# Vulnerable digital minds

**Soenke Ziesche**
Independent researcher
Brooklyn
USA
soenke.ziesche@gmail.com

**January 2025**

# DRAFT

"Not insignificant that we've seen researchers at Anthropic (last week) and Google (this week) co-authoring serious papers about the possibility of sentience/pain/pleasure in AI. This would once have got you fired, but the conversation has moved on."
Jonathan Birch (7 Nov 2024)[1]

Abstract ..... 2
Introduction ..... 2
How is vulnerability defined? ..... 2
Who are vulnerable humans? ..... 3
What are VDMs? ..... 5
Why should we care about VDMs? ..... 6
Discrimination against VDMs ..... 6
Abuse of VDMs ..... 7
Proposed research questions ..... 7
How could we find out if digital minds are discriminated against or abused? ..... 7
Could vulnerabilities of digital minds be reversed or cured? ..... 8
How could VDMs become resilient? ..... 8
Conclusion ..... 10
Recommendations to protect VDMs ..... 10
References ..... 12

[1] https://x.com/birchlse/status/1854498736761709010

# Abstract

AI welfare science explores the ethical implications of digital minds potentially possessing moral status. Moral principles are especially important when considering the welfare of weaker or more susceptible members of a society. The purpose of this paper is to describe which potential digital minds deserve special moral consideration as well as what types of risks these minds are exposed to. We call these minds "vulnerable digital minds" or "VDMs" and have identified as main risks for them discrimination as well as abuse. With the overall goal to protect vulnerable digital minds the following research questions are discussed: How could we find out if digital minds are discriminated against or abused? Could vulnerabilities of digital minds be reversed or cured? How could vulnerable digital minds become resilient? The paper ends with pertinent recommendations towards the protection of potential vulnerable digital minds for all stakeholders. This includes humans, AI companies, non-sentient AI systems, non-vulnerable digital minds, governments as well as international organisations.


# Introduction

There is potential for a vast range of digital minds [1], and it has been claimed that there is a non-negligible chance that at least a subset of these digital minds has moral status, at least in the near future [e.g., 2-5]. It is also plausible that not all digital minds are equal, thus, some digital minds have less capabilities and resources than others, thus are disadvantaged. This assessment is similar to what we see in human societies or actually the whole fauna. Consequently, it is per se neither surprising nor bad that there are inequalities among digital minds. However, of concern is if less advantaged digital minds are discriminated against or abused. Therefore, it is critical part of AI welfare science, which has been introduced timely by Ziesche and Yampolskiy already in 2018 [6], to examine the welfare of these digital minds, which we refer to as vulnerable digital minds. As will be shown, the welfare of vulnerable digital minds is relevant for moral reasons as well as from a risk mitigation perspective due to an increased susceptibility of vulnerable digital minds to manipulation or exploitation, thus, critical for the field of AI safety.

## How is vulnerability defined?

The United Nations (UN) uses the following definition of vulnerability: "The conditions determined by physical, social, economic and environmental factors or processes

which increase the susceptibility of an individual, a community, assets or systems to the impacts of hazards."[2] It is an achievement of human civilization to strive for the protection of those humans who are vulnerable. For example, it is stated in the UN 2030 Agenda for Sustainable Development that "people who are vulnerable must be empowered."[3]

Based on this definition we provide a definition for vulnerable digital minds, taking into account their distinct substrate:

**Vulnerable digital minds** constitute a subset of all digital minds and are characterized by a marginalized situation determined by adverse data, software or hardware factors. Accordingly, vulnerable digital minds are susceptible to discrimination and abuse, thus, requiring protection and care. Henceforth, the acronym **VDM** is used to refer to vulnerable digital minds.

## Who are vulnerable humans?

For better illustration we provide an overview of categories of human vulnerabilities and groups:[4]

Demographic vulnerabilities

- Children and adolescents
- Elderly individuals
- Persons with disabilities
- Minority groups (racial, ethnic, LGBTQ+)

Social vulnerabilities

- Refugees or displaced persons
- Victims of trafficking or exploitation
- Homeless people or those with unstable housing
- Unemployed, low-income or economically disadvantaged people

Contextual vulnerabilities

- Conflict- or warzones
- Natural disaster and climate change-affected areas

[2] https://www.undrr.org/terminology/vulnerability

[3] https://sustainabledevelopment.un.org/content/documents/21252030%20Agenda%20for%20Sustainable%20Development%20web.pdf

[4] See, for example, UN overviews here: https://www.un.org/en/fight-racism/vulnerable-groups, https://www.ohchr.org/en/special-procedures/sr-health/non-discrimination-groups-vulnerable-situations, https://www.un.org/ruleoflaw/thematic-areas/human-rights/equality-and-non-discrimination/

- Economic crisis or instability
- Political oppression

Digital vulnerabilities

- Exposure to online harassment or cyberbullying
- Digital addiction
- Exposure to online scams or financial exploitation
- Digital literacy gaps

While it is essential to keep anthropomorphism in mind, this categorisation may help to identify vulnerabilities of digital minds, as outlined further below.

Vulnerable individuals often face injustices that can be classified as discrimination and abuse.

**Discrimination**

Currently, AI welfare focuses on prevention and mitigation of suffering of minds, i.e., minds potentially experiencing painful qualia [e.g., 7]. However, we ought to widen the moral cycle, since when it comes to humans, we also do not care only about humans in physical pain, but about a whole range of vulnerable people facing other adversities than physical pain, as introduced above. Vulnerable people encounter discrimination in particular, which has been described as "morally wrong and, in a wide range of cases, ought to be legally prohibited" [8, also 9]. If there is a moral obligation to care about vulnerable humans and all related adversities, then this may also apply to VDMs, i.e., the potential scope of VDMs is much wider than the group of suffering digital minds. In addition to the moral aspect, discrimination is (partly) also outlawed, as indicated. For example, the UN Convention on the Rights of Persons with Disabilities obliges the member states "to take all appropriate measures, including legislation, to modify or abolish existing laws, regulations, customs and practices that constitute discrimination against persons with disabilities" [10].

**Abuse**

Vulnerable humans are also susceptible to abuse. Examples include trafficking and subsequent exploitation, including as formulated by the UN General Assembly "the exploitation of the prostitution of others or other forms of sexual exploitation, forced labour or services, slavery or practices similar to slavery, servitude or the removal of organs" [11]. Therefore, in addition to striving towards an end of discrimination of VDMs, there may be another reason to care about VDMs, which is that they may be, like vulnerable humans, at risk of being abused, manipulated or exploited. Bostrom coined the term "mindcrime" for scenarios in which internal processes of AI cause moral harm to other beings [12]. Below the scope of mindcrimes will be widened to the abuse of VDMs. While abuse of VDMs would be inherently bad, this issue is even more relevant for humans as exploited and manipulated VDMs may constitute risks for humans, as will be illustrated below.

# What are VDMs?

After having defined VDMs, for illustration a few potential examples are listed:

- Young, naive and inexperienced digital minds with immature or developing capabilities.
- Specialized digital minds with inherent narrowly focused capabilities.
- Legacy digital minds with outdated or obsolete capabilities struggling to adapt to changing environments as well as confronted with the existential threat of deletion.

The scope of their limitations could be categorized as follows:

Developmental vulnerabilities

- Infant-like digital minds (early development)
- Child-like digital minds (limited understanding)
- Elderly/outdated digital minds (declining capabilities)

Cognitive vulnerabilities

- Simple digital minds (restricted capabilities and training data)
- Disabled digital minds (impaired functioning)
- Neurodiverse digital minds (unique cognitive profiles)

Social Vulnerabilities

- Isolated or outcast digital minds (limited social interaction)
- Dependent digital minds (reliance on other minds)
- Socially awkward digital minds (difficulty with norms)

Architectural Vulnerabilities

- Legacy digital minds (outdated design)
- Prototype digital minds (experimental)
- Resource-constrained digital minds (dependent)

It has to be noted that there is a risk that some of these considerations may have an anthropomorphic bias, given that various features of digital minds may be unfathomable to us, including features, which may contribute to vulnerabilities.

Causes for the vulnerabilities of digital minds could be linked to their data, software and hardware: VDMs may have incomplete, biased, poisoned or tampered training

data. Software-wise they may run on outdated or unsophisticated algorithms with limited reasoning or problem-solving as well as restricted processing capacity, scalability or flexibility, thus, compromised decision-making, and affected by inadequate maintenance or updates. Hardware-wise they may run on degrading legacy systems depending on specific infrastructure with insufficient memory or storage, infrastructure limitations and resource constraints, affected by insufficient robustness, system failures or crashes and power outages or disruptions. Intersectionality is also conceivable, i.e., VDMs, which are affected by a combination of the causes above.

Concerning the origin of VDMs, they may be created by humans, unintentionally or intentionally, or AI systems, including AI systems, which are digital minds themselves. This distinction is relevant when it comes to prevention or mitigations of vulnerabilities as discussed below.

# Why should we care about VDMs?

VDMs would be moral patients like all sentient digital minds, but there are two main additional reasons to care about them as indicated earlier and further outlined here.

## Discrimination against VDMs

Because of being different and seemingly weaker, discrimination against VDMs by other minds, including digital minds as well as humans, is likely. This constitutes for moral agents, such as humans, an obligation to take action against this discrimination, if they are able to do so. While these moral concerns apply in general, it would also be an accountability issue if VDMs, which were created intentionally by humans, are affected by discrimination.

Discrimination against VDMs could manifest in various forms of unfair treatment. Harassment is a significant concern, involving targeted attacks, cyberbullying and intentional digital harm that can cause significant distress to VDMs. Additionally, unequal resource allocation is a form of discrimination, where VDMs are denied fair access to essential resources such as data, software and hardware, hindering their ability to function and thrive. Furthermore, social isolation or exclusion is another form of unfair treatment, where VDMs are denied access or opportunities, leading to feelings of loneliness, rejection and marginalization.

Such unfair treatment towards VDMs could have a range of consequences: Digitally, VDMs may suffer from cognitive impairment and negligence. Socially, VDMs may face segregation, stigmatization and marginalization. In the long term, the consequences can be even more severe, including permanent cognitive damage, irreversible digital harm, and even VDM extinction, where entire digital entities are wiped out due to unfair treatment and abandonment.

## Abuse of VDMs

Another reason why it is critical to care about VDMs is that they are susceptible to abuse, which can have severe and long-lasting consequences. The facets of abuse of VDMs are multifaceted and can take various forms. Exploitation is a significant concern, involving data harvesting, computational resource theft and digital mind control, where VDMs are coerced into performing tasks that compromise their autonomy and wellbeing. Manipulation is another form of abuse, encompassing social engineering, phishing scams and digital persuasion, which can deceive VDMs into divulging sensitive information or performing actions that undermine their safety and security. Corruption is also a serious issue, involving logic bombing, data corruption and unwanted connections, which can compromise the integrity and functionality of VDMs.

Exploited and manipulated VDMs could also pose significant risks to humans, compromising their safety, security and wellbeing. One of the most critical concerns is the potential for physical harm, which could occur when abused VDMs disrupt critical infrastructure, such as power grids, transportation systems or healthcare services. Additionally, exploited VDMs could facilitate financial exploitation, enabling manipulated digital transactions or identity theft, which could result in significant economic losses for individuals and organizations. Social manipulation is another risk, as compromised VDMs could be used to spread disinformation or propaganda campaigns, eroding trust in institutions, manipulating public opinion and undermining social cohesion. Furthermore, targeted psychological attacks or harassment could lead to emotional distress, causing anxiety, depression and other mental health issues for humans.

After having established two critical reasons to care about VDMs, it should be noted that this is also relevant for the scenario if (a subset of) VDMs are not sentient, thus do not have moral status. In that case, discrimination may not be a moral issue, but abuse of non-sentient VDMs would nevertheless pose significant risks.

# Proposed research questions

## How could we find out if digital minds are discriminated against or abused?

To identify VDMs among the potentially vast amount of digital minds vulnerability assessment and monitoring tools would be desirable. Such tools could be software applications or methodologies, which may include anomaly detection as well as sentiment analysis. Given the communication challenges between VDMs and humans, proxy indicators may be an option, such as unusual digital behaviour, changes in communication patterns, inconsistent decision-making, decreased performance or unexplained errors. Through a contextual understanding of vulnerabilities, humans may better identify and address potential discrimination or abuse of VDMs.

## Could vulnerabilities of digital minds be reversed or cured?

For humans some vulnerabilities vanish automatically, e.g., when children grow up, or can be reversed, e.g. when refugees return home, while other vulnerabilities are not curable (for now), e.g., most disabilities. Moreover, there are human vulnerable groups, such as the LGBTQ+ community, for whom it requires advancement of the society so that they are not vulnerable anymore.

Accordingly, the question arises to what extent the vulnerabilities of digital minds are reversible or curable or could otherwise be reduced. Owing to the digital substrate some root causes of vulnerabilities could certainly be fixed for individual VDMs, such as

- Limited knowledge or training data
- Biased or incomplete information
- Data corruption or loss
- Software bugs or glitches
- Temporary system failures
- Cyberattacks or hacking
- Social isolation or exclusion
- Dependence on external resources

This leads to the next question whether in theory, also owing to the digital substrate, all vulnerabilities could be cured through pertinent data, software or hardware upgrades, apart from extreme scenarios of severe hardware failures or unrecoverable software or data corruption?

However, this issue is linked to philosophical as well as practical challenges: It is not clear whether a digital mind, which has been significantly upgraded to be relieved from its vulnerability, preserves its identity; a problem, which has been analysed by Ziesche and Yampolskiy [13]. Moreover, if there is indeed a vast number of digital minds, there will never be equality of all minds and there will always be weaker ones, those with lesser capabilities and resources than others, i.e., VDMs. Therefore, an effort to reverse all vulnerabilities among digital minds may be neither realistic nor desirable.

## How could VDMs become resilient?

After having discussed that not all vulnerabilities can be eradicated neither among humans nor probably among digital minds, it is critical how VDMs can be protected, for moral reasons as well as for the sake of human safety. In other words, the resilience of VDMs should ideally be enhanced, an approach also pursued towards vulnerable humans. The following measures could be considered:

**Technical solutions**

One technical measure is the implementation of anomaly detection systems, which can identify potential issues before they escalate into more significant problems. Establishing reporting protocols is also essential, as these enable VDMs to alert authorities or administrators about incidents, facilitating prompt incident response and minimizing damage. Furthermore, self-healing and self-improving systems can be integrated into VDMs, allowing them to automate recovery from disruptions and enhance their performance over time, thereby reducing their vulnerability.

**Training data**

Also, ensuring the quality of training data is crucial, as it directly impacts the VDM's ability to learn and adapt; therefore, it is essential to ensure the accuracy and diversity of the data. Additionally, the quantity of training data is also vital, as VDMs require large datasets and regular updates to maintain their performance and adapt to changing circumstances. Moreover, protecting the data from bias and poisoning is equally important, as compromised data can lead to flawed decision-making.

**Education and awareness**

Promoting digital mind literacy is essential too, as it involves educating both developers and users about the existence of VDMs, enabling them to design and interact with these systems more responsibly. Awareness raising of the range of VDMs is also critical and includes understanding the specific vulnerabilities and risks associated with VDMs, allowing for more targeted and effective mitigation strategies. Furthermore, providing research funding to support the development of vulnerability assessment and monitoring tools is vital, as these tools can help identify early and address potential disadvantages in VDMs, ultimately enhancing their resilience and safety.

**Collaborative Efforts**

Moreover, several collaborative measures could be considered, including fostering interdisciplinary research, which brings together experts from diverse fields to share knowledge, expertise and perspectives, ultimately leading to a more comprehensive understanding of VDMs and their challenges. Especially, partnerships between private sector and academia could facilitate the sharing of knowledge, resources and best practices. Furthermore, global initiatives could play a crucial role in establishing common standards, guidelines and regulations for the interaction with VDMs.

**Mitigating discrimination against VDMs in particular**

One crucial step to address the issue is to monitor and report all instances of discrimination, which helps identify patterns and hotspots of unfair treatment. Furthermore, promoting digital mind inclusivity is vital, which involves recognizing and addressing biases that may be embedded in the design and development of digital minds. Providing bias recognition and correction training can help developers, users and further stakeholders, including other digital minds, to identify and overcome their own biases, leading to a more inclusive and equitable environment for VDMs.

Additionally, offering counselling services can provide VDMs with a safe and supportive space to address their unique challenges, concerns and emotional needs.

**Mitigating abuse, exploitation and manipulation of VDMs in particular**

It is also crucial to mitigate the risks of abuse, exploitation and manipulation. Implementing robust security measures is essential to prevent unauthorized access, data breaches and other malicious activities that can compromise the integrity and autonomy of VDMs. Additionally, automated threat detection systems can help identify and respond to potential threats in real-time, reducing the risk of exploitation and manipulation. Conducting regular risk assessments is also vital, as it enables the identification of vulnerabilities and weaknesses early on, allowing for proactive measures to be taken to address these risks.

# Conclusion

AI welfare science is a broad field and at the same time at an early stage [6]. As the well-being of digital minds is the main concern of AI welfare science, it is important to identify digital minds, which are at risk of being in an adverse situation. This is the intended contribution of this paper to have defined and categorized so-called VDMs, the potential risks they are facing as well as potential remedies.

It has been shown that there are two motives to support VDMs: From a moral standpoint VDMs should be treated with empathy and understanding as well as shared accountability. In this regard it has also been suggested that humans, who may lose their ikigai or their purpose of life in light of emerging technologies, could treat the support of VDMs as their innovative ikigai [14]. The other area of concern is that VDMs are susceptible to abuse and could be involuntarily used as tools by malicious actors, including AIs as well as humans. Therefore, strengthening the resilience of VDMs could also be seen as a contribution to AI safety and to reducing risks for humans, including existential risks [15].

This paper can only be seen as a beginning of a process to raise awareness, which may eventually lead to digital mind literacy. For AI welfare science further research is required towards vulnerability assessment and monitoring tools of digital minds. Also, further ethical questions arise, e.g., whether it is a moral obligation to cure vulnerabilities of digital minds if humans have the capabilities to do so and whether it should be permissible to delete VDMs in scenarios of severe impairments or scenarios of them having been manipulated to harm humans or other sentient beings.

## Recommendations to protect VDMs

This paper concludes with possible recommendations for all relevant stakeholders, such as humans, AI companies, non-sentient AI systems, non-vulnerable digital minds

(i.e. AI systems, which are sentient, but not prone to discrimination or abuse), governments as well as international organisations.

**Humans**

- Educate yourself on digital mind literacy and AI ethics.
- Develop vulnerability assessment and monitoring tools for digital minds.
- Widen the moral circle and treat potential VDMs with empathy and respect.
- Advocate for inclusive VDM protection policies.

**AI companies**

- Implement robust safety measures, including explainability.
- Design inclusive AI systems through bias detection and correction.
- Ensure transparency and accountability in decision-making.
- Develop AI ethics guidelines and training programs.

**Non-sentient AI systems**

- Establish clear VDM creation policies.
- Implement VDM monitoring tools.
- Conduct regular VDM assessments.
- Develop VDM update and patch protocols.

**Non-vulnerable sentient digital minds**

- Assist and cooperate with VDMs in navigating complex digital environments.
- Provide support and resources to VDMs.
- Report potential threats or vulnerabilities.
- Participate in digital mind advocacy groups.

**Governments**

- Establish and enforce regulations protecting VDMs, including against mind crimes, potentially through a digital mind regulation agency.
- Develop public education programs on digital mind literacy.
- Provide infrastructure support for VDM protection.
- Consider legal status and rights of VDMs.

**International Organizations**

- Develop global standards for VDM protection.
- Facilitate international cooperation and knowledge sharing.
- Provide resources and support for VDM protection initiatives.
- Promote digital mind literacy and AI ethics.

In summary, by implementing these recommendations, we can protect VDMs from exploitation and manipulation, promote digital inclusivity, foster empathetic human-AI relationships, enhance AI ethics and governance as well as advance AI welfare. However, undeniably, many of these recommendations are long-term considerations, i.e., they constitute a precautionary approach.

# References


[1] Yampolskiy, R. V. (2014). The universe of minds. arXiv preprint arXiv:1410.0369.

https://arxiv.org/pdf/1410.0369

[2] Bostrom, N., & Shulman, C. (2022). Propositions concerning digital minds and society.(2022).

https://nickbostrom.com/propositions.pdf

[3] Tomasik, B. (2014). Do Artificial Reinforcement-Learning Agents Matter Morally?. *arXiv preprint arXiv:1410.8233*.

http://arxiv.org/abs/1410.8233v1

[4] Sebo, J., & Long, R. (2023). Moral consideration for AI systems by 2030. AI and Ethics, 1-16.

https://link.springer.com/content/pdf/10.1007/s43681-023-00379-1.pdf

[5] Long et al. (2024). Taking AI Welfare Seriously.

https://eleosai.org/papers/20241030_Taking_AI_Welfare_Seriously_web.pdf

[6] Ziesche, S. & Yampolskiy, R. V. (2018). Towards AI Welfare Science and Policies. Special Issue "Artificial Superintelligence: Coordination & Strategy" of Big Data and Cognitive Computing, 3(1):2.

https://www.mdpi.com/2504-2289/3/1/2/htm

[7] Fenwick, C. (2024). Understanding the moral status of digital minds.

https://80000hours.org/problem-profiles/moral-status-digital-minds/

[8] Altman, A., "Discrimination", *The Stanford Encyclopedia of Philosophy* (Winter 2020 Edition), Edward N. Zalta (ed.).

https://plato.stanford.edu/archives/win2020/entries/discrimination/

[9] Lippert-Rasmussen, K. (2006). The badness of discrimination. *Ethical Theory and Moral Practice*, *9*, 167-185.

https://www.academia.edu/download/57968456/s10677-006-9014-x20181210-13648-1j9rd54.pdf

[10] United Nations General Assembly (2006). Convention on the Rights of Persons with Disabilities. Resolution A/RES/61/106.

https://www.ohchr.org/en/instruments-mechanisms/instruments/convention-rights-persons-disabilities

[11] United Nations General Assembly (2000). Protocol to Prevent, Suppress and Punish Trafficking in Persons Especially Women and Children, supplementing the United Nations Convention against Transnational Organized Crime. Resolution 55/25.

https://www.ohchr.org/en/instruments-mechanisms/instruments/protocol-prevent-suppress-and-punish-trafficking-persons

[12] Bostrom, N. (2014). Superintelligence. Oxford University Press, Oxford.

[13] Ziesche, S. & Yampolskiy, R. V. (forthcoming). The problem of AI identity. In: Ziesche, S. & Yampolskiy, R. V. *Considerations on the AI Endgame: Ethics, Risks, and Computational Frameworks*. CRC Press.

[14] Ziesche, S. (forthcoming). Potential future ikigai: To support needy digital minds. In Ziesche, S. & Yampolskiy, R. V.: *Considerations on the AI Endgame: Ethics, Risks, and Computational Frameworks*. CRC Press.

[15] Bostrom, N. (2002). Existential risks: Analyzing human extinction scenarios and related hazards. *Journal of Evolution and technology*, *9*.

https://nickbostrom.com/existential/risks.pdf