Transcript: “Smash the vase”

From Anthropic (YouTube Channel),
episode “What should an AI's personality be?” (08.06.2024),
36:21 bis Ende
URL: https://www.youtube.com/watch?v=iyJj9RxSsBY

- Amanda Askell (Alignment Finetuning Researcher at Anthropic)
- Stuart J. Ritchie (Research Communications at Anthropic)

This is the end of an expert interview produced by the US-based AI startup company Anthropic for promotional purposes. Anthropic prides itself to be a company doing “AI research and products that put safety at the frontier”[1]. Its main offering is the Large Language Model (LLM) “Claude”.

Amanda Askell is a well-known expert for AI ethics and AI alignment and serves in various interviews (podcasts and YouTube) as the public figurehead of Anthropic for these topics. Stuart J. Ritchie is a Scottish psychologist also working for Anthropic doing “research focused comms”[2].

The recording of the interview is 37:40 minutes long. Mr Ritchie clearly takes the role of explaining and supplementing Ms Askell’s statements for non-expert listeners. He does this in a rather forceful manner, often strongly interrupting Ms Askell. His backchannels are particularly loud and hearably disturb Ms Askell’s talk at some points. So much so that one YouTube user comments “Hey, Stuart. I'd like to hear Amanda finish a thought, if you don't mind.” With 4 other users chiming in with the same sentiment.

Research questions

I am interested in extended trouble-induced side sequences in interviews launched by the interviewer dealing with problematic (offensive, delicate, mispronounced etc.) utterances. In this excerpt, I am especially interested in:

- How does the sudden topic change during an expert interview play out:
  - Orientation to and handling of the intervention
  - Shift in epistemic stance and status of the two interactants
  - Attempts to return to the institutional context and the current interview topic
- How ist the “issue” acknowledged and the “wrong-doing” corrected?
- Is there a recognizable release of tension during and at the end of the intervention?
- The role of overlap and competing for the floor
- Intonation and laughter marking the handling of trouble and changes in stance
- Gender, Identity, Personality: How do they come into play in this excerpt?

[1] See homepage of anthropic.com (accessed 28/03/2025)
[2] Twitter: https://x.com/StuartJRitchie/status/1744764822011699553

Ama >If you were to< sho: ↑excessive empathy .h
>eemee ya could imagine like someone shewin excessive empathy< to
leyk: (.) to objects in the #wo:rld#. .h >and being [like,]<
Stu [>Ya<]
Ama O:h >you-<you should go to pri:son: if you like if you smash the
vayz ↑>an I'm like< loo:k I think it's↑ good to not get in the habit
of like [h smashing objects.]
Stu [S:orry can I just stop] you [there]?
Ama [>ya ya<]
Stu You:'re Sco:ttish and you just said vay:z.
(.)
Stu That'[s-]
Ama [O:h] is [that?]
Stu [er:]::: >You kent say that< [it's vah]z
Ama [iszat]
=>Iszat Amer[ican?<]
Stu >[Ye-high long] hv you been in America?<
Ama £[I've been in America for like thirteen] yearsh£ hh h h h h
Stu [Yea, >How long have you been in America¿]
Stu TOO: LO:NG [clearly it's been too long]
Ama [.hhi ha ha ha]
Stu =B'cause you say vay:z
Ama =↑£Smash the vayz on the sidewalk£↑ HH [h h]
Stu [>Carry on,<]
Stu ↑Say-o:h my ↓God o:h
(.)
Ama .hh [£o:h->I neva-even<↑]
Stu [(>terrible< I k-)]
Ama =>Uh-uh<-Ive-f:orgotten things £that are Scottish and aren't, so£
[didn va:z, okay, va:z heh]
Stu [Hyea well, we cert'nly don't] say-
=no one in this country's ever said vay:z. >Let me [tell ya.< (.)]
Ama [↑£Okayokay£↑]
Stu A:h
Ama =smash [the va:z.]
Stu [>↑Pleas-carry-on!<]
(.)
Stu [>Pleas-carry-on,<]
Ama °[yea-okay]
=s:o-like u::hm°(.)
Yea, ↑so >i-if ya wa to say to peo:ple<↑ o:h you should like go to
pri:s'n: fer fer smashing a vaz £u::hm: hh the:n£
Stu =>sankya<
43 Ama like ghh h h .hih tha:t >that's gone too far<, =So there is like
risks on all sides here,
=But ↑yeah >[maybe I am] sympathetic to the idea of like< (.)
Stu [<yess>]
Ama Do:nt-n like nee:dlessly like lie to or-or mistreat

Stu >°right°<
Ama =treat (.) <anything and that kind of includes these things
>°even if you think they're #not moral patients#°<.