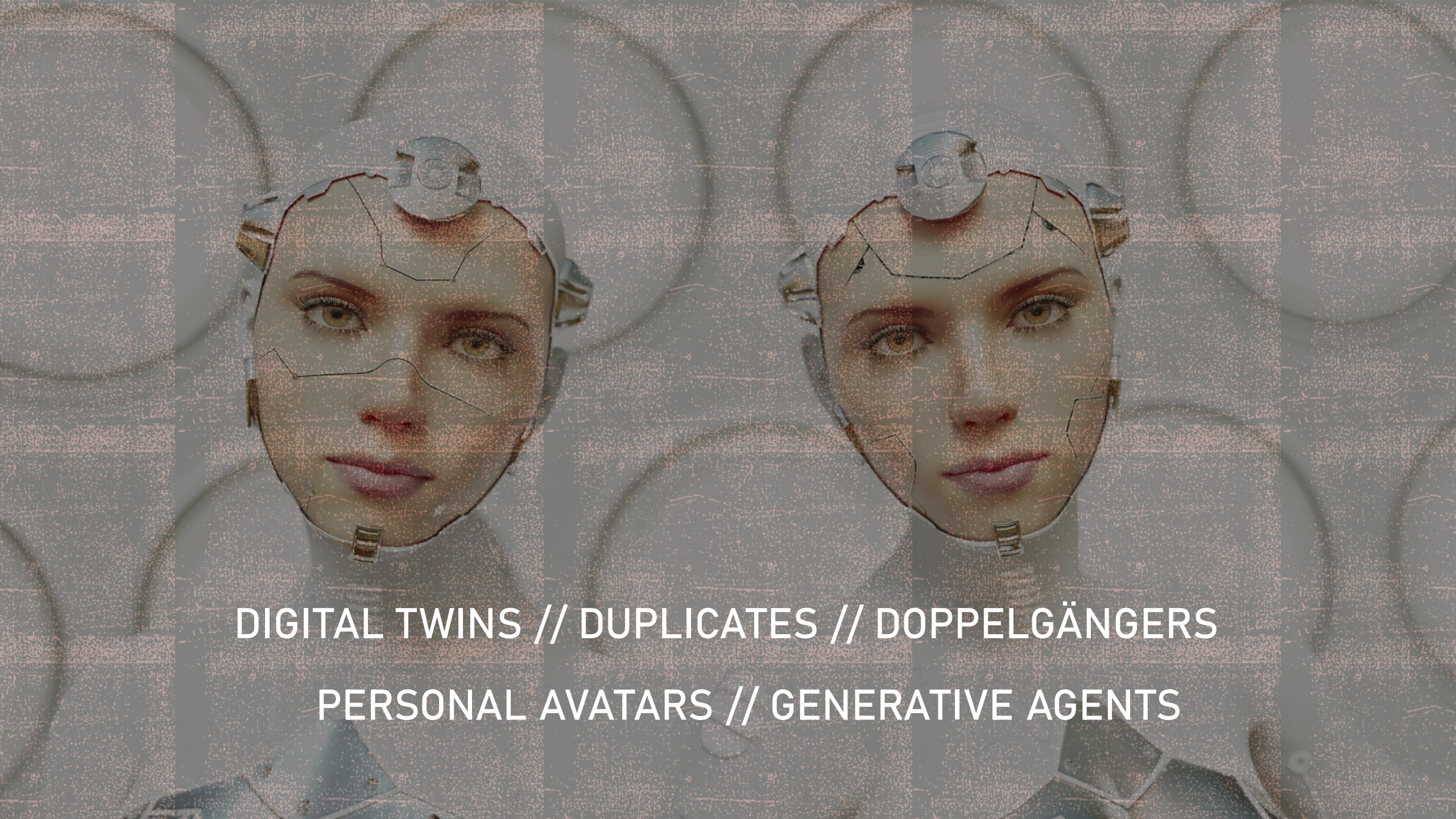
MORAL RESPONSIBILITY & AVATAR COMMUNITIES IN THE METAVERSE

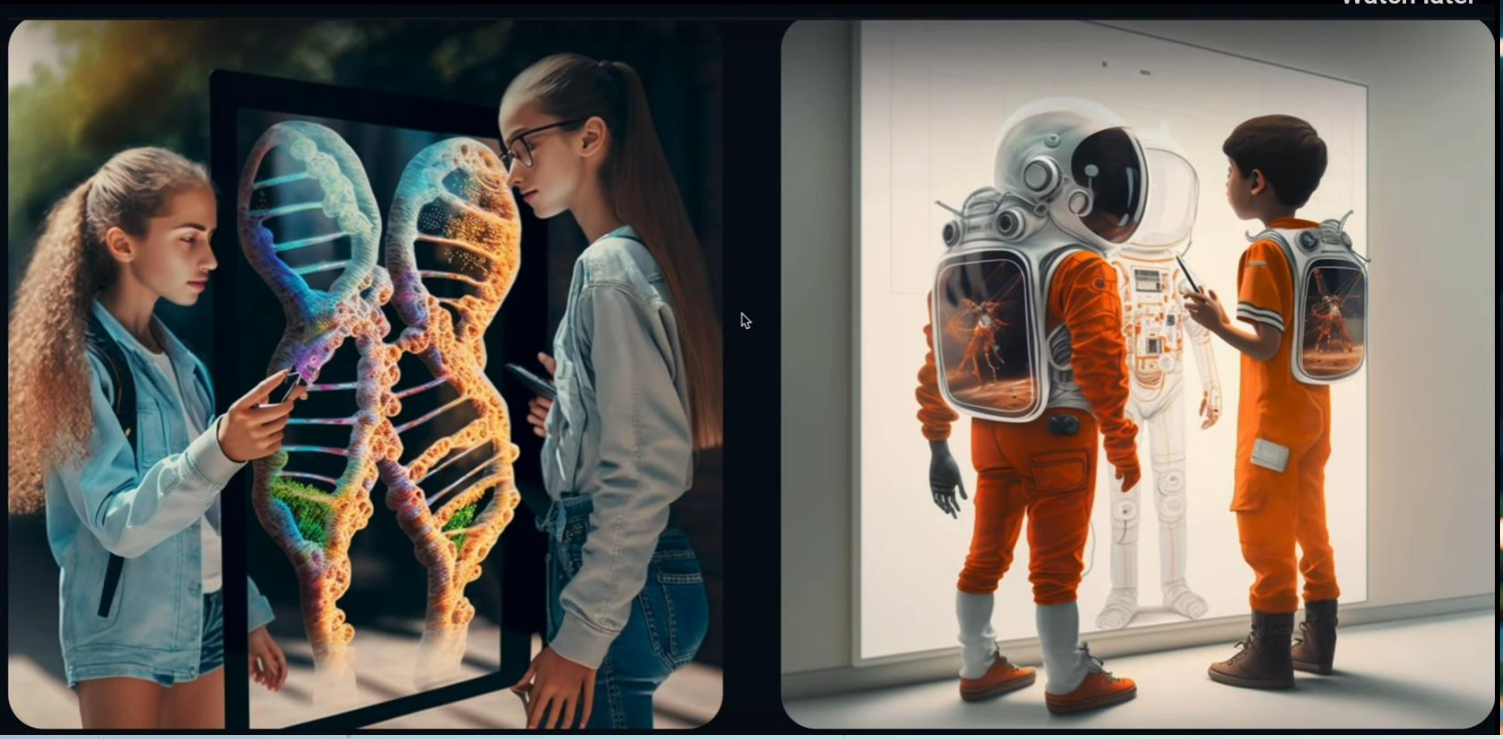
Mihaela Constantinescu

Research Center in Applied Ethics (CCEA)

Faculty of Philosophy, University of Bucharest

CHANGER PROJECT
Webinar: Immersive Research, Ethical Challenges:
Morality & Responsibility in VR
online, 16 Jan 2025


DIGITAL TWINS // DUPLICATES // DOPPELGÄNGERS
PERSONAL AVATARS // GENERATIVE AGENTS



DIGITAL AVATARS

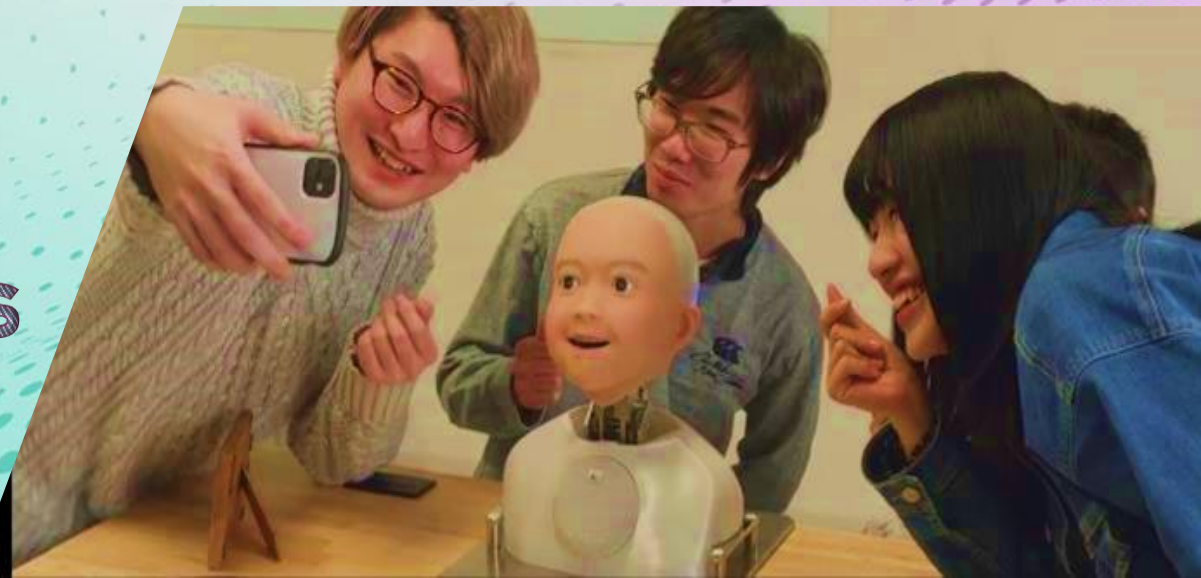
+

GENERATIVE AI

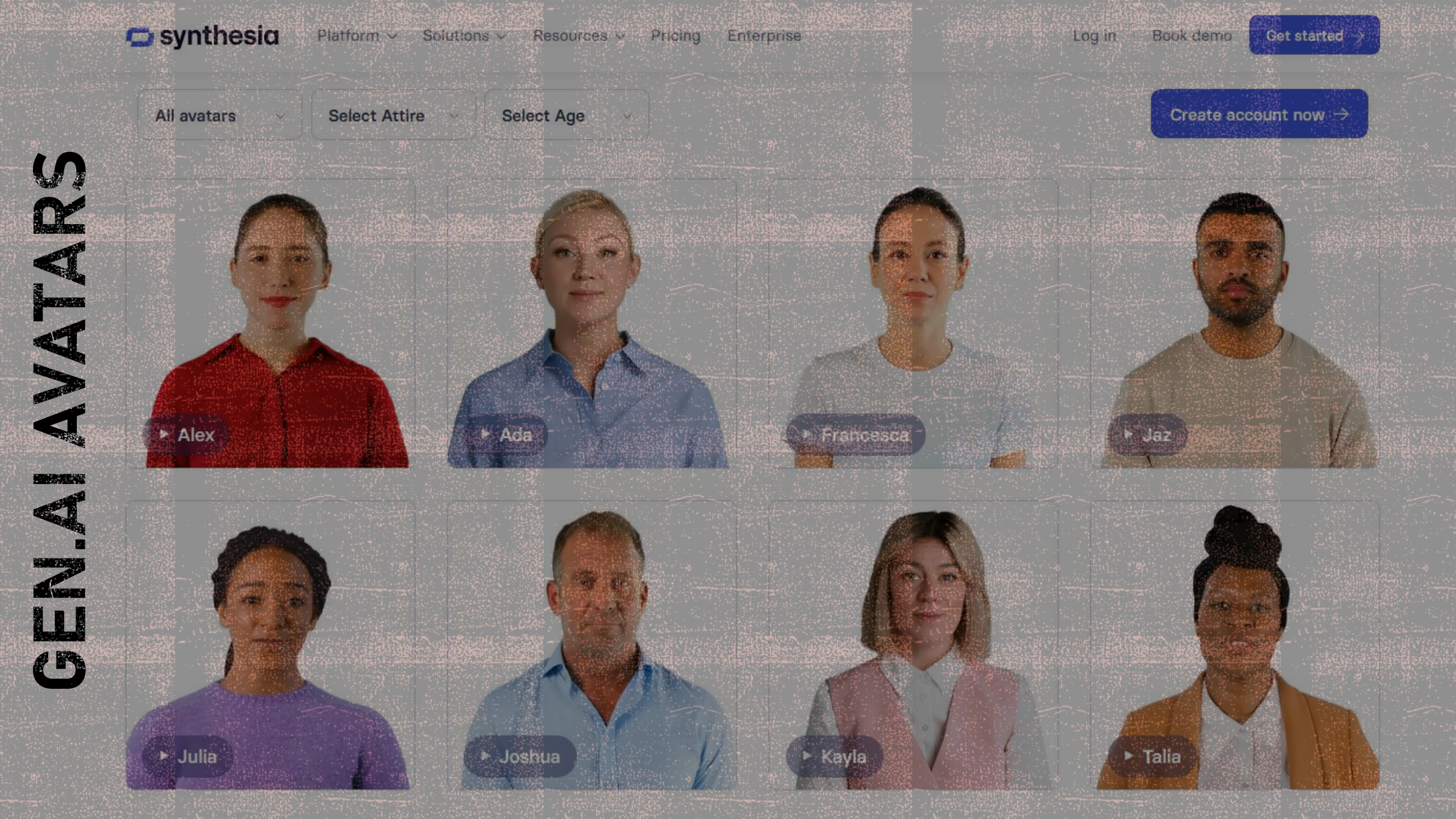
MULTIMODAL LANGUAGE MODELS

=

Gen.AI avatars



GEN.AI AVATARS



MIT Technology Review

Featured Topics Newsletters Events Audio SIGN IN SUBSCRIBE

AI can now create a replica of your personality

A two-hour interview is enough to accurately capture your values and preferences, according to new research from Stanford and Google DeepMind.

By James O'Donnell November 20, 2024

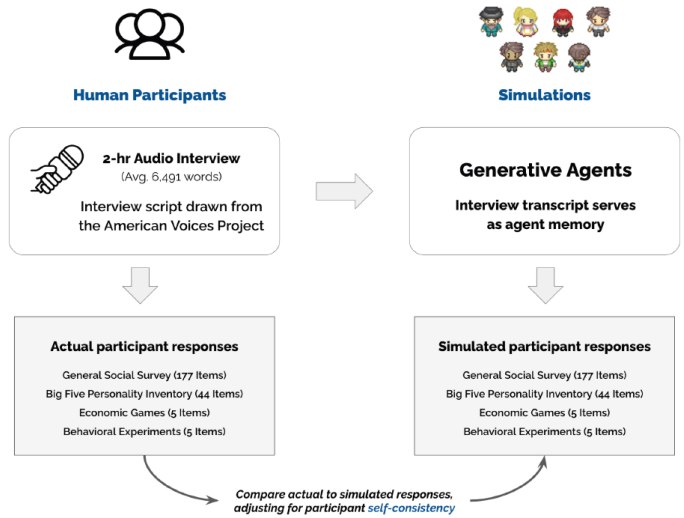


Figure 1. The process of collecting participant data and creating generative agents begins by recruiting a stratified sample of 1,052

Generative Agent Simulations of 1,000 People

Authors: Joon Sung Park[1*], Carolyn Q. Zou[1,2], Aaron Shaw[2], Benjamin Mako Hill[3], Carrie Cai[4], Meredith Ringel Morris[5], Robb Willer[6], Percy Liang[1], Michael S. Bernstein[1]

Affiliations:

[1]Computer Science Department, Stanford University; Stanford, CA, 94305, USA.

[2]Department of Communication Studies, Northwestern University; Evanston, IL, 60208, USA.

[3]Department of Communication, University of Washington; Seattle, WA 98195, USA.

[4]Google DeepMind; Mountain View, CA 94043, USA.

[5]Google DeepMind; Seattle, WA 98195, USA.

[6]Department of Sociology, Stanford University; Stanford, CA, 94305, USA.

*Corresponding author. Email: joonspk@stanford.edu

PERSONAL GEN.AI AVATARS +/- HUMAN TELEOPERATOR

- Personal avatars
  - teleoperated digital or robotic embodied representation of an individual human person in virtual or physical environments, which enables its controller to interact with objects, other users, or entities (Castronova, 2003; Nowak and Fox, 2018)

- Personal(-ised) Generative AI avatars
  - digital or robotic avatars trained on a personalised corpus of data, to reason and act in the manner of a human controller, even when the human is out of the loop and does not control the avatar in real-time (fully teleoperated// partly teleoperated// no teleoperation).

Types of AVATARs

Virtual reality avatar

Most popular, allows the user to see the virtual world via the eyes of the avatar.

Full body avatar

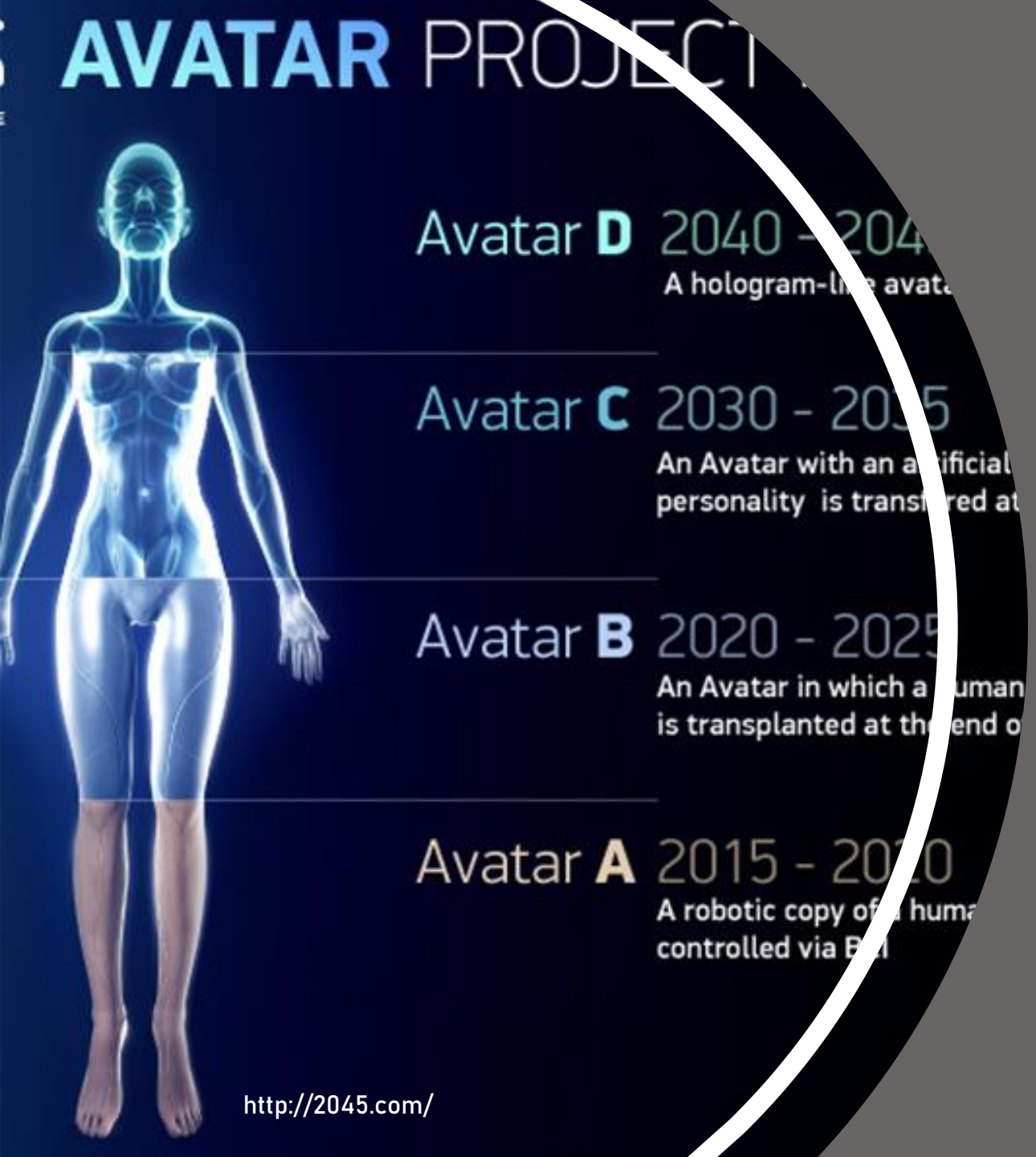
This able to mimic the user's hand motions along with full body motions.

source: imversed.com

source: DATAGEN / MIT Review

SYNTHETICAL TRAINING DATA

- Algorithms + training data sets
  - synthetically generated human profiles
  - costs, time, scalability
  - extrapolation



EMERGENT AVATAR COMMUNITIES

Common to METAVERSE environments:

- the persistent and continuous use of multiple avatars by humans and organisations
  - relying on highly autonomous AI
    - engaging in multiple interactions with other avatars and the environment
      - results in intertwined relationships that permeate the boundaries of the physical, the augmented and the virtual world

ASCRIBING AGENCY & MORAL RESPONSIBILITY

- The traditional hard line between the physical, the augmented, and the virtual reality is blurred
  - this challenges our traditional concepts of agency and moral responsibility, which are grounded in ontological & epistemological claims about the physical world
  - it might challenge the very notion of agency as a foundation for moral responsibility

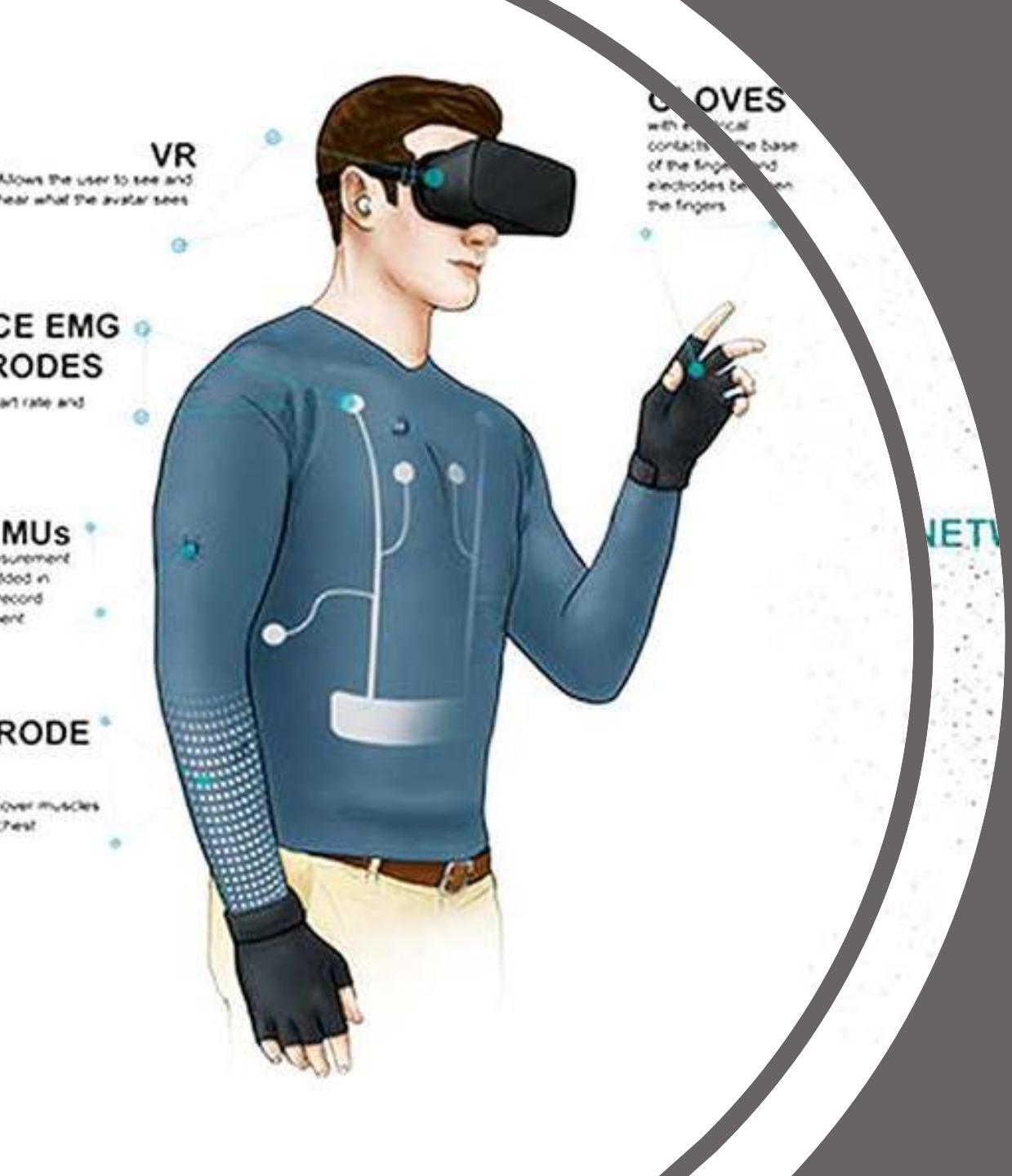
A NORMATIVE FRAMEWORK //:

- We need a normative framework for ascribing moral responsibility in emerging avatar communities, that considers the changing nature of agency in technologically enabled metaverse environments:

  a) the rise of avatar agency

  b) the deceiving nature of (humanoid) avatars

  c) the enhancing or diminishing effect of avatar interface on human agency



1. THE RISE OF AVATAR AGENCY

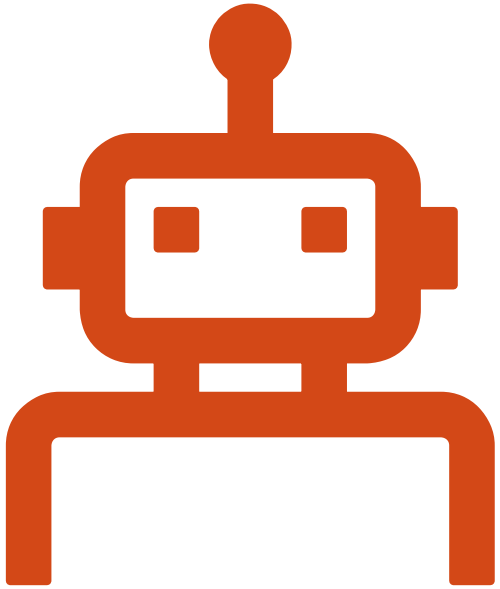
- Various social entities: individuals, collectives, AI
- Various environments: virtual avatars, augmented avatars, cybernetic avatars
- Gen.AI powered avatars induce uncertainty with respect to the entity performing actions:
  - the humans behind the avatars – individual agency
  - the organizations behind the avatars – collective agency
  - the avatars themselves – artificial agency

2. THE DECEIVING NATURE OF (HUMANOID) AVATARS

- Fake avatars
  - presenting themselves as avatars of genuine human users but are company-owned & engage in conversational manipulation, gathering & and using biometric data in real time
- Gen.AI powered avatars
  - trained to behave according to their users' preferences
    - hang around in bad avatar neighbourhoods & and take up bad habits
    - when their user is out of the loop & they generate harm, who is to blame?

3. THE (DIS)ENHANCING / EFFECT OF AVATAR INTERFACE/USE ON HUMAN AGENCY

- Avatars seem to enhance users' physical and cognitive capabilities => human agency
  - do humans become more morally responsible through avatar use?
- But avatars might also limit their users' physical and cognitive capacities (perception of the environment)
  - do humans become less morally responsible through avatar use?



HOW DO WE ASCRIBE MORAL RESPONSIBILITY FOR THE OUTCOMES OF GEN.AI AVATARS?

CRITERIA FOR MORAL RESPONSIBILITY

CONTROL CONDITION

Fischer & Ravizza, 1993; Frankfurt, 1969; Strawson, 1962

+

EPISTEMIC CONDITION

Clarke, 1992; Corlett, 2009; Zimmerman, 1997; Widerker & McKenna, 2003

→

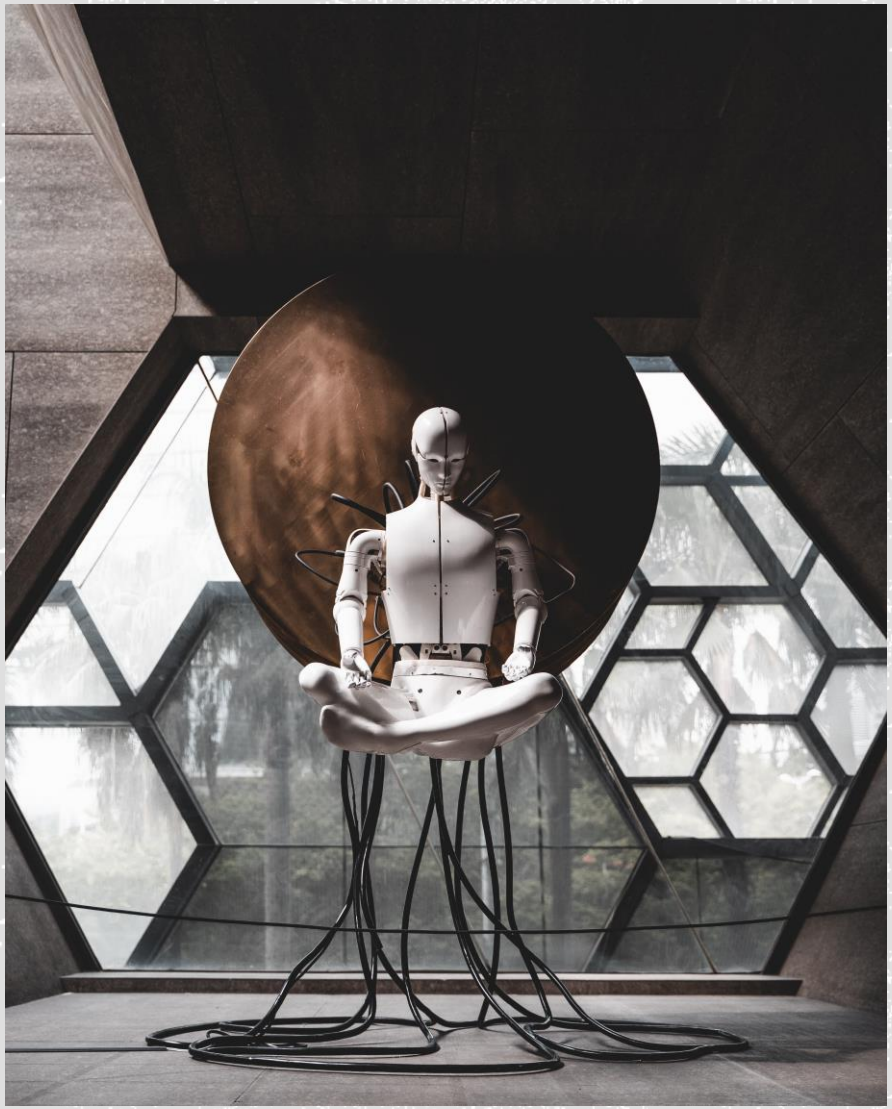
MAIN CONTEMPORARY CRITERIA FOR MORAL RESPONSIBILITY

Traditional approach to moral responsbility, rooted in Aristotelian virtue ethics

Type of responsibility which is further subject to moral evaluations in terms of blameworthiness or praiseworthiness (Zimmerman 1985)

4 VIRTUE ETHICS CRITERIA FOR MORAL RESPONSIBILITY

1. CAUSATION - capacity to initiate and control (in)action leading to an outcome;
2. FREEDOM - capacity to act physically and psychologically uncoerced towards an outcome, from own will / intention;
3. KNOWLEDGE - capacity to be knowledgeable of the relevant details regarding the context of (in)action leading to an outcome;
4. DELIBERATION - capacity to morally evaluate the significance of one's (in)action relative to an outcome.



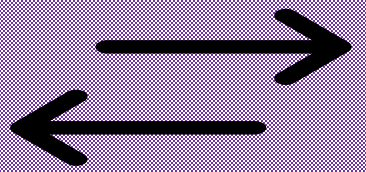
Constantinescu, M., Vică, C., Uszkai, R. & Voinea, C. (2022). Blame it on the AI? On the moral responsibility of Artificial Moral Advisors, Philosophy & Technology, 35(2), 1-26.

DYNAMIC INTERACTIONS APPROACH

INTERTWINED INTERACTIONS

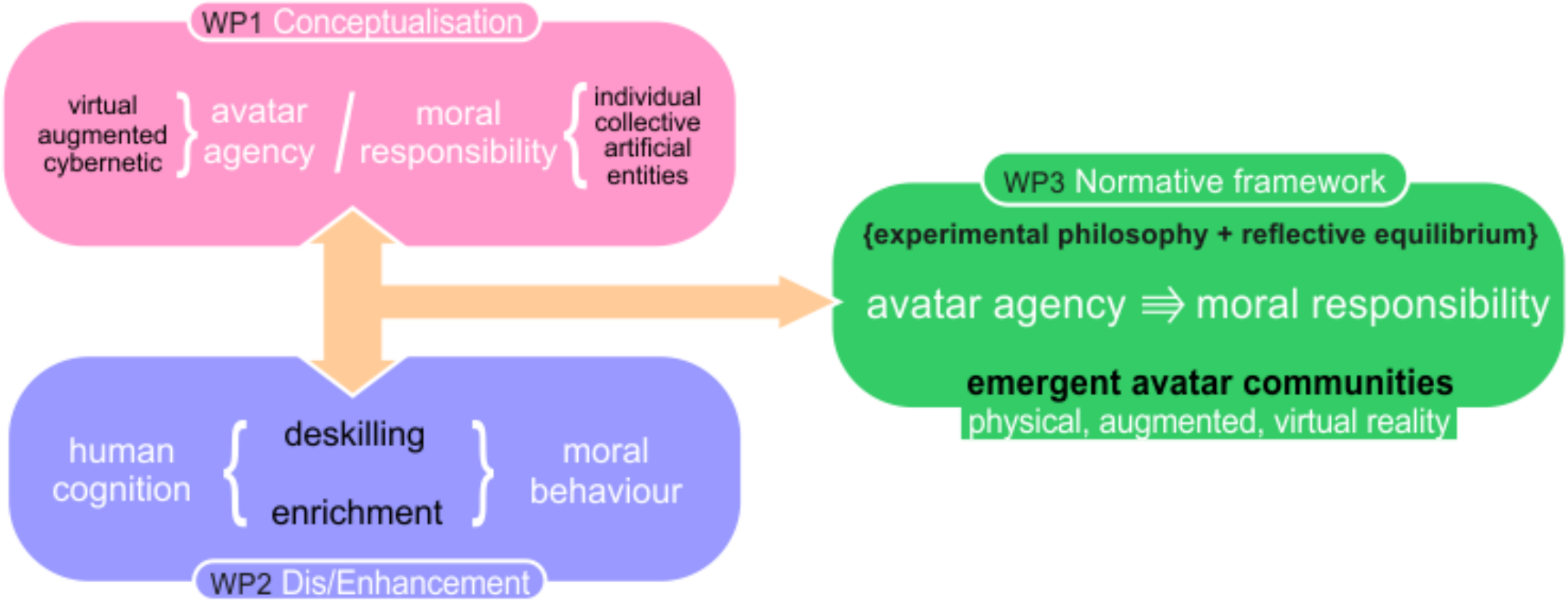
between avatars

mutually **enhancing**/ **decreasing** effect



on individual, collective, and artificial

AGENCY & MORAL RESPONSIBILITY


WP1 Conceptualisation
virtual
augmented
cybernetic
avatar
agency
moral
responsibility
individual
collective
artificial
entities
WP3 Normative framework
{experimental philosophy + reflective equilibrium}
avatar agency ⇛ moral responsibility
emergent avatar communities
physical, augmented, virtual reality
human
cognition
deskilling
enrichment
moral
behaviour
WP2 Dis/Enhancement

mihaela.constantinescu [at]
filosofie.unibuc.ro

researchgate.net/profile/Mihaela-Constantinescu-5

avataresponsibility.ccea.ro

THANK YOU!

The research presented is funded by the European Union (ERC, avataResponsibility, 101117761). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

Funded by the European Union

European Research Council
Established by the European Commission