

Resisting Surveillance AI: Mechanisms of Accountability Evasion and Community Engagement

A Pranav

Trustworthy AI Lab
University of Hamburg, Germany

Motivation

Computer vision methods are increasingly deployed as technosolutionism for ‘safety’ and as security theatre. These systems require massive training datasets, which often rely on low-quality, non-consensual footage of pedestrians in public spaces paired with speculative annotations. Despite these empirical flaws, vendors market these computer vision models using inflated buzzwords to attract venture capital, aggressively pushing them for mass surveillance tasks like person re-identification and biometric tracking. State actors frequently legitimize this tech; in Germany, police operate public cameras in train stations and squares, justifying the surveillance on shoddy performance metrics, while the UK government is currently advancing age-gating legislation that mandates algorithmic identity checks via image uploads. This extended abstract outlines my dissertation, which investigates these cases to examine whether these models actually live up to their technical claims and what their unchecked deployment means for society.

My dissertation asks: *how can technology and society hold each other accountable in the policies they jointly shape?* To answer *how technology informs society*, I evaluate surveillance systems and audit the claims made for them. Through technical audits, I test whether person re-identification and age estimation deliver their claimed capabilities under the conditions in which they are deployed. Through governance audits, I examine how German startups market AI, how age-verification policies specify the systems they mandate, and how benchmarks encode sociodemographic information such as gender and age. To answer *how society informs technology*, I turn to community audits: I document how civic groups contest the surveillance deployments in their districts, and co-develop policies with them. Figure 1 shows an outline of my dissertation which maps the seven audits across the three strands.

Technical Audits

ReID and cross-camera tracking (completed). Public surveillance systems track individuals using three components: a pedestrian detector, a per-camera tracker, and **person re-identification (ReID)**. ReID matches people across

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

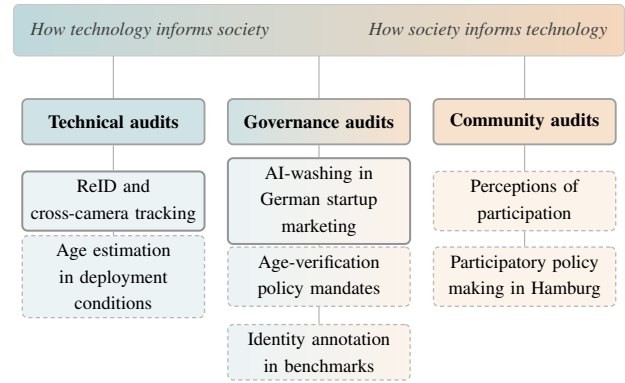


Figure 1: Overview of the audits in my dissertation.

varied contexts, linking images within a single camera feed or across different cameras. Cross-camera tracking relies on ReID to maintain continuous trajectories when individuals leave and re-enter a camera’s view. ReID models are trained on datasets like Market-1501 and MSMT17, constructed from non-consensual pedestrian footage. Developers justify retaining this data by claiming the resulting models enable tracking for policing and retail analytics. I tested whether training on these surveillance datasets actually improves tracking (Pranav and Lauscher 2026). Controlling for all other variables, I fine-tuned various model architectures (from CNNs to transformers) across well-cited ReID datasets and evaluated each inside a pedestrian tracker on three standard benchmarks. **My results show that fine-tuning on surveillance data yielded statistically insignificant tracking improvements, additionally increasing training data provided no improvement.** Instead of accurate tracking, these models merely learn superficial shortcuts and stereotypes, such as clothing colour. From this study, I argue that deploying ReID for mass surveillance is flawed, functioning merely as invasive security theatre.

Age estimation in deployment conditions (planned). Platforms now mandate biometric face scans often paired with uploaded identity documents to gate adult content, which raises questions on privacy and usability. Simultaneously, public surveillance systems estimate age from pedestrian video feeds, and the UK Home Office is trialling these

classifiers on asylum seekers at the border. Research in computer vision has shown that inferring demographics based on physiognomic premises are often flawed, especially on gender classifiers. To evaluate deployment reliability, I will audit the underlying datasets for user consent, the inclusion of minors, and how the demographic labels were originally assigned. I will then test whether these classifiers generalize across datasets or if they exploit visual shortcuts like clothing and backgrounds. Finally, I will determine if these systems are robust enough to maintain stable age predictions for a tracked person in real-world environments.

Governance Audits

AI-washing in German startup marketing (completed). AI-washing is a deceptive practice where companies exaggerate, misrepresent, or obscure the true role of AI in their products. To understand this practice, I annotated the marketing copies of German AI startups on seven dimensions, such as claim specificity and ethics, and interviewed programmers and marketing staff to understand their motivations. The result of my audit showed that 59% of these startups described their AI solely through buzzwords, 43% provided no public evidence, and 79% of numeric claims lacked methodological disclosure. Most importantly, **38% openly marketed applications that the EU AI Act either heavily regulates as high-risk or explicitly prohibits**. This deception is financially motivated; because investors heavily reward AI branding, founders are pressured to drastically overstate their tech's capabilities. This toxic financing loop actively funnels startups into developing invasive surveillance tools, resulting in unverified, oppressive technologies being directly imposed on the public.

Age-verification policy mandates (planned). Building on prior technical evaluations, this project audits the commercial deployment of age estimation systems across digital platforms. Regulatory frameworks are increasingly using facial age estimators for gating the content. I hypothesize that age estimation capabilities are systematically exaggerated, with platforms deploying these tools without independently verifying vendor accuracy reports. I will map a comprehensive audit trail linking age estimation vendors to specific platform deployments. This includes investigating the data pipelines vendors use for training and testing. By compiling a census of vendors, platforms, and policy instruments, I will code every regulatory approval for the actual evidence accepted, evaluating how error liability is structurally deflected across the developer, vendor, and employer supply chain.

Identity operationalization in benchmarks (planned). In this project, I will investigate how the fluid and complex spectrum of human sexuality is codified within language model benchmarks. Models often deploy gender as a discrete variable, whether as a predicted attribute, a feature in downstream decision-making, or a group label for algorithmic evaluation. However, this rigid, categorical encoding systemically marginalizes queer individuals and often gender functions as a superfluous variable within personalization and decision systems. Because benchmarks embed this exclusionary encoding directly into their evaluation

items, and developers explicitly optimize toward these metrics, more models reproduce this erasure at a larger scale. This study audits how current benchmarks operationalize gender and sexuality, critically questioning whether these codified constructs serve any legitimate downstream utility. I will deploy a mixed-methods framework bridging collaborative autoethnography with systematic benchmark auditing. The outcomes of this research would aim to compel AI practitioners and developers toward more ethical, contextually aware, and inclusive paradigms for evaluating language models in the future.

Community Audits

Perceptions of participation (planned). Participatory methods in current ML developments are mostly extractive as systems relegate affected populations to basic data labeling or post-hoc red-teaming, leaving actual design authority exclusively with developers. This study studies this dynamic by locating where the development pipeline can support community input and I will do this via systematic paper annotations with focus groups involving ML practitioners and affected communities. This approach maps the current lifecycle to pinpoint structural barriers and define entry points where communities can shift from passive subjects to active co-designers.

Participatory policy making in Hamburg (planned). Hamburg Police has deployed IVBeo, an automated public CCTV surveillance system, at Hansaplatz and the central station to supposedly detect threats via action classifiers. This expansion of state monitoring lacks technical justification, as audits demonstrate that classifiers operating on mass surveillance datasets fail to yield empirical benefits. While local residents actively oppose this infrastructure, authorities have often rejected their freedom-of-information requests to obscure system efficacy. Through participatory fieldwork with civic groups, I will co-design technical evidence to scientifically contest this deployment and challenge the state's unsupported security claims.

Conclusion

My dissertation shows that surveillance AI does not work under deployment conditions, and that the people marketing it exaggerate and mislead about what it can do. I argue that policy must be designed with the people it affects, especially by centring marginalised communities. This dissertation will co-produce evidence, briefs, and testimony with residents and civic groups, and bring the completed audits into actual governmental policy-making processes.

References

- Pranav, A.; Hartmann, L.; and Lauscher, A. 2026. Auditing AI-Washing in German Startups: A Mixed-Methods Study of Marketing Claims and Perceptions. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*.
- Pranav, A.; and Lauscher, A. 2026. Training on Mass Surveillance Datasets Provides No Empirical Benefits. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*.