TECHNISCHE UNIVERSITÄT
ILMENAU

Evaluating QoE for Social XR

Audio, Visual, Audiovisual and Communication Aspects

Alexander Raake & team
Audiovisual Technology (AVT) Group
Institute for Media Technology
Ilmenau Interactive Immersive Technologies Center (I3TC)
Technische Universität Ilmenau (TUIL)

In collaboration with:
Broll et al.; Brandenburg, Werner et al.; Doering et al.; Groß et al., TU Ilmenau
Fels et al. RWTH Aachen; Klatte et al. TU Kaiserslautern;
Habets et al. FAU Erlangen-Nürnberg;
Fröhlich, Kulick et al. Bauhaus Universität Weimar;
and others...

Audiovisual Technology Group

The SPIRIT of science

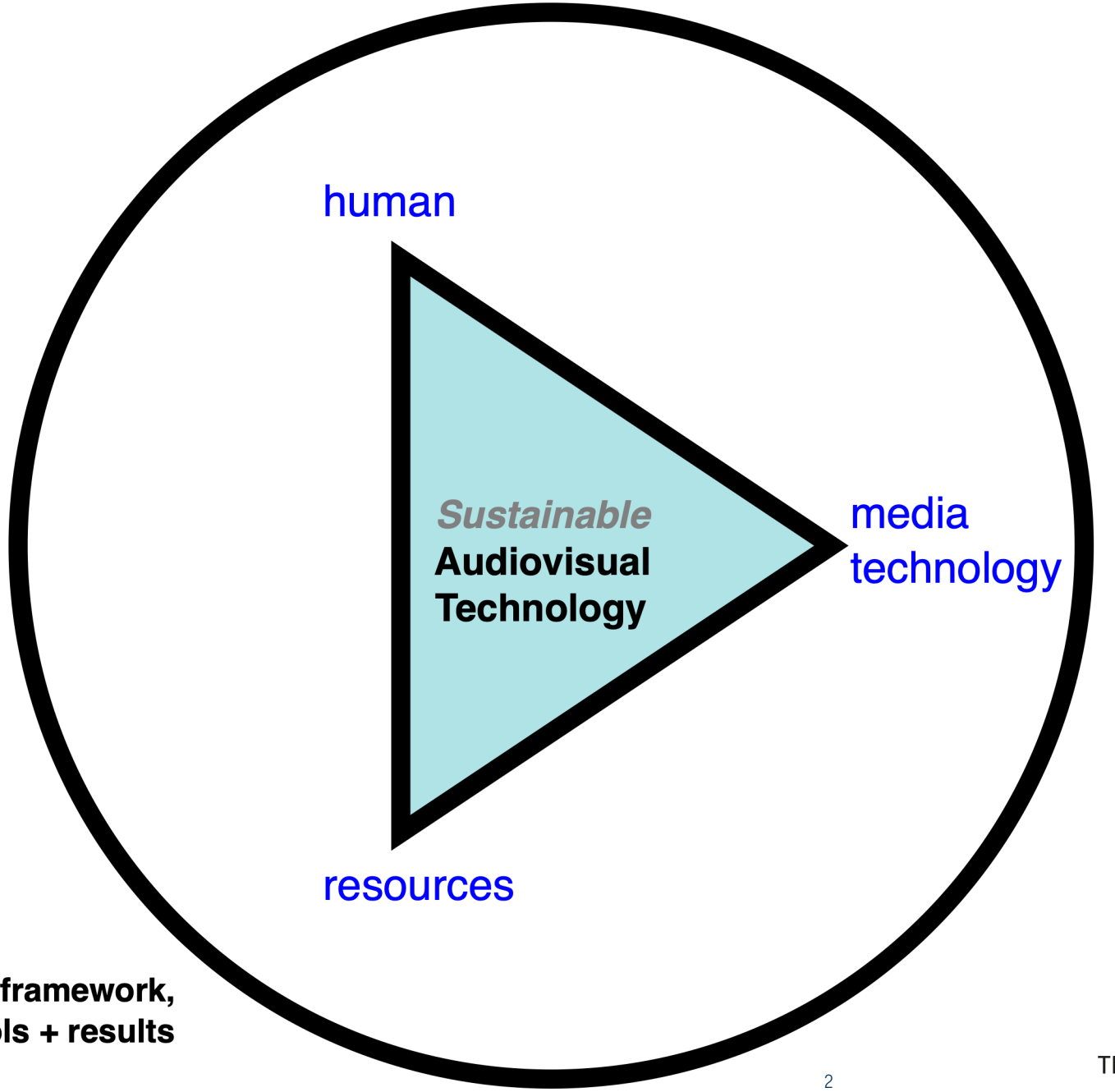
TECHNISCHE UNIVERSITÄT
ILMENAU

Alexander Raake

Audiovisual Technology Group (AVT), Institute for Media Technology

Ilmenau Interactive Immersive Technologies Center (I3TC)
http://www.i3tc.de

Photo: E. Schön, TU Ilmenau



research framework, tools + results

Audio...

… Visual

Conventional Studio

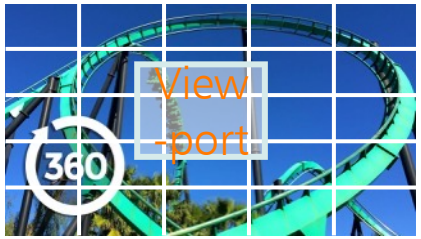
Since Oct. 2023: Greenbox
(DFG Infrastructure project “ILMETA”)

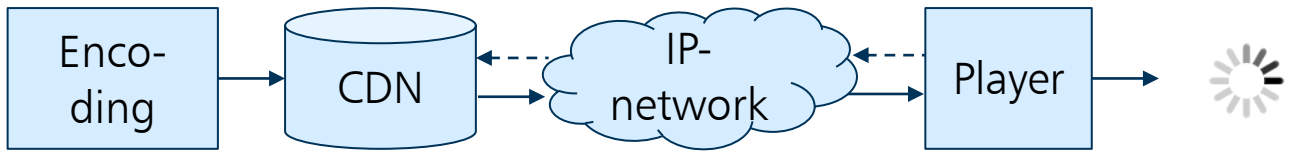
Audiovisual Technology Group

The SPIRIT of science

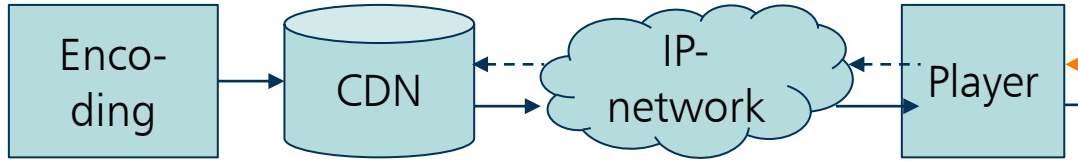
TECHNISCHE UNIVERSITÄT ILMENAU

End-to-end chain: Characterization

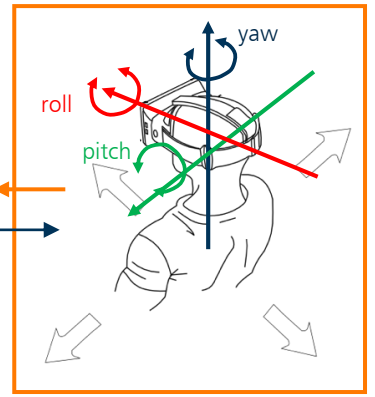


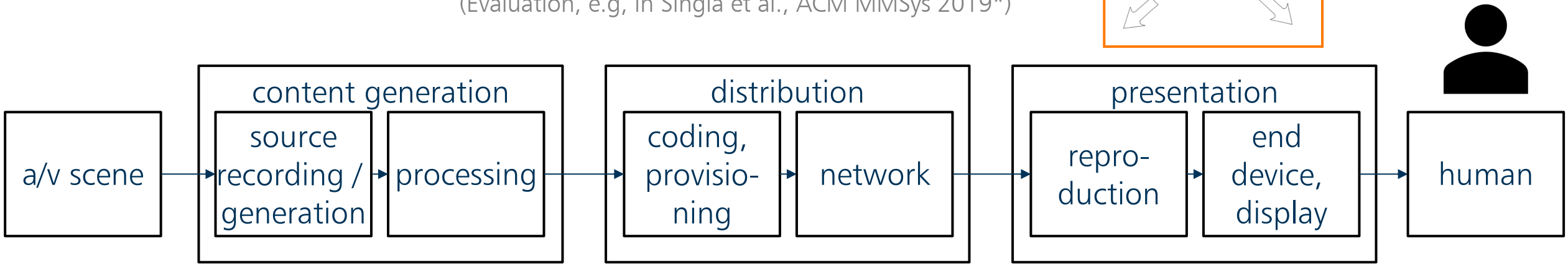


Tile-based encoding
(time, space)



(Evaluation, e.g, in Singla et al., ACM MMSys 2019*)



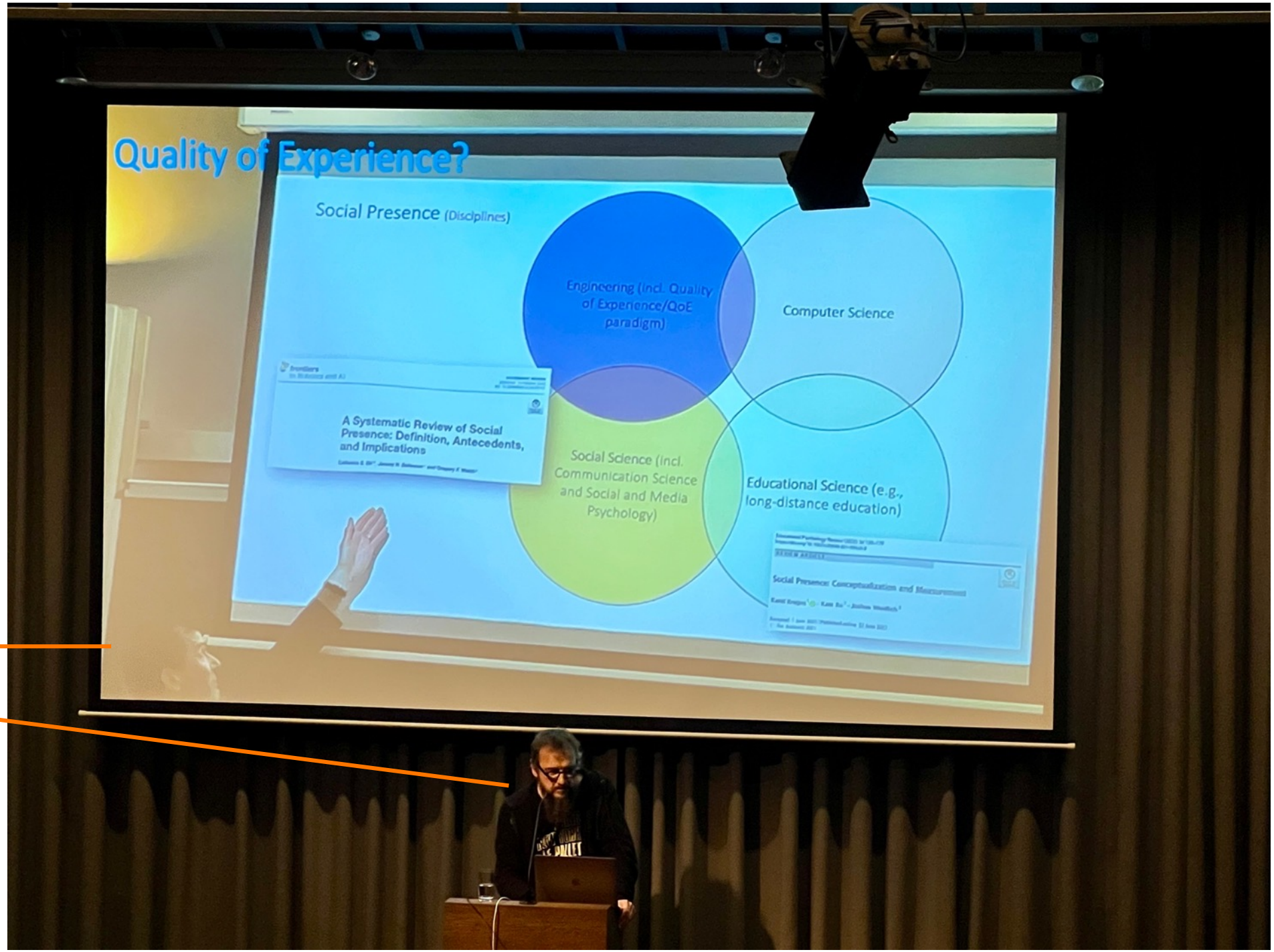


Outline

- Who am I, context
- Part 1: Audiovisual perception & considerations
- Part 2: Quality, QoE
  - Concepts
  - (Social) XR QoE
  - Constituents
  - Methods, tools, datasets, example tests...
- Summary

Building on…

e.g.,
from
Social XR
Spring
School
…



Tilo Hartmann

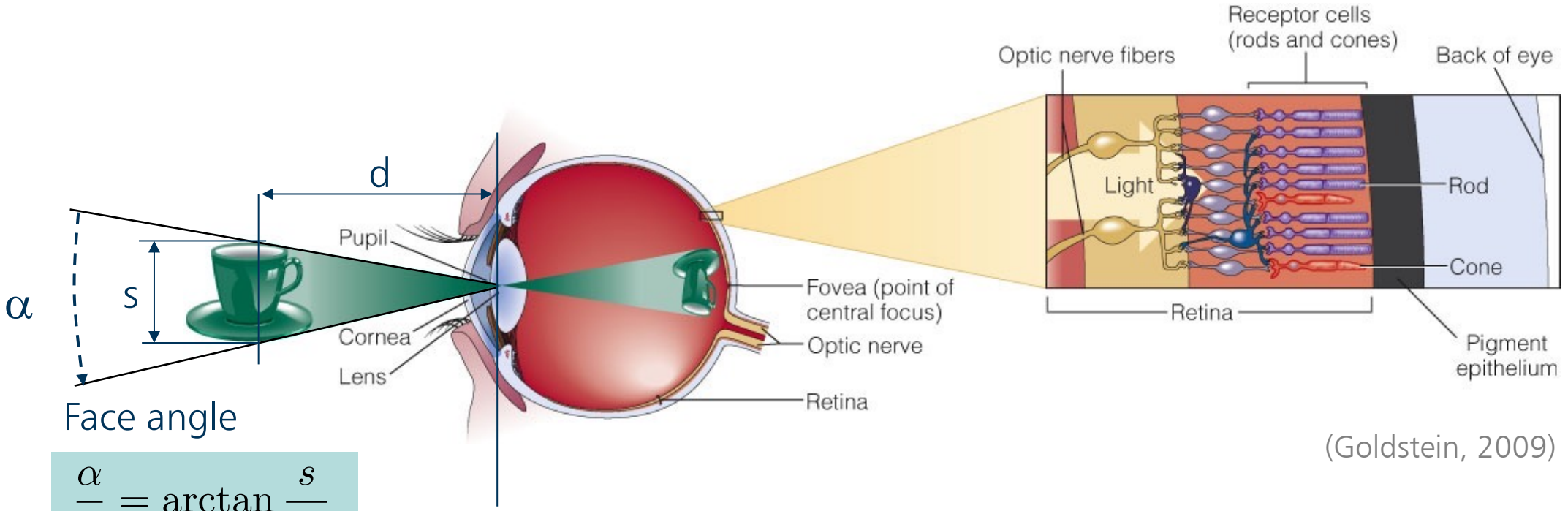
Patrick LeCallet

Mel Slater, Aljosa Smolic, Natasja Paulssen, …

Audiovisual Technology Group

Seeing - spatially organized

Projects on visual perception and quality see https://www.tu-ilmenau.de/mt-avt/



(Goldstein, 2009)

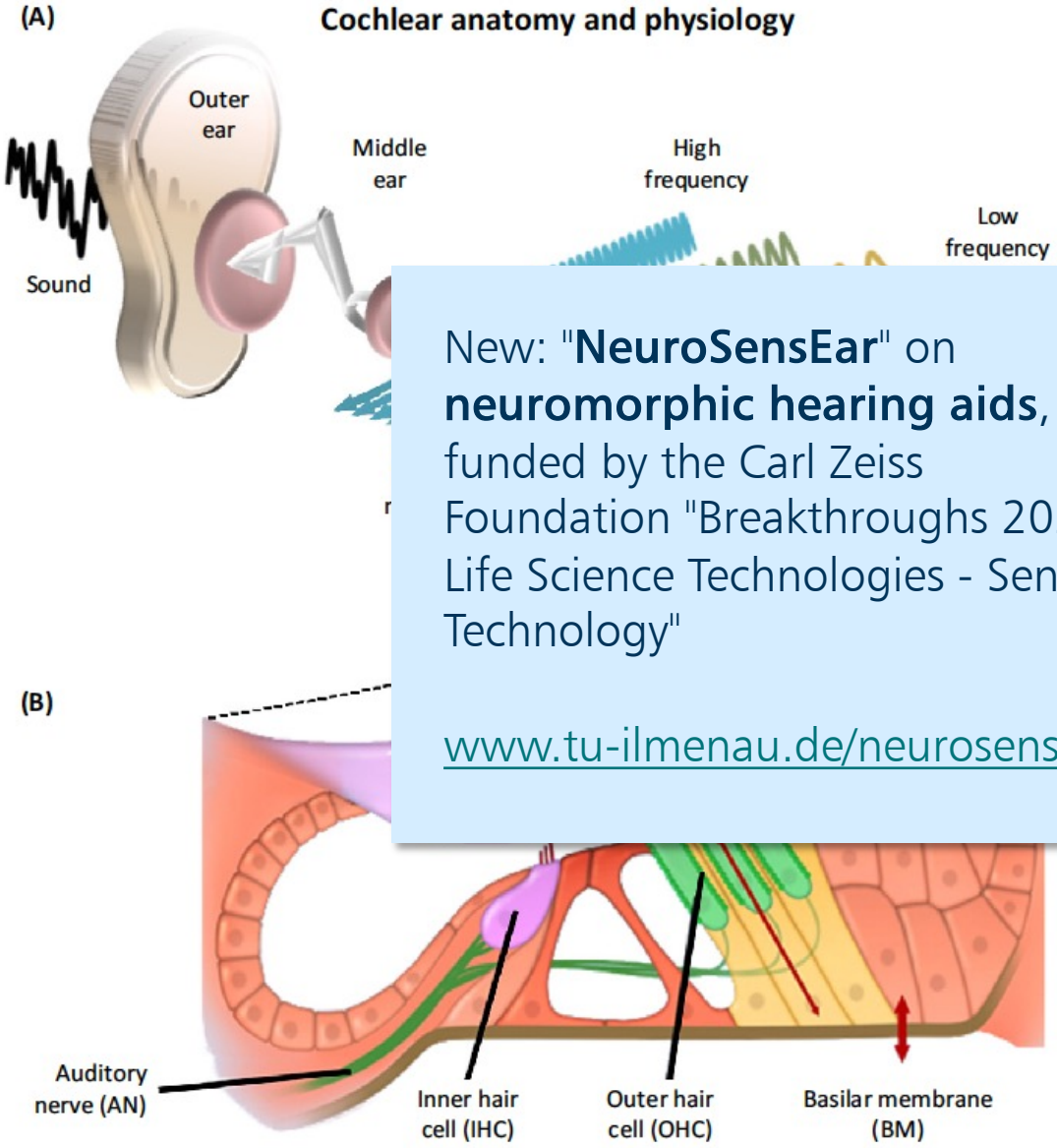
Face angle

$$\frac{\alpha}{2} = \arctan\frac{s}{2d}$$

Thumb size at arm's length: $\alpha \approx 2°$
Visual acuity: 0.4 to 1 arcmin (1 arcmin = 1°/60)

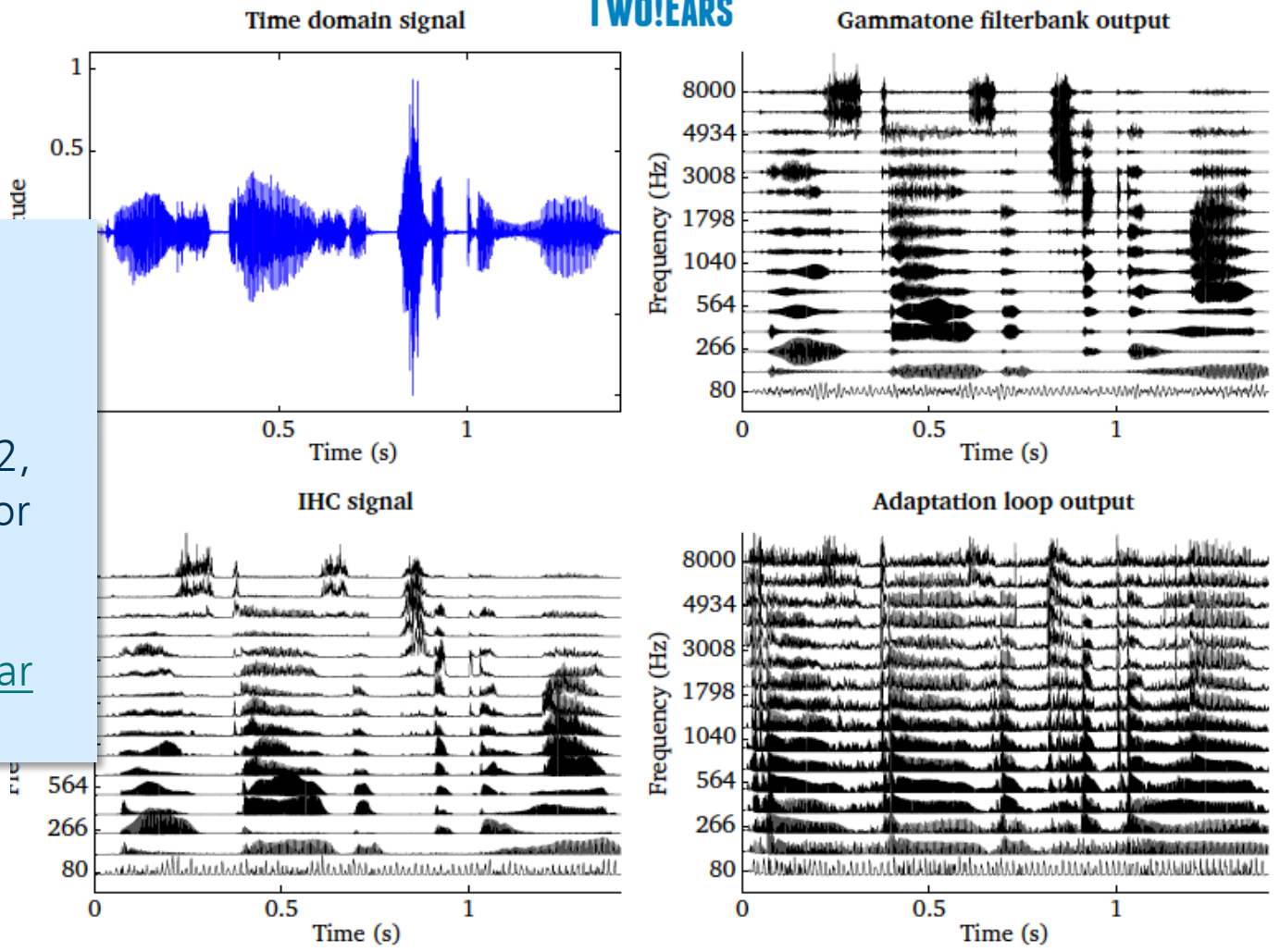
Hearing - temporal, spectral

www.twoears.eu



New: "**NeuroSensEar**" on **neuromorphic hearing aids**, funded by the Carl Zeiss Foundation "Breakthroughs 2022, Life Science Technologies - Sensor Technology"

www.tu-ilmenau.de/neurosensear



(Lesica, Trends in Neuroscience, CellPress Reviews, 41(4), 2018, ad. v. Ashmore, Physiol. Rev. 88, 2008)

(May et al. 2016; Ma et al, 2015; Brown et al, 2017; Wierstorf et al. 2017; Raake et al, 2017, Raake & Wierstorf, 2020)

Multimodal perception

- Human senses (actively) used to recognize signal sources in order to interact with the environment …
  - Objects
  - Events
  - Places / directions

  → All three types are generally multimodal

- Processing multimodal signals
  - Multimodal processing of synchronized inputs is divided into several brain areas
  - Multimodal input is (often) more "salient" and leads to stronger neuronal activity

(Möller & Weiß, in Möller, Weiß & Raake, Lecture MMI 2010, TU Berlin)

Multimodal perception
Congruent signals

- Processing of simultaneous multimodal signals
  - E.g. visual system good for detailed (spatial) information
    → predominantly spatially coded
  - E.g. hearing system good for general (temporal) information (even if not in the field of vision)
    → Information predominantly spectral-temporal coding
    - Warnings, attention control, ...
    - e.g., spatial audio for speech perception useful only for multiple sources
      (Recruitment of visual, spatial information processing, e.g. Shinn-Cunningham et al, AUDICTIVE Conference 2023)
  - Investigation of multimodal human perception: "anomalies" or special effects

(Möller & Weiß, in Möller, Weiß & Raake, Lecture MMI 2010, TU Berlin)

Audiovisual perception - discrepancies

- General observation: If signals - which are expected to come from same origin - are not congruent → Adaptation by human perception, as long as there is no clear evidence that they are not from the same origin!

- Audiovisual effect for counting (Shams et al. 2000)
  - Description: Counting of short events with auditory and visual information discrepancy
  - Example
    - Three visual stimuli in succession (flashes) → How many flashes?
    - More auditory stimuli → More flashes perceived
    - e.g. http://shamslab.psych.ucla.edu/demos/
  - Result: Audio dominates cross-modal perception → temporal component

Audiovisual perception
Discrepancies – ventrilloquist effect

Source: Sascha Grammel, YouTube

Audiovisual localization ("ventriloquist's illusion", e.g. Harris, 1965)

- When auditory and visual information are not congruent, auditory information is usually "captured" by visual information
  - Sources of both modalities associated with location of visual information
- Effect depends on
  - Spatial distance
  - Intensity of the auditory and visual stimulus (e.g. Radeau, 1985)
  - Accuracy of the senses: If visual information is very blurred (approx. 10 degrees), auditory information can dominate (Alais & Burr, 2004)
  - Movement: Visual movement can lead to perceived auditory movement (Soto et al., 2002)

Audiovisual perception
Discrepancies – McGurk effect

Audiovisual localization / identification phon ("McGurk effect", McGurk and MacDonald, 1976)

- Example
  - Auditory: /baba/
  - Visual: /gaga/ /dada/
  - Audiovisual: /dada/ /dada/

Source: BBC, YouTube

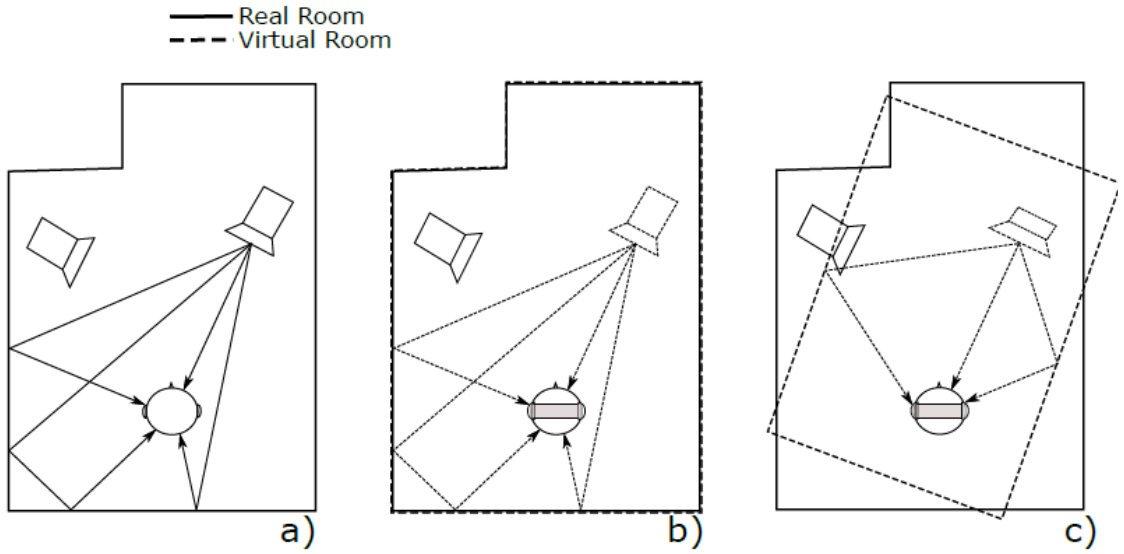
→ Fusion, multimodal integration
  - New articulatory "place"
  - Intonation not affected (/d/, not /t/)
  - But: visual: /baba/, auditory: /dada/ /bdabda/ (combination, not fusion)

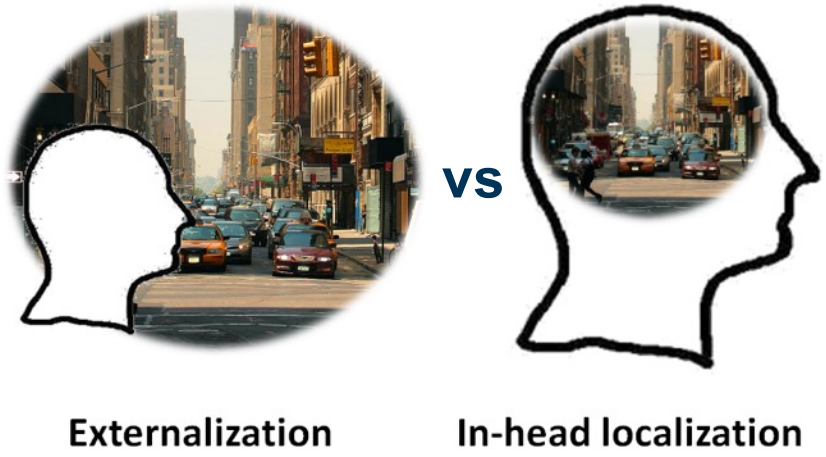
Discrepancies – Spatial divergence effect

Perceptual match

Central research question ...

How exactly must the room acoustic properties be reproduced in order to create a perceptual spatial match?





Left: Neidhardt A, Schneiderwind C, Klein F. Perceptual Matching of Room Acoustics for Auditory Augmented Reality in Small Rooms - Literature Review and Theoretical Framework. *Trends in Hearing*. January 2022. doi:10.1177/23312165221092919

(Slide by Werner et al, FG EMT TU Ilmenau)

Cocktail Party Effect

(Image: Wikipedia.org,
Financial Times - Patrón cocktail bar)

Spatial effects: Multiparty communication – Cocktail Party

- In such situations, **spatial position becomes important for hearing → Spatial Audio**
  - **Intelligibility increased** (mainly auditory effect: Disambiguation of sound sources, denoising: Better-ear listening, interaural effects such as “binaural mask”)
  - **Recruiting neural (spatial) visual network (!)**
    (Michalka et al. Neuron 2015, Michalka et al. Cereb. Cort. 2016,
    Noyce et al. J. Neurosci. 2017; cited in Shinn-Cunningham, Keynote IEEE QoMEX 2021,
    AUDICTIVE Conference 2023)
    - Remembering more precisely who said something (Baldis, CHI 2001; Raake et al., AES 2010; Skowronek & Raake, Speech Comm. 2015)
    - Mental effort reduced...

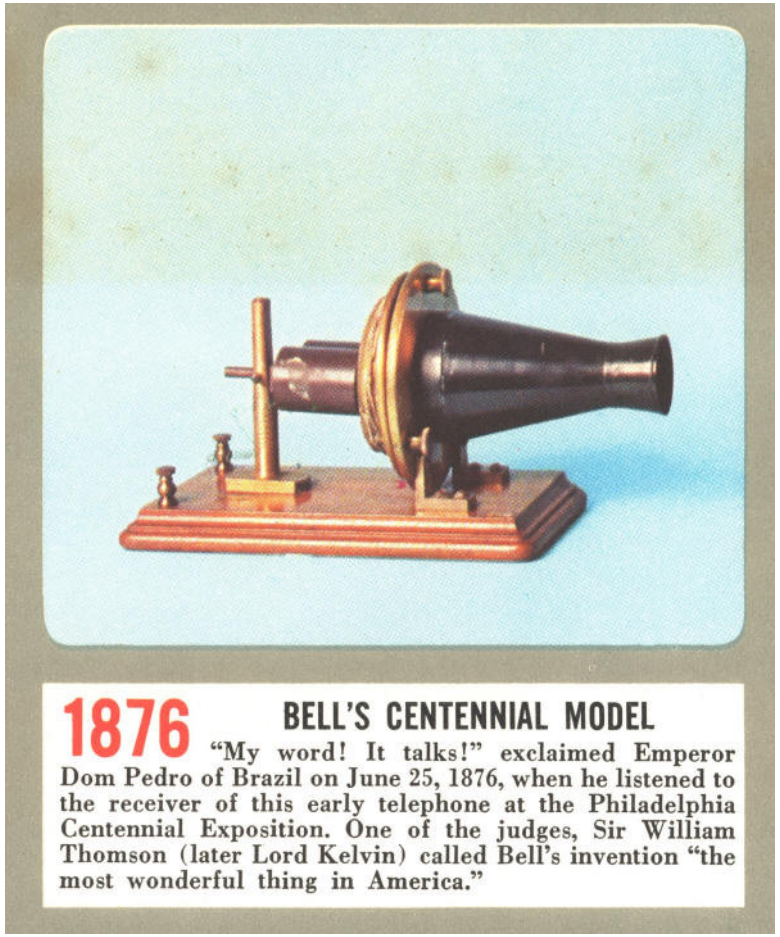
Cocktail Party: Spatial Audio in Communication

- Listening situation
  - Spatial reproduction typically preferred over non-spatial reproduction (Baldis, 2001; Raake et al., 2007)
  - Positive impact on cognitive load, speech communication quality and Quality of Experience (Raake et al., 2010; Skowronek & Raake, 2015)
  - Interactive Virtual Environments: higher immersion and audio-visual quality (e.g. Potter et al., 2022)
- Conversation situation
  - Increased perception of interactivity, shared space, ease of understanding (Nowak et al., 2023)
  - Immersive communication systems
    - Social presence as one commonly used quality indicator
    - Studies show high social presence for such systems, but not considering spatial audio (e.g. Smith and Neff, 2018; Li et al., 2021)

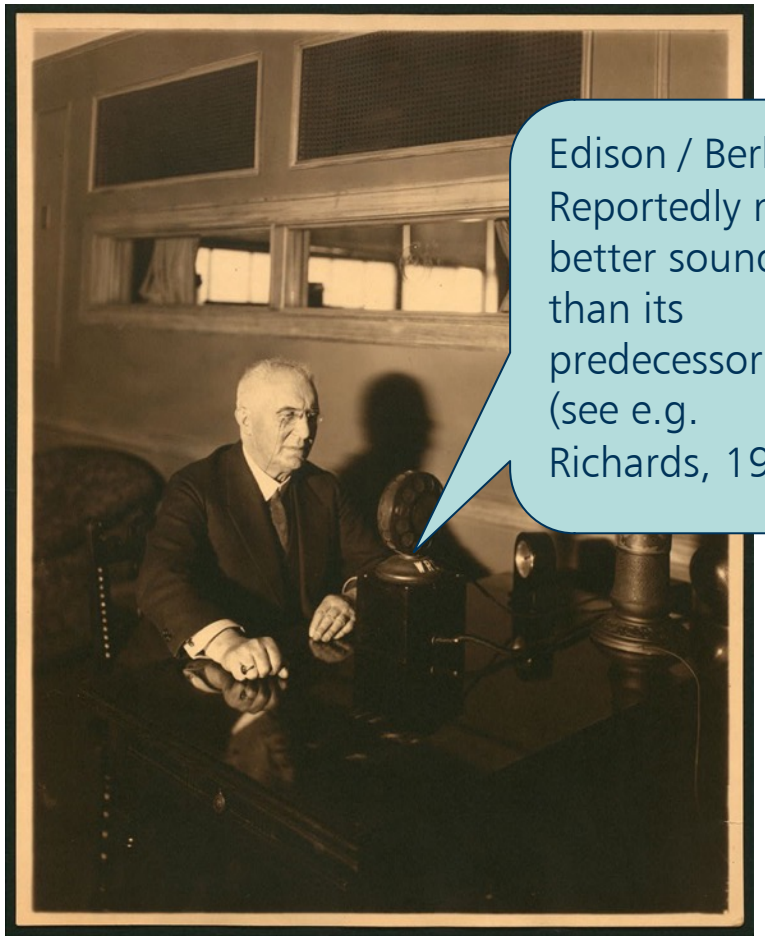
Multimodal perception
Summary / explanations of effects

- Visual dominance?
  - In many early studies, vision dominates other modalities (audio, kinaesthetics)
  - But: mostly spatial tasks in which vision must be precise and reliable.
  - Auditory control
    - e.g. so-called "turn-to-reflex" (see Blauert & Brown, Springer, 2020; Corbetta et al., Neuron, 2008; Petersen & Posner, Review of Neuroscience, 1990, 2012)
    - Auditory event from outside the current field of vision → Attention can be directed by audition (e.g., Singla et al., QoMEX 2023)
- Appropriateness of modality - situational appropriateness determines dominant modality
  - E.g. audio for time-dependent tasks, visual for spatial tasks
- Integration, if applicable, if most expedient / plausible approach
- Recruitment of vision-networks for auditory perception and selective attention improving scene analysis
  - And vice versa...

Quality: From early days of media quality (not the earliest)…

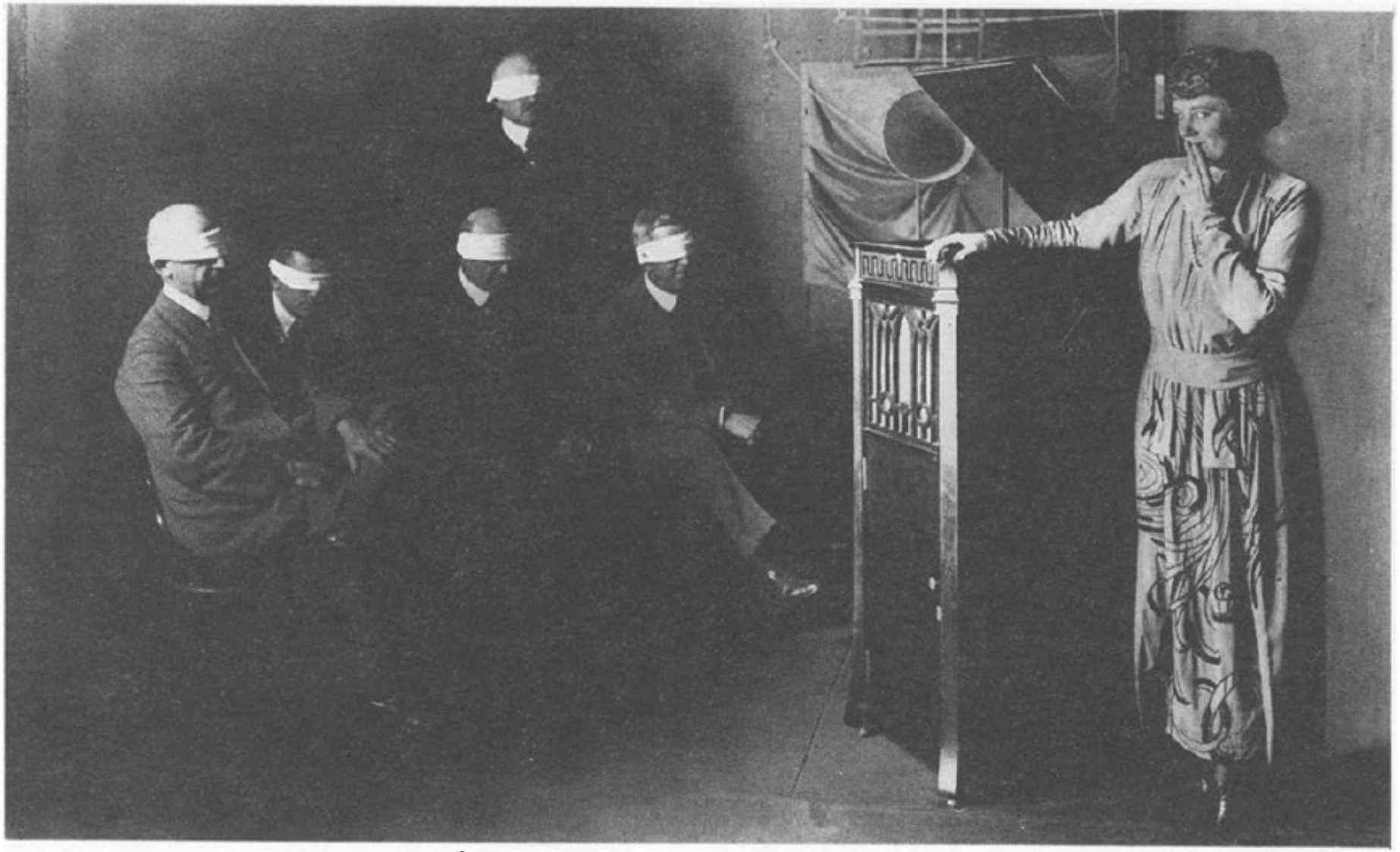


Bell receiver from 1876



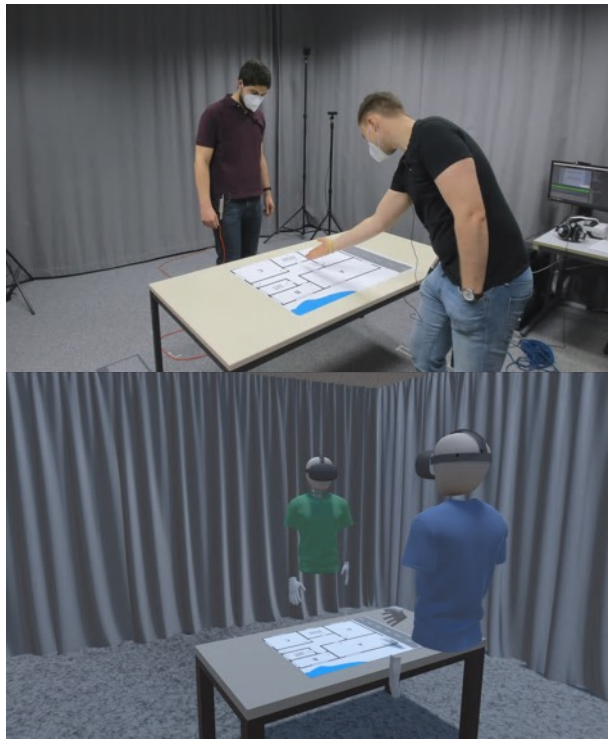
Carbon microphone (here Berliner with his 1877 invention)

Fidelity



Publicity photograph of Frieda Hempel, Edison recording artist, with Edison employees, ca. 1918 (United States Department of the Interior, National Park Service, Edison National Historic Site, from "Machines, Music and the Quest for Fidelity: Marketing the Edison Phonograph in America, 1877-1925", Emily Thompson, Musical Quarterly, 1995)

...towards holistic QoE evaluation of Interactive XR



(Immohr et al., IMX 2023, IEEE VR 2024)

(Source: https://docs.metahuman.unrealengine.com/en-US/UserGuide/, image retrieved Sept. 2022)

https://zenodo.org/communities/audiovisual_scenes?page=1&size=20

(Llorca-Bofí, Vorländer, van de Par, Ewert, Seeber, Kollmeier, Grimm, Hohmann et al., 2018-2022)

Quality

Product / Service quality
"Extent to which **product or service meets / exceeds** customer's **expectations**" (Reeves and Bednar,1994)

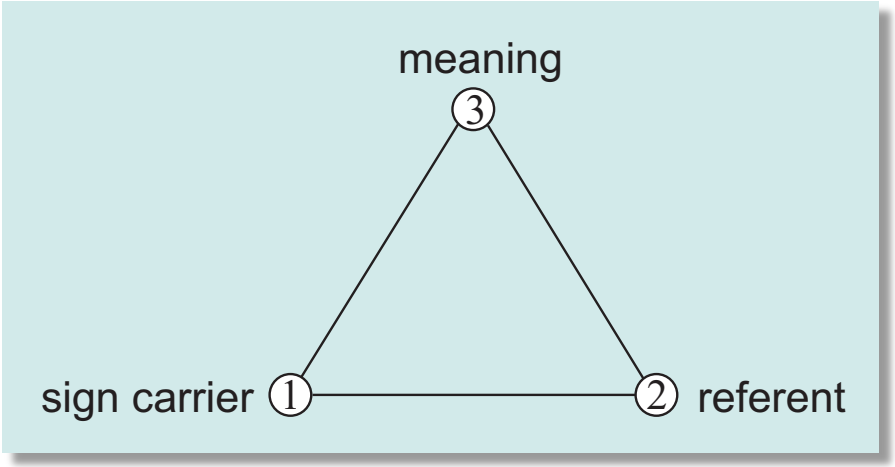
Media quality
"…result of the **judgment of perceived composition** of entity with respect to its **desired composition**." (Jekosch, 2005)

(Photo by Anni Roenkae: https://www.pexels.com/photo/multi-color-painting-2457278/)

Media quality

- Media (i.e. speech) quality according to Jekosch 2005
  - Aka Basic Audio Quality, video quality…
  - Vast and reproducible work re audio & video technology etc. (AES, ITU, VQEG, MPEG...)
  - Good (yet not perfect) understanding of procedures and biases, cf. Bech & Zacharov, Zielinski et al., Pinson et al., …
  - See ITU-R / ITU-T standards such as MUSHRA, P.800, P.910, BT.500, P.130X-series...
- Evaluating person is aware of technical system or at least form / carrier of the media information, assesses it directly, for example
  - …taking part in a media quality test (listening, viewing, ...)
  - …trying out different systems for purchase in store
- Adapted from Mausfeld (2003) "Dual character" of (picture) perception. Person can take 2 perspectives, focusing
  1) on the medium that is employed, in terms of an artifact, for example paying attention to sound features related with audio system (i.e. also in relation to carrier)
  2) on the movie, visual or auditory scene, or musical piece... presented to her (i.e. the content, meaning...)

  → cf. also Hartmann & Hofer, Frontiers in VR, 2023



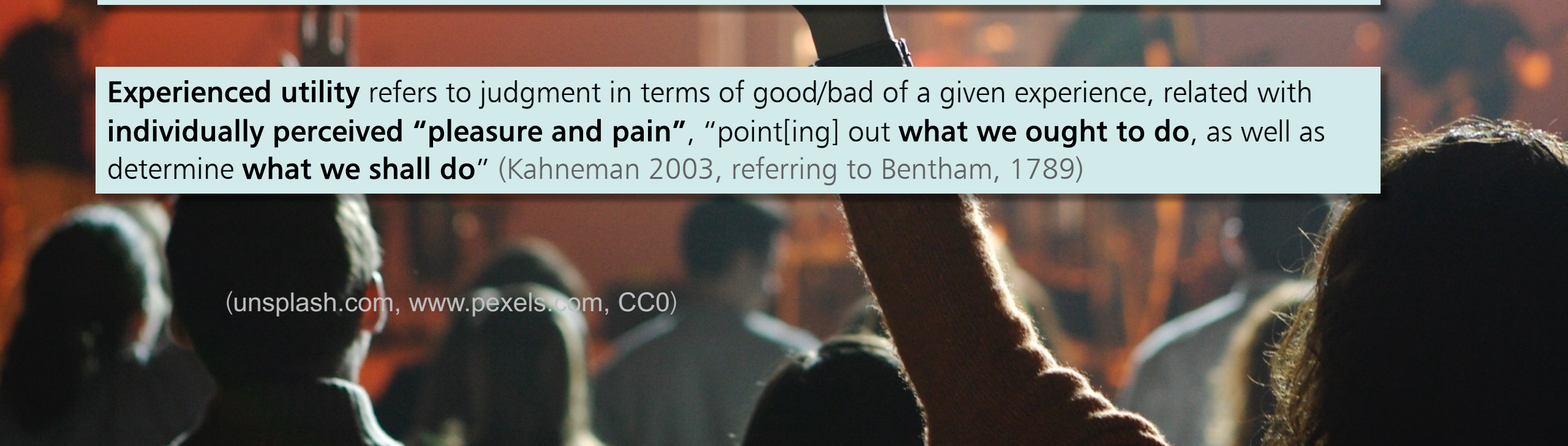
Semiotic Triangle (general form: Nöth, 1990)

Quality, Quality of Experience (QoE)

Quality of Experience (QoE)

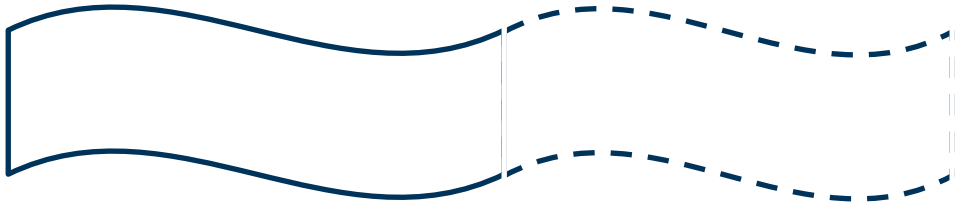
"… **degree of delight or annoyance** of a person whose experiencing involves an application, service, or system. It results from the person's **evaluation of the fulfillment of his or her expectations and needs** with respect to the **utility and / or enjoyment** in the light of the person's context, personality and current state.
(Raake and Egger, 2014, adapted from QUALINET QoE Whitepaper, www.qualinet.eu, 2012. Adopted as QoE definition by ITU-T P.10 in 2016)

Experienced utility refers to judgment in terms of good/bad of a given experience, related with **individually perceived "pleasure and pain"**, "point[ing] out **what we ought to do**, as well as determine **what we shall do**" (Kahneman 2003, referring to Bentham, 1789)



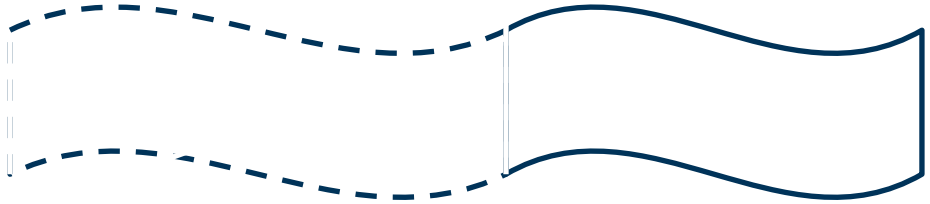
(unsplash.com, www.pexels.com, CC0)

From media quality to QoE
"QoE testing ruler"



„Classic“ media quality testing
- repeated, ca. 5-10s, "neutral" sources
- focus on (technical) quality / carrier

„Ecologically valid“ video quality testing



"QoE" testing

(adapted from Garcia & Robitza, 2016)

User Experience (UX)

User Experience (UX)

(1)

"A person's **perceptions** and **responses** that result from the **use or anticipated use** of a **product, system or service."**

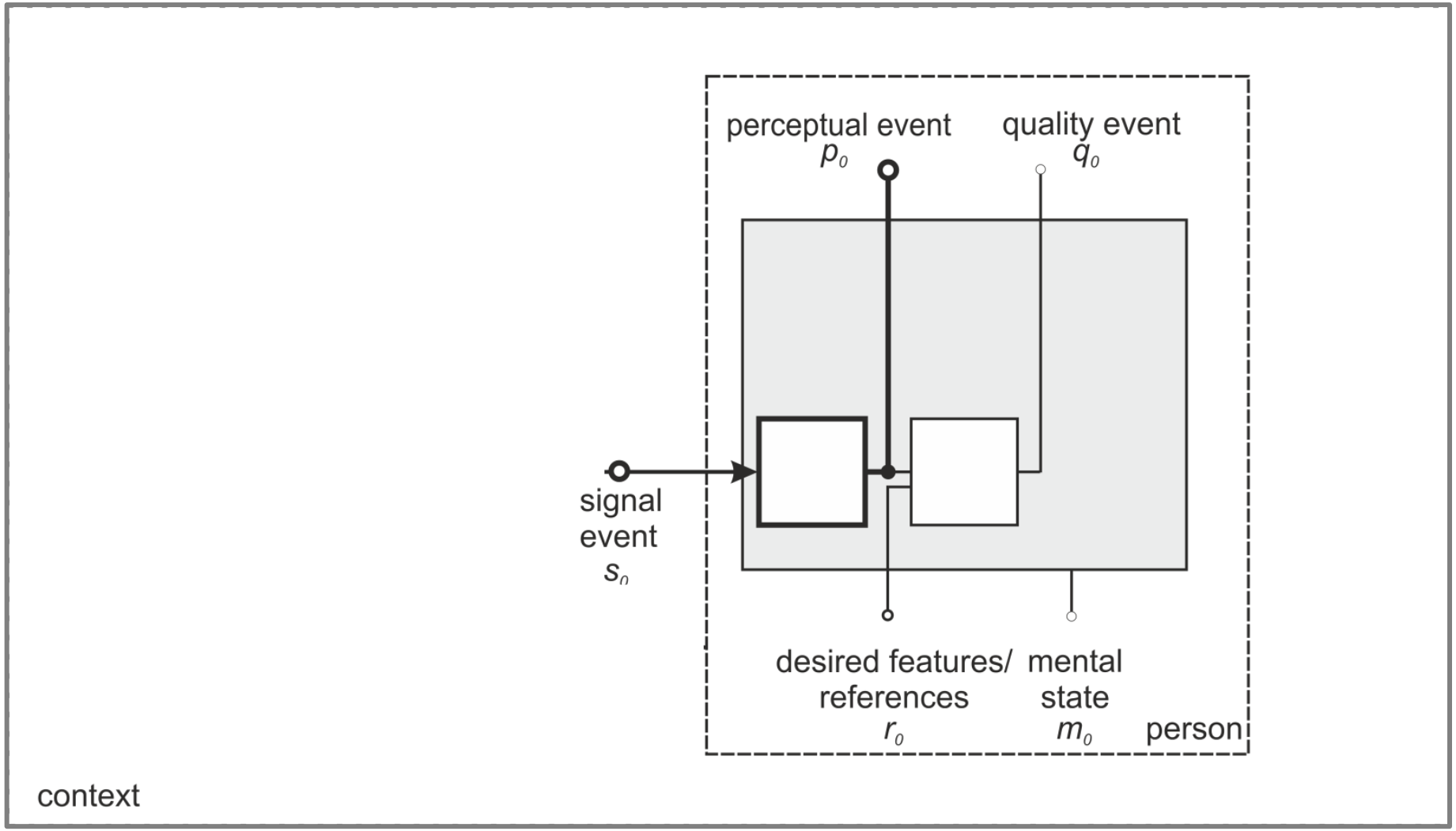
(ISO DIS 9241-210:2008. Ergonomics of human system interaction - Part 210: Human-centred design for interactive systems (formerly known as 13407, now withdrawn), cited from Law et al., CHI 2009)

(2) described as

"dynamic, context-dependent, and **subjective,** stemming from a broad range of **potential benefits users may derive** from a **product"**
(Law et al., CHI 2009)

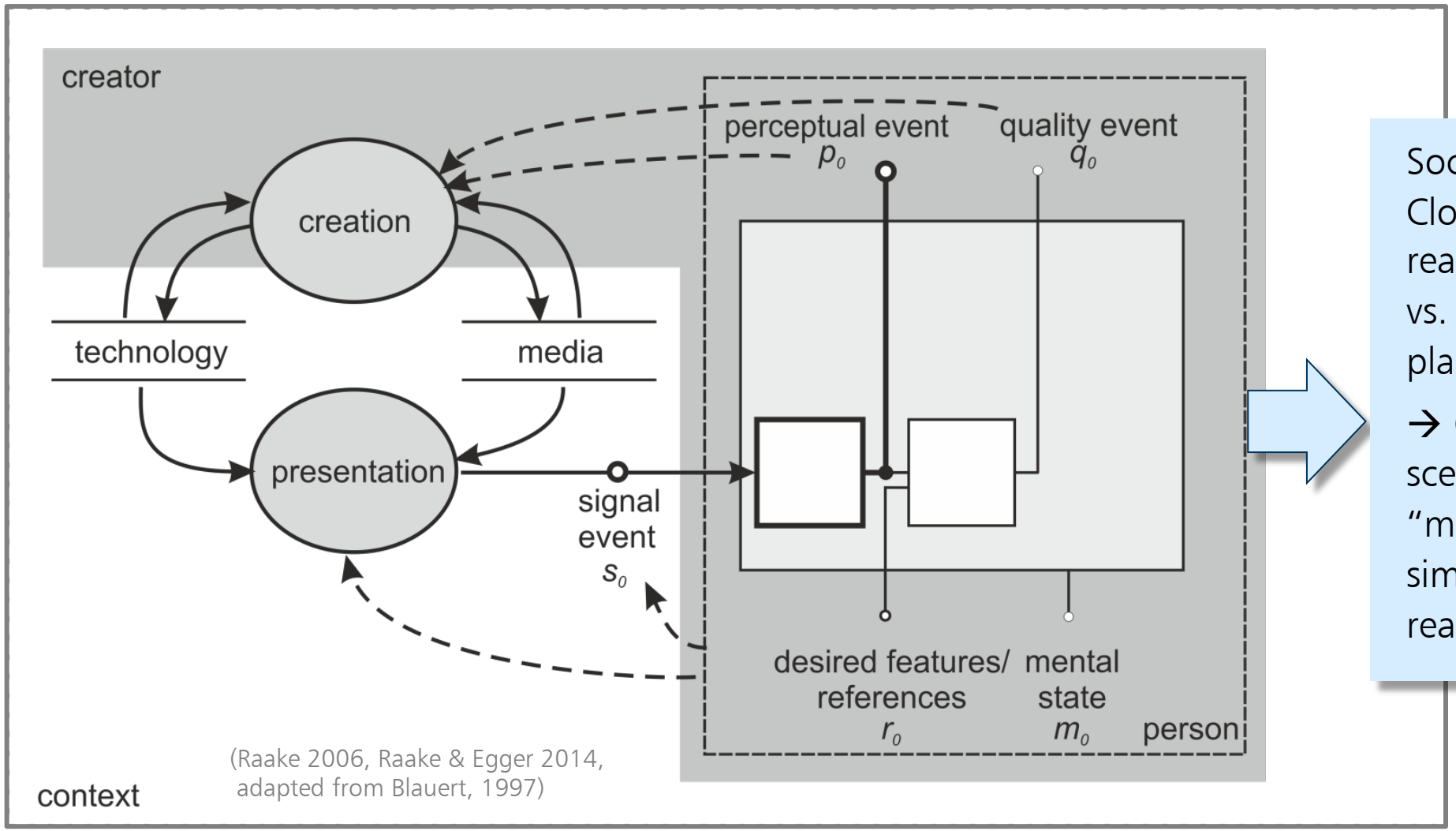
→ see also Wechsung and de Moor, "Quality of Experience vs. User Experience", in: Quality of experience: advanced concepts, applications and methods, Springer, 2014 (Möller & Raake, edts)

Quality / QoE / UX formation
Perception & creation



(Raake 2006, Raake & Egger 2014, adapted from Blauert, 1997)

Quality / QoE / UX formation
Perception and creation, perception by creator

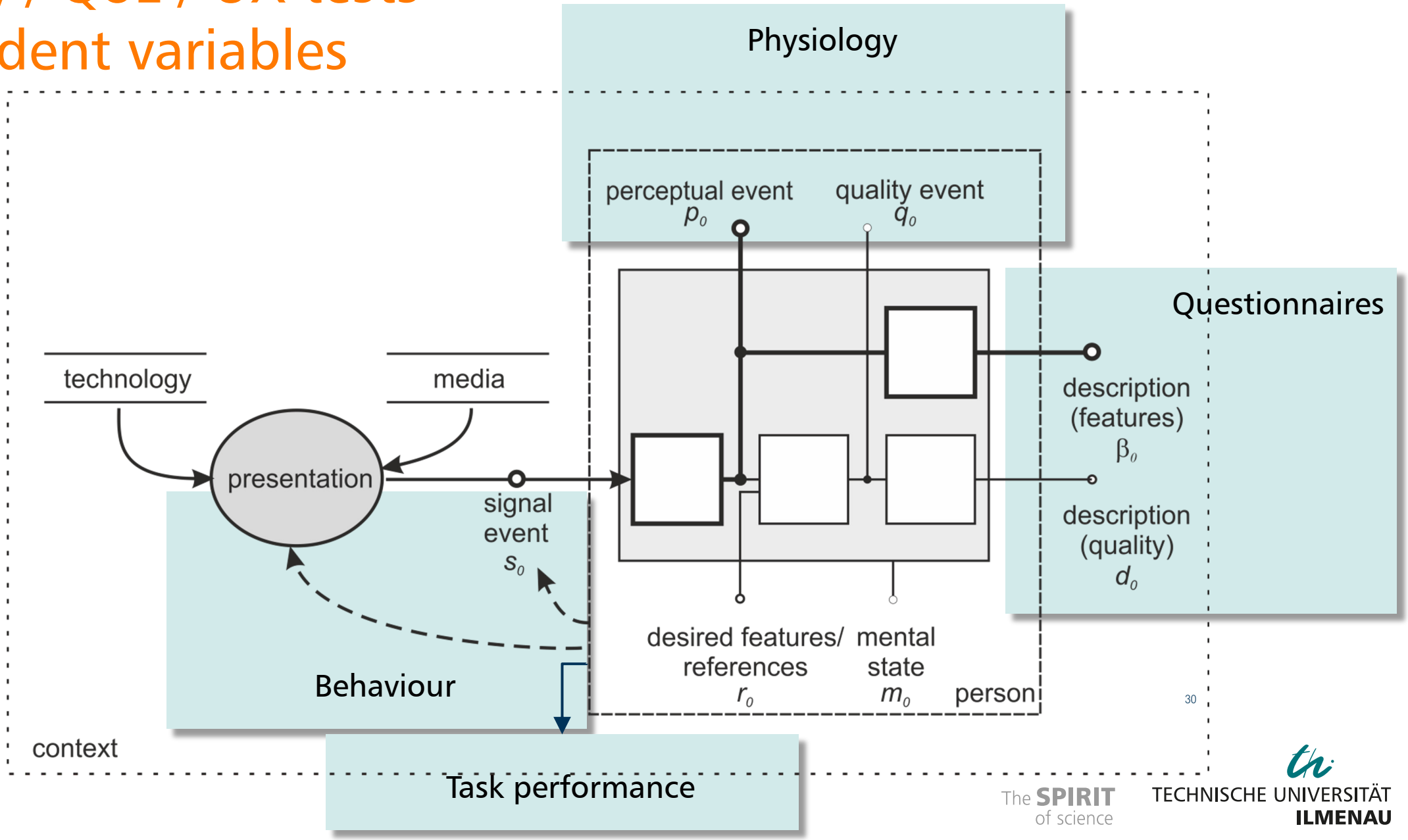


(Raake 2006, Raake & Egger 2014, adapted from Blauert, 1997)

Social XR: Close-to-real-life-scenes vs. gamified places

→ Curated scenes vs. "measured" / simulated reality!

Quality / QoE / UX tests
Dependent variables

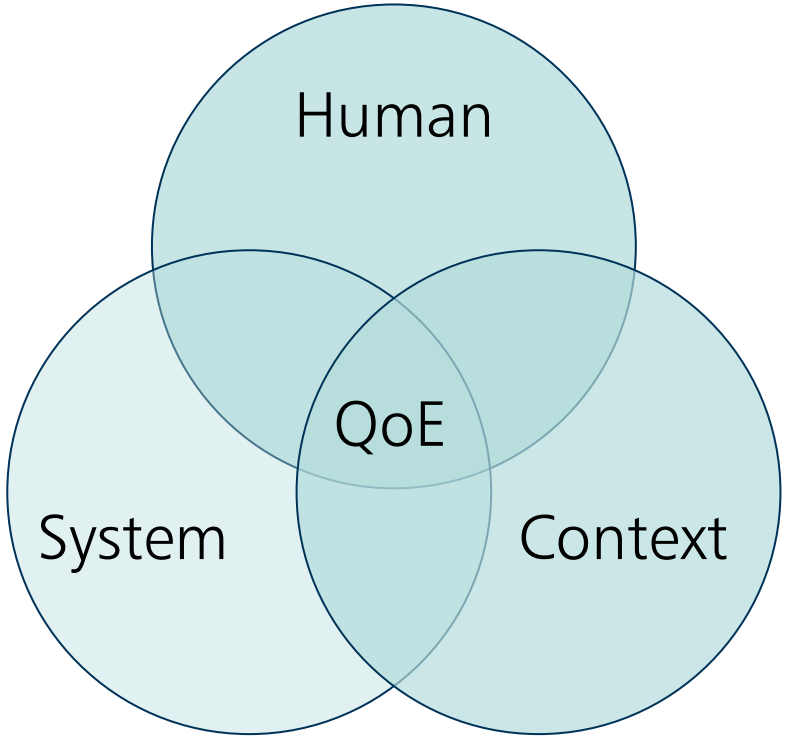


Audiovisual Technology Group

The SPIRIT of science

TECHNISCHE UNIVERSITÄT ILMENAU

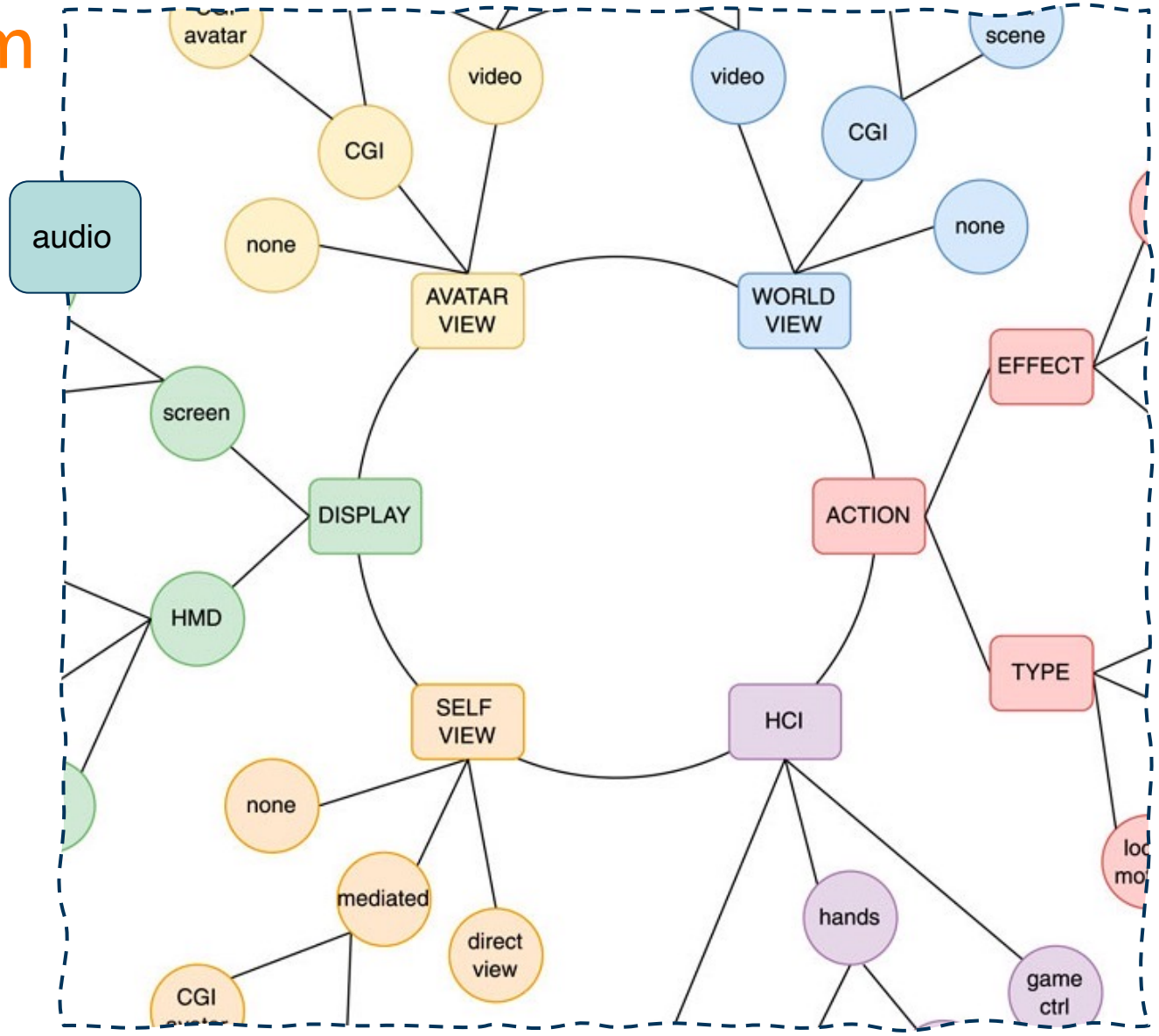
Quality/QoE/UX Influence Factors (“independent variables”)
Note: Some can be re-influenced by QoE!



(vgl. Reiter et al., in Möller & Raake, Quality and Quality of Experience, Springer 2014)

→ cf. talk Patrick Le Callet

Social XR System – components



(adapted from Pérez, Gonzalez-Sosa, Gutiérrez, García, Frontiers Signal Proc. 2022)

What from natural meetings can't be done so easily with "classic" video telephony technology?

From our point of view, for example

- Issues like Zoom / video conferencing fatigue
  (cf. Döring, Schoenenberg, de Moor, Fiedler & Raake, 2022, IJERPH; Raake et al., 2022, arXiv; Bailensson et al.)
- Turning towards - directing attention to individuals, dialog
- Averting → Creativity reduced, possibly because less "defocusing" is possible (Brucks & Levav, Nature 2022)
- Moving more freely in front of a screen (Bailenson, Technology, Mind, and Behavior, 2021) or always keeping the other(s) in view, following someone into another room
- Remembering more precisely who said something (Baldis, CHI 2001; Raake et al., AES 2010; Skowronek & Raake, Speech Comm. 2015)
- Natural interaction, e.g. showing something

Quality/QoE/UX “constituents” for Interactive XR
Dependent variables

Constructs

- Media quality
- Plausibility, authenticity
- Presence: Spatial presence, social/co-presence, self-presence (cf. talks Hartmann, Slater, ...)
- Cybersickness/Simulator Sickness; comfort; fatigue
- Emotional response: Own, perceived
- Cognitive performances & constructs, e.g.,
  - Intelligibility (SRT + corpora, ...)
  - Communication effectiveness and efficiency
  - Scene analysis (localization, time of arrival; identification, ...)
  - Perceived personality traits, states, ...
  - Task effort (NASA TLX, ...)

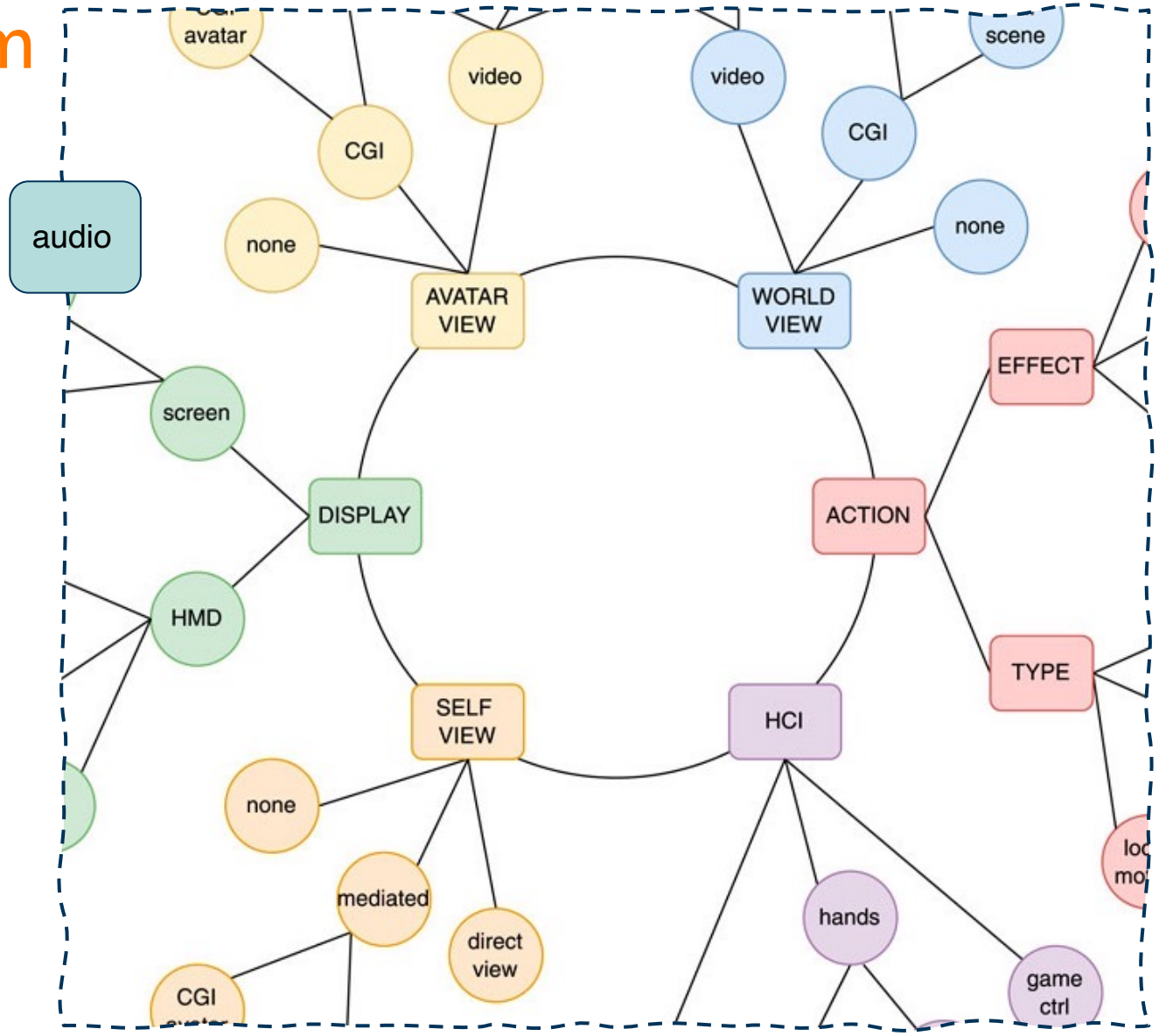
Representation

- Questionnaire-based (ACR, IPQ, SSQ, NASA TLX...)
- Behavior
  - Exploration (head & eye movements, translatory, ...)
  - Human-scene-interaction (“agency”)
    - Objects; other persons (conversation analysis, ...)
  - Emotional (facial gestures, ...)
- Physiological
  - EEG, skin conductance, heart rate, pupil dilation, ...
- Performance
  - Task completion time, error rates, intelligibility in percent or SRT, ...

“direct”

“indirect”

Social XR System – components



needed

- → Joint view of ecosystem to systematically develop evaluation approaches
- → Scenarios
- → Standardization? Open Science?

(adapted from Pérez, Gonzalez-Sosa, Gutiérrez, García, Frontiers Signal Proc. 2022)

Example Standardization: VQEG-IMG / ITU-T P.IXC

- Motivation: How to test bi-directional immersive communication system?
- VQEG-IMG & ITU-T SG12: New **recommendation** for subjective assessment of XR communications:
  - Test design description: System influencing factors to test, how to control context and human influencing factors, which QoE constituents to address.
  - Reduced set of suitable communication-based **interactive tasks**
    - 4 tasks: audio communication, visual communication, object manipulation, and environment exploration
  - Subset of relevant **measures**: questionnaires regarding QoE constituents, physiological measures, behaviour analysis, etc.
- Test plan:
  - 14 international labs running experiments.
  - Tentative schedule: Results reported in the next VQEG meeting and recommendation in 2025
- **Consider participating in VQEG-IMG → Jesús Gutiérrez**

Summary & Outlook

Extension: overview of QoE evaluation methods, tools and resources

- Audiovisual perception – short overview and some effects
- Established methods and tools for trad. media quality and QoE
  - Some open points also there
- More work for QoE assessment needed, ideally using systematic joint effort, e.g.,
  - What actually is Social XR – general system model(s)
  - Scenes
    - Avatar / person representation assessment
      - ➢ Semantics such as facial expressions, … starting point: Volumetric video quality
    - Other scene elements, background, …
  - Inclusion of audio, audiovisual, tactile, …
  - Interaction and behavior
    - Scenarios
    - Trajectories
  - Longer-term studies – QoE also due to novelty vs. long-term gain in Quality of Life
  - Comparing benefits compared to videoconferencing, F2F, gamified vs. real-life XR, …

Thank you!
Comments, questions?

Happy to exchange and collaborate ☺

https://www.tu-ilmenau.de/i3tc
https://www.tu-ilmenau.de/en/audio-visual-technology/