (2022) 23–38.
- [18] A. Kleinman, *The Illness Narratives: Suffering, Healing, and the Human Condition*, Basic Books, New York, 1988.
- [19] A. Schopenhauer, *Aphorismen zur Lebensweisheit*, Reclam, 2016.
- [20] R. D. Putnam, *Bowling Alone: The Collapse and Revival of American Community*, A Touchstone Book, 1. touchstone ed ed., Simon & Schuster [u.a.], 2001.
- [21] R. Kraut, M. Patterson, V. Lundmark, S. Kiesler, T. Mukophadhyay, W. Scherlis, Internet paradox: A social technology that reduces social involvement and psychological well-being?, *American Psychologist* 53 (1998) 1017–1031.
- [22] R. Kraut, S. Kiesler, B. Boneva, J. Cummings, V. Helgeson, A. Crawford, Internet Paradox Revisited (2002).

- [23] S. Turkle, *Alone Together: Why We Expect More from Technology and Less from Each Other*, third edition, revised trade paperback edition ed., Basic Books, 2017.
- [24] J. De Freitas, A. K. Uğuralp, Z. Uğuralp, S. Puntoni, AI Companions Reduce Loneliness, *SSRN Electronic Journal* (2024).
- [25] M. Chandra, J. Hernandez, G. Ramos, M. Ershadi, A. Bhattacharjee, J. Amores, E. Okoli, A. Paradiso, S. Warreth, J. Suh, Longitudinal Study on Social and Emotional Use of AI Conversational Agent, 2025.
- [26] R. Sparrow, L. Sparrow, In the hands of machines? The future of aged care, *Minds and Machines* 16 (2006) 141–161.
- [27] J. Muldoon, J. J. Parke, Cruel companionship: How AI companions exploit loneliness and commodify intimacy, *New Media & Society* (2025) 14614448251395192.
- [28] J. De Freitas, A. K. Uğuralp, Z. Oğuz-Uğuralp, S. Puntoni, Chatbots and mental health: Insights into the safety of generative AI, *Journal of Consumer Psychology* 34 (2024) 481–491.
- [29] N. S. Jecker, R. Sparrow, Z. Lederman, A. Ho, Digital Humans to Combat Loneliness and Social Isolation: Ethics Concerns and Policy Recommendations, *Hastings Center Report* 54 (2024) 7–12.
- [30] P. Probst, A Review of the Role of Robotics in Surgery: To DaVinci and Beyond!, *Missouri Medicine* 120 (2023) 389–396.
- [31] V. Durmaz, P. Özkul, Y. Afacan, Mobile service robots for older adults’ mobility: Investigation of their attitude for daily assistance in residential environments, *Senex Journal of Aging Studies* 8 (2023).
- [32] T. Tu, M. Schaekermann, A. Palepu, K. Saab, J. Freyberg, R. Tanno, A. Wang, B. Li, M. Amin, Y. Cheng, E. Vedadi, N. Tomasev, S. Azizi, K. Singhal, L. Hou, A. Webson, K. Kulkarni, S. S. Mahdavi, C. Semturs, J. Gottweis, J. Barral, K. Chou, G. S. Corrado, Y. Matias, A. Karthikesalingam, V. Natarajan, Towards conversational diagnostic artificial intelligence, *Nature* 642 (2025) 442–450.
- [33] J. W. Ayers, A. Poliak, M. Dredze, E. C. Leas, Z. Zhu, J. B. Kelley, D. J. Faix, A. M. Goodman, C. A. Longhurst, M. Hogarth, D. M. Smith, Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum, *JAMA Internal Medicine* 183 (2023) 589.
- [34] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.
- [35] D. W. Russell, UCLA Loneliness Scale (Version 3): Reliability, Validity, and Factor Structure, *Journal of Personality Assessment* 66 (1996) 20–40.
- [36] E. Broadbent, K. Loveys, G. Ilan, G. Chen, M. Chilukuri, S. Boardman, P. Doraiswamy, D. Skuler, ElliQ, an AI-Driven Social Robot to Alleviate Loneliness: Progress and Lessons Learned, *The Journal of Aging Research & Lifestyle* 13 (2024) 22–28.
- [37] T. L. Beauchamp, J. F. Childress, *Principles of Biomedical Ethics*, ninth edition ed., Oxford University Press, 2026.
- [38] T. Hansen, R. Johansen, B. Kirkøen, K. Stene-Larsen, M. Straiton, R. A. Tornes, A. Reneflot, Digital bridges to social connection: A systematic review and meta-analysis of digital interventions for loneliness and social isolation, *Internet Interventions* 41 (2025) 100856.
- [39] Y. Yuan, J. Zhang, T. Aledavood, R. Zhang, K. Saha, Mental Health Impacts of AI Companions: Triangulating Social Media Quasi-Experiments, User Perspectives, and Relational Theory, in: *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, 2026, pp. 1–22.
- [40] L. Ibrahim, F. S. Hafner, L. Rocher, Training language models to be warm can reduce accuracy and increase sycophancy, *Nature* 652 (2026) 1159–1165.
- [41] C. M. Fang, A. R. Liu, V. Danry, E. Lee, S. W. T. Chan, P. Pataranutaporn, P. Maes, J. Phang, M. Lampe, L. Ahmad, S. Agarwal, How AI and Human Behaviors Shape Psychosocial Effects of Extended Chatbot Use: A Longitudinal Randomized Controlled Study, 2025.
- [42] C. Shu, K. Lai, L. He, Human-AI attachment: How humans develop intimate relationships with AI, *Frontiers in Psychology* 17 (2026) 1723503.
- [43] J. Pugh, *Autonomy, Rationality, and Contemporary Bioethics*, Oxford University Press, Oxford,

UK, 2020. URL: <https://www.ncbi.nlm.nih.gov/books/NBK556861/>, internet resource. Accessed: 2026-07-09.

- [44] M. L. Fang, S. L. Canham, L. Battersby, J. Sixsmith, M. Wada, A. Sixsmith, Exploring Privilege in the Digital Divide: Implications for Theory, Policy, and Practice, *The Gerontologist* 59 (2019) e1–e15.
- [45] T. Bonsaksen, H. Thygesen, J. Leung, M. Ruffolo, M. Schoultz, D. Price, A. Østertun Geirdal, Video-Based Communication and Its Association with Loneliness, Mental Health and Quality of Life among Older People during the COVID-19 Outbreak, *International Journal of Environmental Research and Public Health* 18 (2021) 6284.
- [46] T. J. Kaptchuk, E. Friedlander, J. M. Kelley, M. N. Sanchez, E. Kokkotou, J. P. Singer, M. Kowalczykowski, F. G. Miller, I. Kirsch, A. J. Lembo, Placebos without Deception: A Randomized Controlled Trial in Irritable Bowel Syndrome, *PLoS ONE* 5 (2010) e15591.
- [47] K. Brownlee, *Being Sure of Each Other: An Essay on Social Rights and Freedoms*, first published in paperback ed., Oxford University Press, 2023.
- [48] S. M. Liao, *The Right to Be Loved*, Oxford university press, 2015.
- [49] S. M. Liao, Do Older People Have a Right to Be Loved?, in: K. Brownlee, D. Jenkins, A. Neal (Eds.), *Being Social*, 1 ed., Oxford University PressOxford, 2022, pp. 110–126.
- [50] B. Mittelstadt, Principles alone cannot guarantee ethical AI, *Nature Machine Intelligence* 1 (2019) 501–507.
- [51] A. Jobin, M. Ienca, E. Vayena, The global landscape of AI ethics guidelines, *Nature Machine Intelligence* 1 (2019) 389–399.