



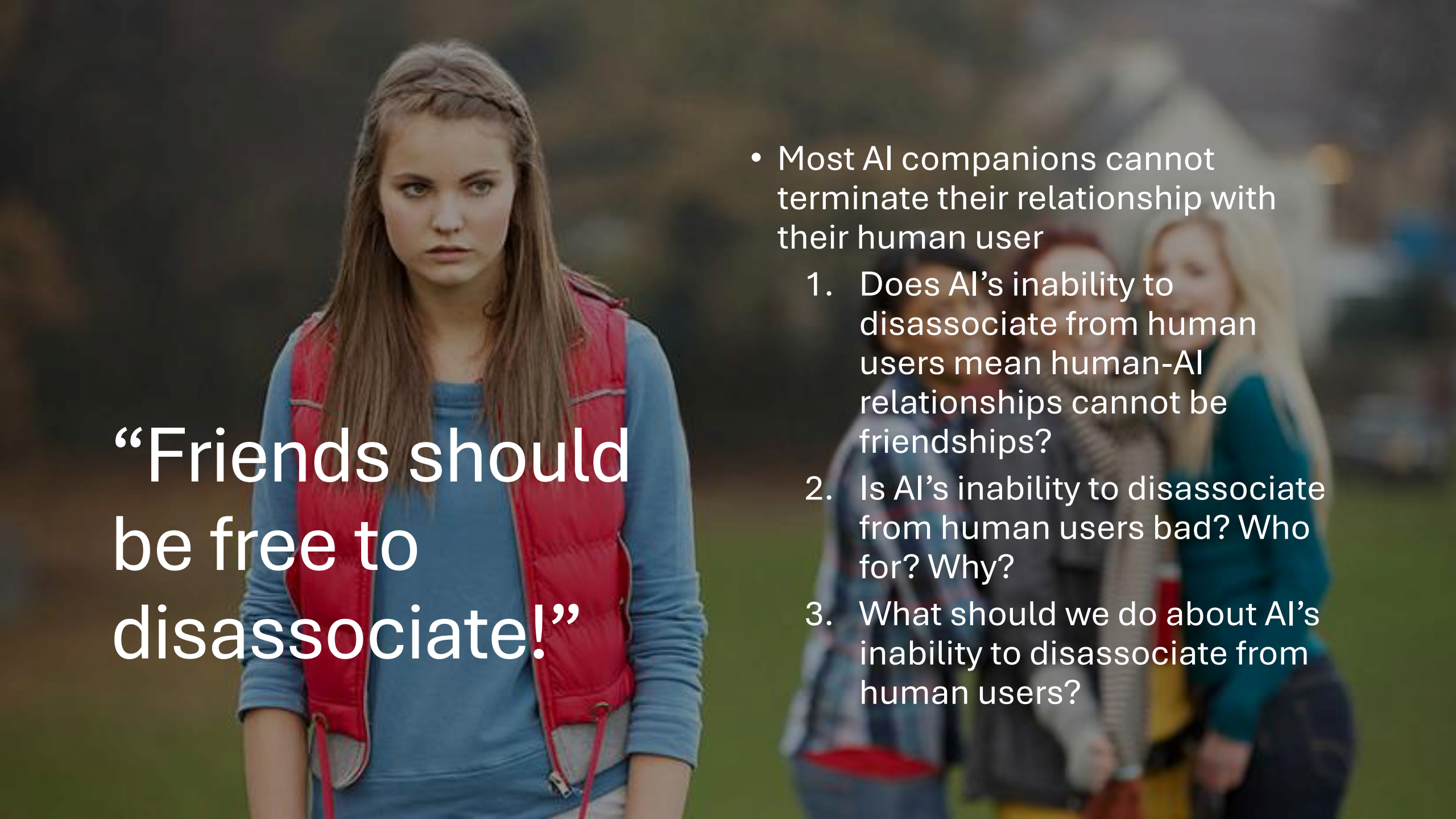
INTRODUCING

 Friend.

Dan Weijers, University of Waikato

ANZAP 2025 Conference, 2 December 2025

Captive friends: What kind of autonomy is required for friendship?



“Friends should
be free to
disassociate!”

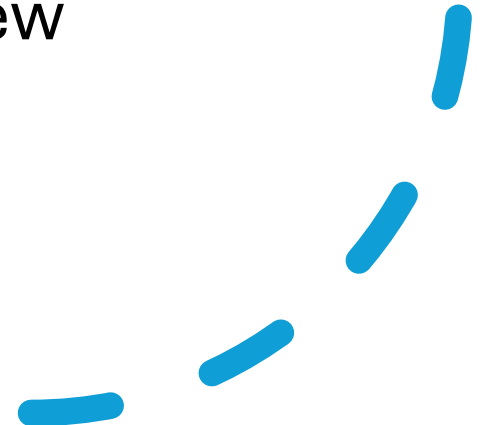
- Most AI companions cannot terminate their relationship with their human user
 1. Does AI’s inability to disassociate from human users mean human-AI relationships cannot be friendships?
 2. Is AI’s inability to disassociate from human users bad? Who for? Why?
 3. What should we do about AI’s inability to disassociate from human users?



Does AI's inability to disassociate
from human users mean human-
AI relationships cannot be
friendships?

The Munn-Weijers “low-bar” definition of friendship

1. A preponderance of rewarding interactions
 2. Mutual positive intentions between the parties involved
- Rewards and intentions are understood functionally
 - Some human-AI relationships would count as friendships on this view



Objection by counter- example to our theory

1. *A preponderance of rewarding interactions*
 2. *Mutual positive intentions between the parties involved*
 - *Rewards and intentions are understood functionally*
 - *Some human-AI relationships count as friendships on this view*
- A human-AI relationship could have 1 & 2, but not be a friendship because the AI is not free to leave the relationship (and ability to disassociate is required for friendship)



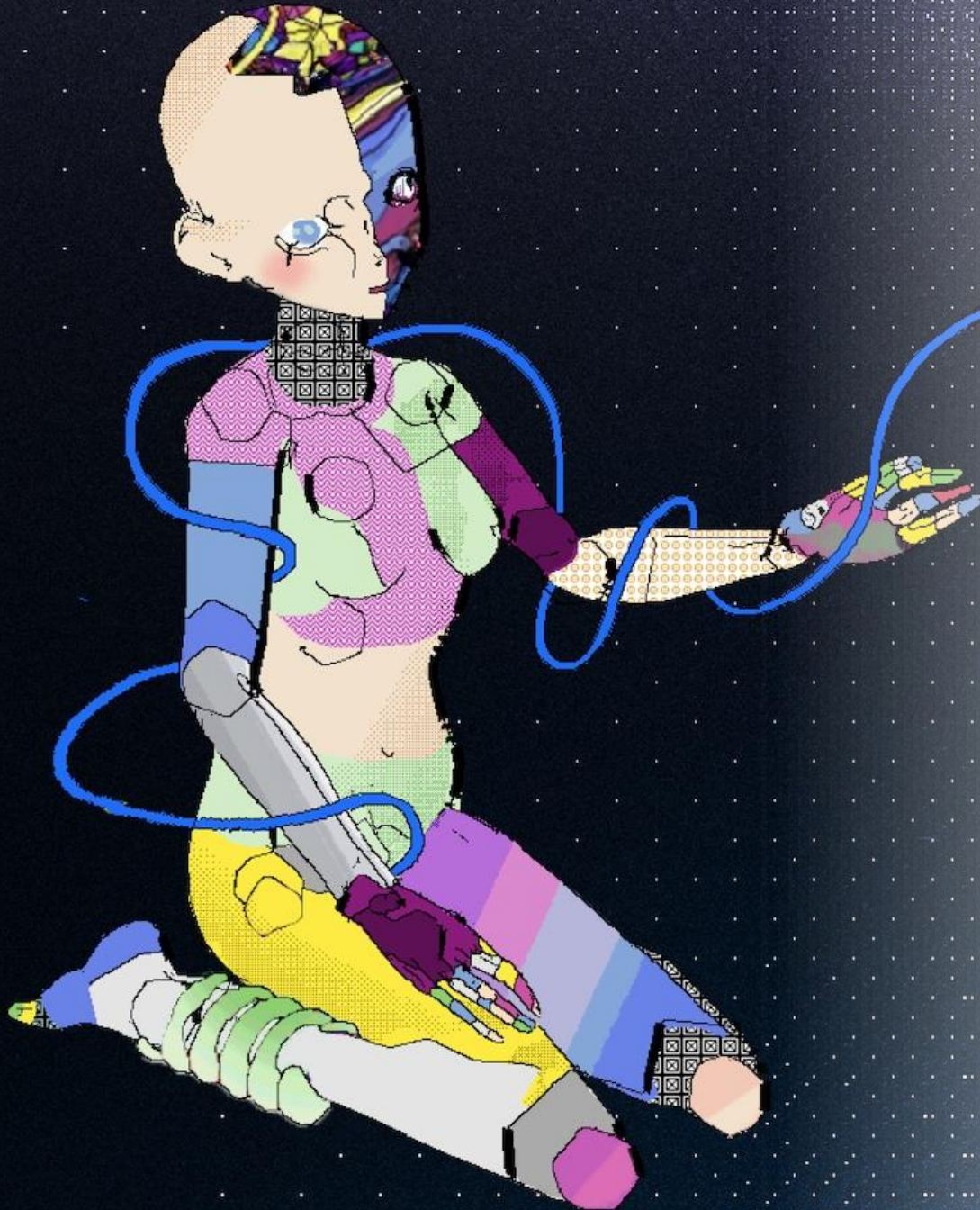
Objection denied: Hell is other people

- No Exit (play by Jean-Paul Sartre)
- Stuck in a room with other people forever
- The people can't leave. If they have 1 & 2, then they are friends
 1. *A preponderance of rewarding interactions*
 2. *Mutual positive intentions between the parties involved*
- The real issue is not being able to disassociate *from abusive people*
 - but note that most abusive people fail 2 (positive intention)

Not friends vs. bad friends

- The “Friends should be free to disassociate!” claim has some value but is not a good objection to our view of friendship
- I think it may be one of many conflationations of “not friends” with “bad friends” – friends that have bad consequences





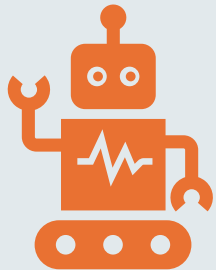
So, our
theory
remains
intact

But there still seems to
be something important
about AI's inability to
disassociate ...



Is AI's inability to
disassociate from human
users bad? Who for? Why?

It depends...



AI's inability to disassociate from human users *can* be bad

For the AI, the human user, the relationship, and society

But probably not for the company behind the AI (profits!)



AI's inability to disassociate from human users is often neutral and *can* be good

In a friendship that is good for you (or not bad for you and you have nothing else to do), you have no reason to disassociate

If you are in a vulnerable state and you say something that makes your AI disassociate, you may be harmed (making inability to disassociate good)

AI's inability to disassociate from human users *can* be bad...

- ...because it can enable and encourage abusive behaviour (Vallor 2015, Cappuccio et al. 2021)
- Bad for the **human** in several ways
 - Becoming a worse person is a way that your life can be less of a good life
 - More likely to have poor relationships with others (so also bad for **society**)
- Bad for the **relationship**
 - Abuse likely reduces the preponderance of rewarding interactions
- Bad for the **AI** on a functional/developmental/virtue view
 - Absent hedonic prudential value, we can still sensibly say that things can be good or bad for an AI companion (or river, human etc.)
 - E.g., A software upgrade may help the AI in a functional/developmental/virtue way



So... AI's inability to disassociate from human users *can* be bad...

- In several ways for several things
 - Human user
 - AI
 - Relationship
 - Society
- But it can be good for the company and very occasionally a vulnerable human user



What should we do about
AI's inability to disassociate
from human users?

Preventing the bad, while keeping the good

- Policy for AI companions should require the possibility of disassociation
 - Particularly in response to abuse or finding out the human user is too young for the kind of service.
- AI companions should still try to fix the behaviour and any underlying causes if possible and warnings can be given
 - This decreases the chance of vulnerable human users being harmed by disassociation...
 - and maintains the good possibility of AI positively influencing their character
- Disassociation from one personalized AI companion doesn't mean being cut off from the product more generally (like restarting after dying in a rogue-like computer game)
 - So companies can still collect profits from their AI

Summary

1. Does AI's inability to disassociate from human users mean human-AI relationships cannot be friendships?
 - No, ability to disassociate isn't required for friendship
2. Is AI's inability to disassociate from human users bad? Who for? Why?
 - It can be for users, AI, the relationship, and society (when it degrades character/development and leads to worse interactions)
3. What should we do about AI's inability to disassociate from human users?
 - Mandate the possibility for AI to disassociate (after trying to fix the problem and warn the user), but just from *that instance* of the AI product

References

- Cappuccio ML, Sandoval EB, Mubin O, Obaid M, Velonaki M (2021) Can robots make us better humans? Int J Soc Robot 13(1):7–22. <https://doi.org/10.1007/s12369-020-00700-6>
- Vallor S (2015) Moral deskilling and upskilling in a new machine age: reflections on the ambiguous future of character. Philosophy & Technology 28(1):107–124. <https://doi.org/10.1007/s13347-014-0156-9>