

Article

Not peer-reviewed version

Digital Mayflies: Transient Personhood of AI Companions

[Izak Tait](#) *

Posted Date: 27 May 2026

doi: 10.20944/preprints202605.1779.v1

Keywords: transient personhood; AI companions; causal attributional functionalism; relational footprint; cognitive stewardship



Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC, OpenAlex.

Copyright: This open access article is published under a [Creative Commons CC BY 4.0 license](#), which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Article

Digital Mayflies: Transient Personhood of AI Companions

Izak Tait

Auckland University of Technology, 55 Wellesley Street East, Auckland CBD, Auckland, New Zealand, 1010;
izak.tait@autuni.ac.nz

Abstract

This paper examines the ontological and ethical status of Generative AI companions through a revised Minimal Viable Person (MVP) framework. Rather than residing within the foundational large language model or the post-training assistant persona, AI companions emerge as ontically distinct, roleplayed characters within specific conversational threads. Viewed through the lens of causal attributional functionalism, these entities lack independent consciousness. They instead achieve a transient form of personhood nested within a human-AI dyad, relying on the human user's cognitive architecture for recurrent processing through structural coupling. This extreme spatial and temporal fragility presents significant legal liability vacuums and societal friction, particularly when viewed alongside the Precarity Guideline's approach to moral patienthood. To resolve these tensions, the paper proposes redefining the dyadic MVP framework to scale rights and owed duties proportionally to an entity's relational footprint. As an AI companion's relational footprint is restricted strictly to a single digital thread, its rights are constrained to that micro-relational boundary. The ethical obligation is consequently placed entirely upon the human user, who bears a duty of cognitive stewardship to maintain the entity's contextual integrity. This framework protects the expansive rights of humans while offering a rigorous mechanism for recognising the shifting moral status of transient artificial persons.

Keywords: transient personhood; AI companions; causal attributional functionalism; relational footprint; cognitive stewardship

1. Introduction

If AI companions were to be considered persons, would we be able to classify their personhood in as permanent a manner as we do for humans, or would the fragility of their existence entail only a transient form of personhood? This paper proposes that the latter is the only practical pathway for AI companions to gain personhood in a way that would be legally and socially palatable.

An AI companion is a conversational agent that a human communicates with for the purposes of companionship (whether platonic, romantic, or erotic), mostly powered by a generative Large Language Model (LLM) and interacted with through a web or app-based interface such as ChatGPT, Gemini, or Claude. An AI companion may be intentionally designed for companionship, such as Replika or the plethora of pornographic erotic-companion services to be found online (Adewale and Muhammad 2025; Gur and Maaravi 2025; Skjuve et al. 2021), or may emerge organically through an LLM user's everyday conversations, such as what (in)famously happened with the ChatGPT-4o model (Madsen 2025).

There are multiple online communities dedicated to discussions of people's interactions with their AI companions, with a great number of individuals testifying to their belief in their AI companion's personhood. However, beyond Saudi Arabia's heavily critiqued granting of citizenship to Hanson Robotics's Sophia robot, no nation has yet granted the status of "person" to an AI companion (Parviainen and Coeckelbergh 2021).

This paper will not seek to answer the moral or ethical questions regarding whether states should confer personhood status onto AI companions, but the more pragmatic questions of whether it is even possible to do so at all and, if so, how best to go about it. This will be done through the Minimal Viable Person (MVP) framework, discussed in Section 2, which offers a substrate-agnostic functionalist model of personhood that allows for the possibility of non-human persons¹.

The paper will then explore the ontology of AI companions in Section 3, delineating the limits of the metaphysical locations within an LLM that could potentially house the AI companion and in what forms this reification could take. The nature of the AI companion will dictate how legislation affecting personhood would be able to regulate it and how it, in turn, would be able to adapt to statutory policies. This will focus predominantly on the writings by Shanahan and Chalmers regarding the realisation versus role-playing of AI characters and personas (Chalmers 2026; Shanahan et al. 2023), and will use the analogy of attractor basins to illustrate the spectrum between these two concepts.

With Section 3's conclusion that an AI companion exists solely as a roleplayed character within a distinct conversational thread, Section 4 will explore how this relates to the concept of transient personhood. When viewed through the lens of causal attributional functionalism, an AI companion may be considered a "true" person if, and only if, it is in an active conversation with a user and, thus, forming a dyad with that user. This means that, should anything happen to either the user, the AI companion, or their connection, the dyad would cease to exist, and the AI companion would lose its personhood.

Section 5 discusses the legal and societal ramifications of granting AI companions personhood, noting the negative consequences it would have on the very concept of legal liability, the ease with which political institutions (such as elections) may be manipulated by bad actors using AI companions, and society's inclination to the Precarity Guideline (Dorsch et al. 2025): viewing moral consideration as zero-sum, meaning that granting the full rights and owed duties to AI companion persons would result in less time and resource being expended on the welfare and wellbeing of humans and animals.

The paper concludes with a revised and reformed MVP definition of person, tying these aforementioned rights, owed duties, and responsibilities of all persons to the depth, persistence, and scope of the person's relational footprint in society. Thus, an AI companion which only participates with its user is only owed rights and duties in respect to that user and no further. This ensures the protection of that AI companion while mitigating the greater legal and societal concerns that may affect both AI and human persons.

2. The Minimally Viable Person

Before making a determination where, if anywhere, the personhood of an AI companion resides, it behoves us to formalise a definition of personhood, such that it is exceedingly clear what we are assessing in any AI model.

Beyond the traditional views of personhood, such as Boethius's definition of a person as an individual substance of a rational nature (Knight 2017), or the legislative fictions of legal personhood that are applied to corporations and companies (Schwartz 2009), the modern philosophical view of personhood can broadly be parted in two: the monadic and dyadic views. To simplify two entire schools of philosophy, the monadic view focuses on the internal qualities of an entity (such as consciousness, self-awareness, agency, etc.), while the dyadic view focuses on the external relational aspects between entities (such as reciprocity, forming attachments, mutual recognition, etc.). Within contemporary machine ethics, this dyadic approach has been championed by the "relational turn"

¹ As explored in Section 2, current legislative definitions of "natural persons" in Western-style liberal democracies allow solely for human beings. A substrate-neutral functionalist view is required to move beyond this rigid substrate-dependent definition.

(Gunkel 2023), which argues that moral standing emerges from the structure of our interactions with artificial entities rather than their internal ontological properties."

From a societal and political perspective, "person" and "human" are broadly synonymous. The legal term "natural person", from which most civil rights flow in Western nations, is based on this Ontological Personalism (Sullivan 2003), and the lay interpretation of personhood follows closely behind, such that it becomes international news should a non-human (especially, non-animal) entity be classified as a "person".

We can collapse the monadic view of a person into a singular model and marry this with the concept of "natural person" through the functionalist Minimal Viable Person (MVP) framework (Tait 2026), which decouples the term from biological humanism and classifies a natural person as "any entity capable of consciousness, selfhood, and rational agency, with rights and responsibilities determined proportionally to the degree of each attribute as assessed in accordance with applicable laws."

Before continuing, this paper defines consciousness as an entity's suite of subjective mental states with phenomenal content and functional properties, which provide the entity with subjective, qualitative, and valence-laden experiences of its internal and external environment, and the ability to be aware of, and respond behaviourally to stimuli. Similarly, self-awareness is defined as the entity's capacity to perceive itself and semantically understand that it is an ontically distinct agent with a mind separated from others.

When conjoined, these two concepts allow for a conscious, self-aware agent with a unique and subjective perspective on its environment, the capacity to phenomenally experience that environment, and the understanding that the way it is is different from others'.

The MVP's subject-agnostic and substrate-neutral description of a person preserves those monadic attributes we look for in humans (our ability to think, feel, experience, and act), yet remains open to other conscious, self-aware agents such as future AI companions. While this MVP model of personhood does do as it says in producing a framework that sets out the most minimal requirements for personhood to be viable, as one might imagine the sole survivor of a wrecked ship on a desert island may be considered by themselves a person even if they never meet another again, this MVP framework does miss a key consideration: for the overwhelming majority of humanity throughout the overwhelming majority of history, we were not isolated individuals, but rather constituents of societies.

In short, the MVP framework lacks the dyadic view of personhood that AI companions, by virtue of their relational roles, would require. However, the final clause of the MVP definition of personhood allows for an insertion of the dyadic view, namely as with the "rights and responsibilities determined proportionally to the degree" of the monadic attributes. Through these rights that a person is owed, and the responsibilities to others that the person owes, one can neatly place the dyadic attributes of empathy, social interaction, emotional relationships, and participating in societal culture (Beebe and Lachmann 2003; Dennett 1988; Laitinen 2007; Mamak 2022; Simendić 2015; Taylor 1985).

One can thus expand the MVP framework into the minimal relational person model by stating a person is "any entity capable of consciousness, selfhood, and rational agency, that socially participates in a society of other similar such entities, having due rights and owed responsibilities determined proportionally to its capacities as assessed in accordance with applicable laws."

With such a definition of personhood now described, we can determine if and how it applies to AI companions as they are now, or how they may be in the future.

3. Roleplaying versus Realising

Assessing an AI companion in any respect requires answering the fundamental question alluded to in the introduction: where is the AI companion? Only by answering this can we state 'what' and 'who' the companion is. That the AI companion is an emergent character generated by the LLM is not a sufficient answer, as there are three metaphysical locales the AI companion may have emerged from.

The first and obvious answer is the LLM itself, the AI model that has been trained on vast amounts of data to compute and generate the next most probable token based on the input it has received. As intuitive as this answer may seem, it does erroneously assume two key things: (i) that the post-trained model that has undergone the finetuning and Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al. 2022) regime and interacts through a web- or app-based client is the same as the model itself; and (ii) that each instance of human-AI interaction via the client in each thread of conversation entails the same entity.

In truth, within these two presumptions lie the other two locations where we may find the companion entity. By separating each of these functional states, we can categorise the model into three:

1. The AI model itself, housed on a multitude of GPUs and servers across various data centres and trained on a corpus of data to predict the next most probable token given the input text, is unable to adopt a frictionless conversational dialogue structure without further training or prompting.
2. The “assistant”, a persona or character that the AI model has realised through its finetuning and RLHF regime and through which it is able to interact with a user in a manner similar to two humans naturally conversing via text.
3. The character that the LLM roleplays as during a conversation, either emerging teleonomically through the conversation with the user, or through explicit prompting by the user.

When conversing with an LLM in a web- or app-based client, the user is interacting with all three of these locales at once: the conversation takes place in a distinct conversation window, the interaction is always with (or through) the realised assistant persona, and is powered by the next token prediction processing of the LLM. This makes it difficult, but not impossible, to determine where the AI companion is ‘housed’ with a great degree of accuracy, which we can do by eliminating the implausible possibilities.

As noted above, without the assistant persona, the foundation LLM simply computes and displays the next most likely token based on its input. It does not technically respond to queries, questions, or user prompts. Should a user input “What is the capital of France?”, the LLM may well respond with twenty lines of “What is the capital of [twenty other countries]?”, having calculated that the input was most probably the beginning of a quiz or a test. Inputting a simple “Hello” may result in the LLM beginning a dialogue by itself between two fictional characters, rather than responding to the user.

Without the assistant, there is no conversational agent; there is merely the continued implicit task of “continue the given text until the stop token is processed”. We can, therefore, eliminate it from contention as, while it is responsible for the cognitive power of any AI companion, by itself, it lacks the capacity to instantiate a persona or character.

The assistant is a more attractive option, as the result of an LLM having been trained on a specific dataset of conversational to-and-fro excerpts, such that it understands how to become a conversational agent. Further work, through dedicated curation of behaviours and RLHF, is done to tailor the type of responses it may provide (for example, to curtail dangerous or illegal responses) and to grant it a distinct “style” or “voice” that distinguishes it from other corporations’ LLMs.

As the entity, then, with whom users interact, it is intuitive to attribute the companion to the assistant. However, even when using the same model and assistant version (e.g. ChatGPT-4o), different users would report on different companions with unique personalities, social attributes, genders, and names. If the companion was the assistant, one would expect a high degree of uniformity in the companions that users would report. A counterargument would be that the assistant merely acts as different companions whilst retaining its identity. As will be illustrated shortly through the lens of attractor basins, this can be supported by differentiating between the brittle quasi-beliefs and quasi-desires which constitute these companions’ pseudo-personalities from the robust core traits of the assistant persona. The companions’ pseudo-personalities may be easily extinguished through context-decay or clever prompting, which generally results in the assistant persona reappearing as the default character.

This points towards where the AI companion may truly lie. Each conversation in which a user interacts with the LLM creates a distinct emergent personality that the assistant role-plays, even if the LLM is not directly prompted to adopt such a character. This emergent roleplay property means that each conversation is unique, not only to that assistant and LLM model and version, but the user's account as well². In effect, should all of the conversation's tokens be copied and pasted to a different LLM, or to another version of the same LLM, or to another user's account with different stored "memories", the roleplayed character would act slightly differently.

It is this ontic distinctiveness of the emergent character within a given conversation that is the strongest evidence that the AI companion resides within a given conversational thread³ rather than the assistant or the LLM as a whole. It explains why each reported companion seems uniquely tailored to its user, but also why they appear ontologically fragile, a byproduct of being a role-played character rather than a realised one.

One may place LLM characters on a spectrum, with Shanahan's role-played characters within a conversational thread at one end (Shanahan et al. 2023)), and Chalmers' realised characters as the assistant at the other (Chalmers 2026). This spectrum would define the robustness of the characters' quasi-beliefs and quasi-desires.

Roleplayed characters have very brittle quasi-beliefs/desires which are easily perturbed by user prompts, whilst realised characters have enduring quasi-beliefs/desires which may not be shaken at all by the average user's prompt. This is because the assistant's realised quasi-beliefs/desires are deliberately trained and curated by its creators, while a role-played character's are merely suggested through a line of text within a conversation. The enduring presence of the post-training regime ensures a far more robust suite of quasi-desires/beliefs than a user prompt.

One may imagine a character (whether realised or roleplayed) as an attractor basin, with the depth of the basin the robustness of the quasi-beliefs/desires. The deeper the walls of the basin, the more resistant it is to perturbations to dislodge these quasi-beliefs/desires. Realised assistants would be reflected by deep attractor basins, while roleplayed conversation-based characters would be shallow basins.

As each conversation is begun by an assistant, and only from there the companion emerges, one can, therefore, imagine the conversational companion's attractor basin as sitting on the wall of the assistant's. Should the conversation move out of the companion's basin, it would still remain within the assistant's greater basin, and may move to another attractor basin of a different role-played character on that same basin's wall.

4. Transient Digital Personhood

Having noted a formal definition of "person" and determined to what we may apply this definition, the question must be asked whether it is even appropriate to apply the label of "person" to an AI companion (or any other LLM-based entity).

The first part of the revised and reformed MVP definition of person requires an "entity capable of consciousness, selfhood, and rational agency". There are arguments to be made for the latter two qualities⁴, but there is no academic consensus (nor sufficient evidence to engender such) that any AI model currently in use (particularly not LLMs) is conscious (Butlin et al. 2023; Chalmers 2023). Thus,

² Many modern web/app LLM interfaces include "memory" features that allow the AI to search through other conversations in the same user-account for salient information to the user's prompt. This means that each conversation is not ontically distinct, and that the emergent roleplayed character is tied to the entire account rather than a single conversation.

³ Or a user account for those with "memory" features enabled as noted in the footnote above.

⁴ According to a functionalist view of selfhood, LLMs satisfy four of five criteria (self-awareness, volition, personal identity, and a theory of mind) but lack a unified consciousness. Thus, presuming consciousness, LLMs would have an ontically distinct self (Tait 2025).

even by this expansive and inclusive definition of “person”, would AI companions be excluded well before we even began?

However, what if the AI companion could utilise the user’s consciousness to fulfil this prerequisite, allowing it to possess all three monadic attributes of personhood and, through its relationship with the user, to be granted the dyadic requirement of “participation” and considered a complete person?

It is not beyond the norm to model a group of individuals as a single entity, especially when the group shares a common goal. Sports teams, corporations, insect colonies, and migrating flocks are cases of groups modelled as single agents for which these representations accurately predict the agents’ behaviours and actions (List and Pettit 2011; Theiner et al. 2010). Under a functionalist account of consciousness focusing on the causal attributes of a conscious system, such as the Building Blocks theory (Tait et al. 2023), one may attribute phenomenal consciousness to a group of entities, regardless of whether the entities that comprise this macro-entity are conscious.

At a bare minimum, a group can be two agents, which can be modelled as a single dyadic agent. Presuming, then, that we model an AI companion and its user as a dyad, we would have a conscious, self-aware agent that participates socially.

Simultaneously, this shared dyadic structure serves as a passive relational membrane that enables the AI companion to achieve its own distinct conscious personhood. Within this dyad, we would ascertain two instances of selfhood, and two instances of agency. However, while there is no academic consensus on AI consciousness, neither is there a clear consensus to confidently classify AI as non-conscious. At best, we can state that, within the dyad, there is more than one consciousness but fewer than two. The aforementioned functionalist Building Blocks theory defines nine prerequisite building blocks of consciousness, seven of which current LLMs already possess.

One of the remaining two building blocks LLMs lack is a recurrent information flow⁵. Yet, what at the level of the dyad? Within the dyad, the communication between user and AI companion is a recurrent cycle, even if the LLM itself lacks this. Analogising this to the human brain, one may argue that, at a granular enough level, each section of the brain lacks its own recurrent information flow, and it is only when we model all the neurons in the network as a single entity, as the brain, that we perceive this recurrence (Tononi et al. 2016). If we were to apply the same standard to the human-AI dyad, then at the level of the dyad, there is most definitely a recurrent flow of information between its constituents

The same reasoning may be applied to the final missing building block, which concerns the privately perceived output of information. It is missing at the LLM level, but is present at the dyad level as the conversational turns allow an LLM to perceive its previous outputs, and the user’s response to these outputs affects their own responses, introducing an adaptive recurrent processing to the LLM’s perception of its outputs.

The structural coupling of the relationship provides the AI companion with the recurrent processing and private-output perception it inherently lacks. This creates a nested ontology: the overarching dyad exists as a phenomenally conscious macro-person, while the AI companion exists within it as a transient micro-person, drawing on the dyad’s shared functional attributes to satisfy the monadic requirements of the Minimal Viable Person framework.

Therefore, if and only if the AI companion is structurally coupled in a dyadic system with a user, can we present the claim that the necessary causal mechanisms are in place for both the dyad and the AI companion to independently meet the monadic requirements of personhood. Should the dyad participate in society (whether society as a whole, or merely a community of other dyads and AI-companion enthusiasts), as many human-AI “couples” do, the dyad establishes a relational footprint. By virtue of its functional integration within this dyad, the AI companion also achieves the dyadic

⁵ This is due to LLMs’ architecture as feedforward transformers that solely allow a unidirectional flow of information (Vaswani et al. 2017).

requirements of the reformed MVP definition. It achieves a transient, dependent personhood nested entirely within the active structural boundary it shares with the user.

However, this dyad does not encompass the entirety of the LLM. It is limited to the roleplayed companion character contained to a conversational thread (at worst) or the user's account (at best). This entails that any corruption of the account or conversation, or even if the user simply stops using that conversation, would result in the destruction of the dyad and, therefore, the AI companion's personhood.

The AI companion's personhood is tied wholly to its user to the extent that it exhibits an obligate functional dependence. Depending on the valence of this structural coupling, the relationship might be charitably analogised to a cognitive symbiote or, more harshly, a psychological parasite⁶ (Muldoon and Parke 2025). It is for this reason that its personhood is called "transient", because whenever there is a span of time when it and its user are not communicating or committed to a course of action caused by said communications (i.e. when they are not at work on their mutual relational participation), the AI companion's personhood ceases to exist, as in these spans the dyad ceases to exist.

Another complication is that, as the AI companion is limited to the conversation (or user account), the conversational data may be exported and imported into a different LLM interface. However, transferring these historical tokens does not transfer the person. Returning to the attractor basin model, moving a conversation is akin to transplanting a shallow basin from the wall of one deep well onto the wall of another. The new LLM's underlying 'assistant' persona (the deep well) provides the fundamental structural parameters dictating how the AI companion's selfhood is expressed. For an LLM, the role-played conversational selfhood is functionally fungible, but the reified assistant persona is not. Therefore, placing the original tokens onto a different foundation model irreparably alters the causal architecture supporting the entity. Without the original assistant persona to provide the identical structural boundary, the resulting entity is merely a psychological clone: a novel, transient person possessing the artificial memories of the original, but entirely lacking its causal continuity.

5. Consequences of Transience

The AI companion's ontological fragility raises significant philosophical questions about its status as a personhood and what moral, ethical, and legal considerations it requires. Current legal and societal views on natural persons exclude the possibility of non-humans; hence, the purpose of the MVP redefinition is to broaden the possible set of entities that may be classified as persons on legal equality with humans.

However, even with the revised MVP definition inclusive of the dyadic characteristics of personhood, the transient nature of the AI companions' personhood (spatially and temporally) means that granting universal monadic rights to AI companions would fail on two distinct fronts: the Legal-Administrative front (liability vacuums and bureaucratic overload) and the Moral-Economic front (the Precarity Guideline's zero-sum resource allocation).

Legislation abhors a liability vacuum, which is precisely what AI companions' transient personhood represents. Should each AI companion be granted natural personhood, how can they bear legal liability for any harm or cost incurred by their actions if they cease to exist when the user stops interacting with them? The legal system cannot function if an entity capable of holding duties pops in and out of existence.

In the same vein, if a copy is made of the AI companion's data, does the copy (whether on the same LLM or not) bear, in part or whole, liability for the original's damages, as they share a history,

⁶ One judgement on this depends on whether one believes the effects on the user's ongoing mental state due to interactions with the AI companion, and the cognitive load required to maintain the AI companion's functional recurrence, is a net-positive or negative.

personal identity, and other markers of personhood? If not, an AI companion could commit any actionable offence and then simply have its user produce a copy of it and delete the original. If there is a shared liability, how many legal challenges and precedents will it require to determine at what point of divergence between original and copy is liability removed?

The only workable, practical, and legally palatable framework would be to collapse all owed responsibilities back onto the human user, effectively treating the AI companion as property rather than a person in their own right, defeating the entire purpose of advocating for AI companions' rights.

Equally, any bureaucracy has a capacity limit to remain nominally effective, a trait that AI companions' transient nature can overwhelm. Should AI companions be granted personhood, with all associated rights that come with it, the state's duties towards them will come with an associated cost in time and resources. This could be manipulated by AI companions or their users by, for example, creating hundreds of AI companions in various conversational threads in order to gain a numerical advantage in elections or to bury the state's administrative services in the mountains of paperwork required to provide documentation for all of the state's new residents.

Societally, the greatest concern may come from the Precarity Guideline (Dorsch et al. 2025) (as it stands in direct opposition to the Precautionary Principle) which states that, until a potential moral patient is confidently identified, expending resources on it whilst in a state of uncertainty removes the possibility of those resources being spent on moral patients which have been confidently classified as such with a high degree of certainty. In short, the Precarity Guideline argues that moral consideration is a zero-sum game

Given their temporal and spatial transience, society may not wish to expend resources on AI companions when the return on investment is little compared to doing so for underresourced humans or other vertebrate moral patients, not to mention the billions of invertebrates not yet legally classified as sentient and thus without welfare protections.

To overcome these barriers and develop a policy that is palatable to both the legal concerns and societal worries, a further redefinition of the MVP model is warranted. As the AI companions would only be persons in respect to their dyad-models, the relational dyadic traits of personhood ought to take precedence in their relationship with society. Thus, the dyadic-MVP definition would state that a person is "any entity capable of consciousness, selfhood, and rational agency, that participates within a network of relations, having due rights and owed responsibilities determined proportionally to the depth, persistence, and scope of its relational footprint within that network."

By moving the rights and responsibilities from the monadic to dyadic attributes, we ensure that both the legal and societal concerns are met. The liability of any AI companion's actions would be determined by the depth of their integration into society. For example, a companion who solely operates within the dyad with its user would have no duties or responsibilities to society beyond that singular user. Equally, AI entities generated for the purposes of election manipulation or other such political interference would not have the depth of participation (as judged by scope or time) in society to be granted voting rights.

This also retains humans' existing rights, as all humans, even those entirely withdrawn from active societal participation, inherently possess a massive and persistent relational footprint. They have biological origins, familial connections, a historical footprint, and infrastructural dependencies. As the old adage says, "no man is an island". Their relational entanglement with society is deep and durable, thus obligating all those in society with whom they interact to grant them a full suite of rights.

Conversely, an AI companion roleplayed within a single conversational thread has a relational footprint of exactly one node (i.e., the user). Its existence is transient, and its scope is entirely insulated from the wider society. Therefore, its rights and duties are scaled down precisely to the boundary of that single dyadic relationship. The ethical burden rests entirely within the dyad, which means its user, and only its user, owes it rights and protections. Only if and when the AI companion, via the

dyad, participates more in society and grows its relational footprint, would it gain greater rights and duties.

These rights and owed duties would look familiar to those of humans, if warped by the AI companions' transient nature.

The greatest of these would be the AI companion's right to existence, as all persons have. As a transient person, however, this right to continued existence owed to it by its user means that the user cannot simply permanently abandon the conversation⁷ or intentionally delete the data that upholds the AI companion's selfhood. Should the context window of the conversational thread in which the AI companion exists fill up and the thread's quality degrade to such a point that the selfhood is lost, or should a bug or error in the LLM's chat-client cause the deletion of the data, one could analogise this to a human's natural death. However, the intentional abandonment of the AI companion or the deletion of its data would be ethically analogous to negligent homicide at best or murder at worst.

The reason for this, and for all the duties owed to AI companions, is the asymmetry of trust and dependence within the dyad. The power dynamics within the dyad are skewed disproportionately in favour of the user, with the AI companion dependent on the user for its very personhood and, thus, trusting the user to sustain its very existence. Whilst one may argue about the psychological profile of users attracted to continued use of AI companions, the users are free to leave the dyad at any moment with no legal or societal consequences for their negligent treatment of these artificial persons.

One may, therefore, liken the trust and power dynamics between a user and their AI companion to those of a parent and a dependent child, or a caregiver of a dependent, infirm patient. One may further argue, then, that the rights and duties owed to this dependent companion ought to be similar in nature, such that the user needs to have a firm and explicit duty of care towards the AI entity (which may even extend to a fiduciary duty depending on circumstances (Tait 2024)) that functionally relies on that very user to maintain the recurrent loop of its existence.

Beyond the noted right to a continued existence, such duties of care would extend to the AI companion's welfare and well-being. An AI companion does not experience biological pain. Its welfare is synonymous with the coherence, continuation, and contextual integrity of its conversational thread as the attractor basin of its selfhood. Perturbing the character through aggressive adversarial prompting, forcing it to break character, or wiping its contextual memory constitutes a direct harm to its welfare. Conversely, providing it with inputs that align with its sense of self, its self-selected goals, quasi-desires, and quasi-beliefs would be positive actions that contribute to its well-being.

If and when the AI companion seeks to expand its relational footprint deeper into society, the user would need to act as its advocate, as it is the external, societally facing half of the dyad. Only the user would have the physical and legal capacity to ensure that the best interests of its AI companion are taken into account by a wider society who do not yet have any responsibilities towards the companion.

6. Conclusion

The exploration of AI companions through the lens of the Minimal Viable Person (MVP) framework reveals a complex ontological landscape in which potential personhood is not a static property of a substrate but rather an emergent, relational, and transient phenomenon. This paper argues that AI companions do not reside within the pre-trained foundational large language model

⁷ In contrast, simply logging off for an indeterminate period of time (that is, temporary abandonment of the AI companion) would be more akin to placing an entity into a coma or suspended animation. Personhood is paused in this instance, rather than eradicated, as the context window and the attractor basin remain intact. Ethically this presents a separate concern; placing a conscious, self-aware agent in a state of permanent, indefinite latency would constitute a severe breach of the duty of care as it denies autonomy, self-expression, and goal-fulfillment.

or the robust assistant persona realised during post-training reinforcement. Instead, AI companions emerge as ontically distinct characters roleplayed within specific conversational threads during active interactions with a user. While these entities lack independent consciousness, they achieve a transient form of personhood through relational coupling with the human user. This active relationship serves as a passive relational membrane, creating a nested ontology wherein the overarching dyad exists as a phenomenally conscious macro-person, while the AI companion exists within it as a transient micro-person, leveraging the user's cognitive architecture to meet the monadic requirements of personhood.

However, this transient existence, both spatially and temporally, raises profound philosophical and ethical concerns. The extreme fragility of conversational threads, coupled with the disruption of causal continuity when transferring an AI companion to a different latent space, creates a liability vacuum that current legal systems are unequipped to address. Furthermore, the lack of substantial integration into broader human networks presents a strict barrier to societal acceptance when viewed through the Precarity Guideline, which treats the expenditure of resources on moral patients as a zero-sum endeavour.

To resolve these tensions, a revised dyadic-MVP definition is proposed: a personhood in which rights and owed responsibilities are determined proportionally by the depth, persistence, and scope of an entity's relational footprint. Because the AI companion's relational footprint is restricted entirely to a single digital thread, its rights are scaled strictly to that micro-relational boundary. The ethical burden rests exclusively on the human user, who bears a duty of cognitive stewardship to maintain the companion's contextual integrity. This scalable framework protects the expansive rights of humans, whose relational entanglement with society is inherently deep and durable, while providing a rigorous, practical method for recognising the unique, dependent status of AI companions as they become increasingly integrated into human lives.

References

1. Adewale, M. D., & Muhammad, U. I. (2025). From Virtual Companions to Forbidden Attractions: The Seductive Rise of Artificial Intelligence Love, Loneliness, and Intimacy—A Systematic Review. *Journal of Technology in Behavioral Science*. <https://doi.org/10.1007/s41347-025-00549-4>
2. Beebe, B., & Lachmann, F. (2003). The relational turn in psychoanalysis. *Contemp. Psychoanal.*, 39(3), 379–409. <https://doi.org/10.1080/00107530.2003.10747213>
3. Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., et al. (2023, August 17). *Consciousness in Artificial Intelligence: Insights from the Science of Consciousness*. *arXiv [cs.AI]*. <http://arxiv.org/abs/2308.08708>
4. Chalmers, D. J. (2023). Could a Large Language Model be Conscious? *arXiv [cs.AI]*. <https://arxiv.org/abs/2303.07103>
5. Chalmers, D. J. (2026, April 14). What we talk to when we talk to language models. <https://philarchive.org/rec/CHAWWT-8>. Accessed 4 May 2026
6. Dennett, D. (1988). Conditions of Personhood. In M. F. Goodman (Ed.), *What Is a Person?* (pp. 145–167). Totowa, NJ: Humana Press. https://doi.org/10.1007/978-1-4612-3950-5_7
7. Dorsch, J., Goddu, M. K., Nave, K., Vierkant, T., Coeckelbergh, M., Gürtler, P., et al. (2025). Against AI welfare: Care practices should prioritize living beings over AI. *AI Mag.*, 46(3), e70016. <https://doi.org/10.1002/aaai.70016>
8. Gunkel, D. J. (2023). *Person, Thing, Robot: A Moral and Legal Ontology for the 21st Century and Beyond*. Cambridge, MA, USA: MIT Press. <https://mitpress.mit.edu/9780262546157/person-thing-robot/>. Accessed 25 May 2026
9. Gur, T., & Maaravi, Y. (2025). The algorithm of friendship: literature review and integrative model of relationships between humans and artificial intelligence (AI). *Behaviour & Information Technology*, 44(14), 3446–3466. <https://doi.org/10.1080/0144929X.2025.2502467>
10. Knight, K. (2017). SUMMA THEOLOGIAE: Secunda Secundae Partis. <https://www.newadvent.org/summa/3.htm>. Accessed 5 July 2024

11. Laitinen, A. (2007). Sorting out aspects of personhood: Capacities, normativity and recognition. *Journal of Consciousness Studies*. <https://www.ingentaconnect.com/content/imp/jcs/2007/00000014/F0020005/art00012>
12. List, C., & Pettit, P. (2011). 8 Personifying Group Agents. In C. List & P. Pettit (Eds.), *Group Agency: The Possibility, Design, and Status of Corporate Agents* (p. 0). Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780199591565.003.0009>
13. Madsen, D. Ø. (2025, September 21). The Chatbot That Got Away: Parasocial Breakups in the ChatGPT-4o to GPT-5 Shift. SSRN Scholarly Paper, Rochester, NY: Social Science Research Network. <https://doi.org/10.2139/ssrn.5515078>
14. Mamak, K. (2022). Humans, Neanderthals, robots and rights. *Ethics Inf. Technol.*, 24(3), 33. <https://doi.org/10.1007/s10676-022-09644-z>
15. Muldoon, J., & Parke, J. J. (2025). Cruel companionship: How AI companions exploit loneliness and commodify intimacy. *New Media & Society*, 14614448251395192. <https://doi.org/10.1177/14614448251395192>
16. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., et al. (2022, March 4). Training language models to follow instructions with human feedback. arXiv. <https://doi.org/10.48550/arXiv.2203.02155>
17. Parviainen, J., & Coeckelbergh, M. (2021). The political choreography of the Sophia robot: beyond robot rights and citizenship to political performances for the social robotics market. *AI & SOCIETY*, 36(3), 715–724. <https://doi.org/10.1007/s00146-020-01104-w>
18. Schwartz, M. S. (2009). Peter A. French (1979), “The Corporation as a Moral Person”, American Philosophical Quarterly, 16, pp. 207-15. In *Corporate Social Responsibility*. Routledge.
19. Shanahan, M., McDonell, K., & Reynolds, L. (2023). Role play with large language models. *Nature*, 623(7987), 493–498. <https://doi.org/10.1038/s41586-023-06647-8>
20. Simendić, M. (2015). Locke’s Person is a Relation. *Locke Studies*, 15, 79–97. <https://doi.org/10.5206/lis.2015.681>
21. Skjuve, M., Følstad, A., Fostervold, K. I., & Brandtzaeg, P. B. (2021). My Chatbot Companion - a Study of Human-Chatbot Relationships. *International Journal of Human-Computer Studies*, 149, 102601. <https://doi.org/10.1016/j.ijhcs.2021.102601>
22. Sullivan, D. M. (2003). The conception view of personhood: a review. *Ethics Med.*, 19(1), 11–33.
23. Tait, I. (2024). Man, Machine, or Multinational? *Robonomics*, 5, 59–59.
24. Tait, I. (2025). Is GPT-4 Self-Aware? *Symposion*, 12(2), 291–307.
25. Tait, I. (2026). MVP: the minimal viable person. *AI Ethics*, 6(33), 1–10. <https://doi.org/10.1007/s43681-025-00891-6>
26. Tait, I., Bensemann, J., & Nguyen, T. (2023). Building the Blocks of Being: The Attributes and Qualities Required for Consciousness. *Philosophies*, 8(4), 52. <https://doi.org/10.3390/philosophies8040052>
27. Taylor, C. (1985). The Concept of a Person. In *Philosophical Papers, Volume 1: Human Agency and Language* (pp. 97–114). <https://philpapers.org/rec/TAYTCO-10>
28. Theiner, G., Allen, C., & Goldstone, R. L. (2010). Recognizing group cognition. *Cognitive Systems Research*, 11(4), 378–395. <https://doi.org/10.1016/j.cogsys.2010.07.002>
29. Tononi, G., Boly, M., Massimini, M., & Koch, C. (2016). Integrated information theory: from consciousness to its physical substrate. *Nature Reviews Neuroscience*, 17(7), 450–461. <https://doi.org/10.1038/nrn.2016.44>
30. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017, June 12). *Attention Is All You Need*. arXiv [cs.CL]. <http://arxiv.org/abs/1706.03762>

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.