# Subversive AI: Resisting automated algorithmic surveillance with human-centered adversarial machine learning

**Sauvik Das**
Georgia Tech
sauvik@gatech.edu | @scyrusk

**Abstract:**

How can we balance the power dynamics of AI in favor of everyday Internet users, particularly those from populations who are disproportionately harmed by automated algorithmic surveillance? I argue that, when advanced from a human-centered perspective, "adversarial" machine learning (AML) can make way for "subversive" AI (SAI). The goal of SAI is to empower end-users with usable obfuscation technologies that protects the content they share online against automated algorithmic surveillance without affecting how that content is consumed by the intended human audience. SAI employs a human-centered design process that spans three-phases of work: (i) modeling lived threats; (ii) exploratory co-design; and, (iii) implementation with human-centered evaluations. I outline a research agenda for Subversive AI to help orient interested researchers and practitioners.

## 1. Introduction:

Today, AI primarily benefits a few powerful institutions—governments, financial institutions and big tech—while its costs are primarily borne by "the people": the masses of individuals subject to ubiquitous, expansive and impersonal surveillance.

Ubiquitous surveillance is everywhere—not just on one's phone, or one's web browser, but in the home, in the car, on the street. Expansive surveillance goes beyond the shallow—not just clickstreams or which websites one visits, but inferences about one's politics, sexuality, driving habits, and one's influence on one's friends. Impersonal surveillance operates at scale—not just carried out by a specific human analyst on a person of interest, but carried out dispassionately on all people at all times.


**Figure 1:** Subversive AI spans a three-phased human-centered design process that integrates the voices of human users that are subject to and concerned about automated algorithmic surveillance.

With state-of-the-art computer vision, for example, it is possible for law enforcement to find the online presence of an anti-police brutality protestor by using facial recognition on photos circulated online [1]. With advanced anomaly detection algorithms, it is possible for banks to construct profiles of what constitutes "normal" purchasing behavior for an individual [2]. With vocal analysis, it is possible for social networking services to infer binary gender and mental state from a voice clip [3-4]. In short, the existing ethos of AI research is to construct an automated algorithmic surveillance infrastructure in the name of enhanced profits, security, and even "social good."

The impact of this ubiquitous, expansive, and impersonal algorithmic surveillance can produce widespread chilling effects that stifle free expression and exacerbate systemic inequities. In the U.S., for example, over 60% of internet users believe their online activity is monitored by the government [5]. Moreover, this surveillance disproportionately affects historically oppressed populations [6]. In China, facial recognition has been used to track religious minorities and protestors [7]. In Russia, NLP has been used to censor "objectionable" online content, targeting the Russian LGBTQ+ community [8]. In the U.S., predictive policing is used to justify excessive law enforcement presence in communities of color [9].

In short, advances in AI, to date, have been instrumental in maintaining existing power structures. But there are alternative trajectories for AI that can subvert, rather than enhance, the power of surveillance capitalists and intelligence agencies.

Consider adversarial machine learning (AML)—a field of research which examines methods to minimally perturb data (e.g., images, audio files) in a manner that is nigh imperceptible to humans, but can thwart state-of-the-art machine learning classifiers [10]. AML could be used to provide people with obfuscation tools

that allow them to share personal data online with reduced fear of algorithmic surveillance: e.g., to share photos of protestors invisible to face detection, but in which humans can see protestors' faces clearly.

Brunton and Nissenbaum call such obfuscation a "weapon for the weak"—used in both nature and throughout human history to subvert predatory beings and powerful institutions [11]. But, that vision is not what drives AML research today. Indeed, an examination of its lexicon is telling. "Adversaries" are parties who dare to challenge AI systems; attacks are methods adversaries use to accomplish their goals; defenses are methods to protect AI systems against adversarial attacks. This lexicon may seem like it has been innocuously borrowed from cybersecurity, but it is important to be clear about its implications: that AI systems—often developed in partnership with or funded by large surveillance institutions—are "good" and require protection from the "bad" people who dare deviate from their learned decision boundaries.

When the salient threat model is not rogue hacker groups but the AI systems and the surveillance institutions who employ them, I argue that AML techniques are not merely adversarial: they have the potential to be subversive. But this transformation from adversarial to subversive requires a human-centered approach in which the needs of the people are considered and checked against the systems that are produced. To that end, I outline a research agenda for Subversive AI (SAI)—a human-centered enhancement of adversarial machine learning to shift the power dynamics of AI away from surveillance institutions and towards the people.

**2. Resarch Agenda for SAI:**

SAI can be thought of as a natural extension of efforts in Human-centered AI (HAI) as applied to adversarial machine learning (AML). Riedl defines HAI as "a perspective on AI and ML that intelligent systems must be designed with awareness that they are part of a larger system consisting of human stake-holders." [12] SAI, then, can be distinguished from research in adversarial machine learning in its application of a human-centered design process to drive the development of people-facing obfuscation tools that are adversarial to algorithmic surveillance.

SAI has three goals: technical, social and ethical. The technical goal of SAI is to create obfuscation filters that people can apply to the content they share online in a way that minimizes the differences in how that content is consumed by intended human audiences but that thwarts algorithmic surveillance in reliable ways. To accomplish this goal, I envision adapting black-box evasion and poisoning attacks in adversarial machine learning (e.g., projected gradient descent [13]) so that they may be applied to use-cases and threat models derived from real end-users.

The social goal of SAI is to empower people, particularly those from communities who disproportionately bear the negative effects of algorithmic surveillance (e.g., LGBTQ+ activists [8] and religious minorities [14]), to use online services to mobilize their communities and share their voices with reduced fear of algorithmic surveillance and content moderation. To accomplish this goal, we must employ a human-centered design process with diverse stakeholder groups and evaluate our designs directly with these stakeholders.

Finally, the ethical goal of SAI is to reduce the power imbalance of machine learning in which large corporate and civil institutions reap the benefits of advances in artificial intelligence at the expense of the privacy of individual end-users. To accomplish this goal, it will be necessary to move beyond theory and contained applications—the focus must be on creating deployable obfuscation tools and conducting robustness evaluations against AI systems actually used to de-identify and profile user-generated content.

*2.1 Threat Model*

The broad threat that subversive AI attempts to address is automated algorithmic surveillance: the computational processing and transformation of large datasets into sensitive outputs, inferences and predictions about individuals, often invisibly and without consent [15-16]. Importantly, the *automated* qualifier refers to surveillance in which there is *no human analyst in-the-loop* who is on the lookout for a specific individual: SAI obfuscation tools are meant to confuse algorithms, but keep content legible for humans. This attribute allows for legitimate law enforcement or national security efforts — those in search of specific persons with a warrant, for example — to continue, even if more slowly.

One concern is that SAI may only be ephemerally effective—obfuscation strategies that work now may be broken later. This is a general concern for all secure systems since security claims are unfalsifiable [17]—i.e., security cannot be guaranteed given future advances. As such, SAI should not be thought of as a panacea for automated algorithmic surveillance. Rather, it is one tool within a broader suite of tools to protect against automated algorithmic surveillance more completely (e.g., end-to-

end encryption, onion routing, VPNs). Nevertheless, even with SAI, individuals must exercise judgement in deciding what level of risk they are willing to tolerate with sharing content online and what protections provide sufficient relief against those risks. At the very least, sharing content with SAI protections should raise the costs of automated algorithmic surveillance, making it less appealing and/or profitable to do en masse.

*2.2 Human-Centered Design Process*

Given a context in which people are concerned about algorithmic surveillance on the content they might share online (e.g., photos of protesters; anonymously authored texts), the goal of SAI is to produce obfuscation tools that allow intended human audiences to consume shared content undeterred while preventing AI from automatically inferring sensitive, identifiable information (e.g., faces, interests, attributes). Accordingly, fundamental to Subversive AI is a human-centered design process that spans three-phases of work: (i) lived threat modeling with relevant stakeholder groups; (ii) exploratory co-design workshops, specifically seeking out and including marginalized voices; and, (iii) implementation with human-centered evaluations.

*2.2.1 Modeling lived threats*

All too often, formal and rigid threat models map poorly onto lived experiences of threat. Since a core goal of subversive AI is to create usable obfuscation technologies that help real people curb real threats of automated algorithmic surveillance, we must first understand the threat as perceived and experienced by affected populations.

To do so, it is imperative to employ mixed-methods formative studies with a diverse range of end-users who must regularly share sensitive, identifiable information (SII) online —e.g., activists who share images of protests, whistleblowers who share anonymous texts, journalists from historically marginalized communities. The specific methods employed can include, for example, semi-structured interviews to obtain broad understanding of threat and context for relevant stakeholders; scenario-based design prompts to funnel participants attention to worst-case hypotheticals; and, diary studies in order to obtain in situ information about perceived threats.

The analysis goals of this phase of research are three-fold: (i) to typify the algorithmic surveillance threat models varied stakeholders harbor in sharing SI content; (ii) to describe behavioral responses to these threat models; and, (iii) to highlight how various aspects of identity and content intersect with threat models and behavioral responses.

*2.2.2 Participatory co-design workshops*

The next phase of research should include exploratory co-design workshops, using the results from the formative studies to scaffold the explorations. Co-design, which has its roots in Scandinavian participatory design, is the act of designing with stakeholders by engaging them in a dialog around the types and form of the desired solution [18]. Co-design can help ensure that an implemented solution closely matches what stakeholders need.

Using the lived threat models uncovered in the first phase of research to guide design explorations, researchers should construct "seed" scenarios describing a character from a stakeholder group, the SII they would like to share, and the threats about which they are concerned. Participant groups should, in turn, collectively ideate design constraints and obfuscation solutions. Note that co- and participatory design works when relevant stakeholders are present in the design process; thus, a key gooal of these co-design workshops should be recruiting diverse and marginalized voices.

Researchers should then use their domain expertise to convert the generated ideas and the discussion around those ideaas to produce prescriptive design recommendations for SAI filters and tools.

*2.2.3 Implementation & Evaluation*

Only after needs-finding in phase I and co-design in phase II should researchers begin implementation of a SAI tool, starting from techniques in AML. While a full review of AML techniques is out-of-scope for this short treatise, adversarial techniques generate or perturb inputs to ML systems that force those systems to produce erroneous outputs [10,13].

There are three types of AML attacks [10,19]: evasion attacks, in which an attacker perturbs inputs to confuse trained models in a manner that is nearly imperceptible to humans (e.g., changing an image that a person would recognize as a bird to be misclassified as something else); poisoning attacks, in which an attacker contaminates the data on which a model is trained; and, exploratory attacks, in which an attacker attempts to reconstruct an unknown model. Within evasion attacks, there are white-box attacks, in which an attacker knows the underlying model being used, and black-box attacks, in which an attacker does not know the underlying model being used [19].

The broad approach I envision for SAI is an evasion attack. Black-box evasion attacks can sometimes

be effective because of adversarial sample transferability [20]: i.e., the observation that adversarial samples produced to work against a simple model can work against more sophisticated models for the same task. This allows attackers to train a local substitution model that approximates the black-box model and create adversarial samples that work against this model using any white-box attack (e.g., projected gradient descent [13]). Moreover, model inversion attacks can approximate production models with carefully constructed inputs [21], making it possible to create local substitution models that are effectively white boxes. While many defensive techniques have been proposed to make models more robust (e.g., distillation, compression, adversarial training) [10,13,19], none work on all classes of attacks [10].

Finally, it is not enough to implement a subversive defense and test only its technical effectiveness in toy use cases on existing datasets. The developed models must be evaluated against the needs of the human users who motivated its development.

The gold standard would be to implement a working prototype that these human users could test in a field deployment—but care must be taken to first ensure the robustness of the SAI system through lab evaluations. It is of critical importance to recruit a diverse range of users in these initial evaluations, particularly those who are most affected by the identified threats. A mixed methods approach would be ideal: using qualitative methodologies to extract rich experiential data, and quantative methodologies on data collected telemetrically, with consent.

## 3. Conclusion:

Today, there are significant asymmetries in who benefits from advances in AI and who bears the costs. Scholars in science and technology studies (STS) have been steadfast in their critiques of AI for that reason. For example, Zeynep Tufecki argued: "We're building a dystopia just to get people to click on ads." [22] Shoshanna Zuboff argued that advances in AI have given way to a new era of "surveillance capitalism" in which the prime invective of powerful data institutions is to collect and exploit a surplus of personal, behavioral data [16]. Beyond these top-level concerns, scholars like Simone Brown [6] have demonstrated how the negative effects of surveillance disproportionately affect people of color, especially Black people.

But AI need not have only one story. To that end, I have outlined a research agenda for Subversive AI: a human-centered enhancement of adversarial machine learning to shift power away from surveillance institutions and towards the people. It will not be easy—the powers that be have more resources, a significant head start, and can veil their problematic advances under the guise of "social good" [23]. Yet, we must resist.

## 4. Acknowledgments:

My Ph.D. students, Jacob Logas and Youngwook Do, provided helpful comments and feedback for this paper.

## 5. References:

[1] J Wilkin. 2018. Mapping Social Media with Facial Recognition: A New Tool for Penetration Testers and Red Teamers. TrustWave. Retrieved May 8, 2020

[2] Fu, K., Cheng, D., Tu, Y., & Zhang, L. 2016. Credit card fraud detection using convolutional neural networks. In International Conference on Neural Information Processing (pp. 483-490). Springer, Cham.

[3] Meysam Asgari, Izhak Shafran, and Lisa B Sheeber. 2014. Inferring clinical depression from speech and spoken utterances. 2014 IEEE international workshop on Machine Learning for Signal Processing (MLSP).

[4] Robert M Krauss, Robin Freyberg, and Ezequiel Morsella. 2002. Inferring speakers' physical attributes from their voices. Journal of Experimental Social Psychology 38, 6: 618–625.

[5] Brooke Auxier, Lee Rainie, Monica Anderson, Andrew Perrin, Madhu Kumar, and Erica Turner. 2019. Americans and privacy: Concerned, confused and feeling lack of control over their personal information. Pew Research Center: Internet, Science & Tech (blog). November 15: 2019.

[6] Simone Browne. 2015. Dark matters: On the surveillance of blackness. Duke University Press.

[7] Bei Qin, David Strömberg, and Yanhui Wu. 2017. Why does China allow freer social media? Protests versus surveillance and propaganda. Journal of Economic Perspectives 31, 1: 117–140.

[8] Kyle Knight. 2019. Russia Censors LGBT Online Groups. Human Rights Watch. Retrieved May 8, 2020 from https://www.hrw.org/news/2019/10/08/russia-censors-lgbt-online-groups

[9] Sheehey, B. 2019. Algorithmic paranoia: the temporal governmentality of predictive policing. Ethics and Information Technology, 21(1), 49-58.

[10] Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, Aleksander Madry and Alexey Kurakin. 2019. On evaluating adversarial robustness. arXiv preprint arXiv:1902.06705.

[11] Finn Brunton and Helen Nissenbaum. 2015. Obfuscation: A user's guide for privacy and protest. Mit Press.

[12] Riedl, M. O. 2019. Human-centered artificial intelligence and machine learning. Human Behavior and Emerging Technologies, 1(1), 33-36.

[13] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2017. Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083.

[14] Isobel Cockerell. 2019. Inside China's Massive Surveillance Operation. URL: https://www. wired. com/ story/inside-chinas-massive-surveillance-operation/(25 September 2019).

[15] Maria Helen Murphy. 2017. Algorithmic surveillance: the collection conundrum. International Review of Law, Computers & Technology 31, 2: 225–242.

[16] Shoshana Zuboff. 2019. The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power: Barack Obama's Books of 2019.

[17] Cormac Herley. 2016. Unfalsifiability of security claims. Proceedings of the National Academy of Sciences 113, 23: 6415–6420. http://doi.org/10.1073/pnas.1517797113

[18] Bo Edvardsson, Anders Gustafsson, and Inger Roos. 2005. Service portraits in service research: a critical review. International journal of service industry management.

[19] Anirban Chakraborty, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, and Debdeep Mukhopadhyay. 2018. Adversarial attacks and defences: A survey. arXiv preprint arXiv:1810.00069.

[20] Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. 2016. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. arXiv preprint arXiv:1605.07277.

[21] Florian Tramèr, Fan Zhang, Ari Juels, Michael K Reiter, and Thomas Ristenpart. 2016. Stealing machine learning models via prediction apis. 25th USENIX Security Symposium (SEC'16) 601–618.

[22] Tufekci, Z. (2017). We're building a dystopia just to make people click on ads. TED: Ideas Worth Spreading,

[23] Green, B. (2019). "Good" isn't good enough. In Proceedings of the AI for Social Good workshop at NeurIPS.