POTs: the revolution will not be optimized?

Seda Gürses, Rebekah Overdorf, and Ero Balsa
COSIC, KU Leuven

1. THE OPTIMIZATION PROBLEM

In the 90s, software engineering shifted from packaged software and PCs to services and clouds, enabling distributed architectures that incorporate real-time feedback from users. In the process, digital systems became layers of technologies metricized under the authority of objective functions that drive selection of software features, service integration, cloud usage, user interactions, user growth, and environmental capture. Whereas previous information systems focused on storage, processing and transport of information, organizing knowledge, and making it accessible —with associated risks of *surveillance*— contemporary systems leverage the knowledge they gather to not only understand the world, but also to *optimize* it, seeking maximum extraction of economic value through the capture and manipulation of people's activities and environments.

The ability of these *optimization systems* to treat the world not as a static place to be known, but as one to sense and co-create, poses social risks and harms such as social sorting, mass manipulation, asymmetrical concentration of resources, majority dominance, and minority erasure. In the vocabulary of optimization, these harms arise due to choosing inadequate *objective functions*, that, among other things, 1) aspire for antisocial or negative environmental outcomes[1], 2) have adverse side effects[2], 3) be built to only benefit a subset of users[2], 4) externalize risks associated with environmental unknowns and exploration to users and their surroundings[3], 5) be sensitive to distributional shift, wherein a system that is built on data from a particular area or domain is deployed in another environment that it is not optimized for [4], 6) spawn systems that exploit states that can lead to fulfillment of the objective function short of fulfilling the intended effect [5], and 7) distribute all of the errors to a specific group of users[6] [2]. Common to both surveillance and optimization systems is the concentration of data and processing power that results in overwhelming political economic leverage enabled by scalability, network effects, and externalizing of risks to populations and environments.

To better illustrate the difference between information and optimization systems and the problems the latter pose, in the rest of the paper we focus, without loss of generality, on location based services (LBS). LBS have moved beyond tracking and profiling individuals to generate spatial intelligence to leveraging this information to manipulate users' behavior and create "ideal" geographies that optimize space and time to customers or investors interests [3]. Population experiments drive iterative designs that ensure sufficient gain for a percentage of users while minimizing costs and maximizing profits.

LBS like Waze provide optimal driving routes that put users in certain locations at a disadvantage. Waze often redirects users off of major highways through quiet suburban neighborhoods not accustomed to heavy traffic. While useful for drivers, it affects quiet neighborhood dwellers by making their streets busy, noisy and less safe. It also affects the towns, that consequently need to fix and police the roads more often. This further shows that even when most users benefit, *non-users* may bear the ill effects of optimization. Users within a system may also be at a disadvantage due to their location. Pokémon Go users living in urban areas see more Pokémon, more *Pokéstops* (to collect resources) and more *gyms* (to beat Pokémon) than users in rural areas. Uber manipulates prices in space and time, constituting geographies around supply and demand that both drivers and riders are unable to control, negatively impacted by price falls and *surges*, respectively. Recent studies report that Uber drivers (who work on commission, sharing a part of the revenue from every ride they complete with the company) make less than the minimum wage in many jurisdictions.

Disadvantaged users have worked from within the system to tame optimization in their favor, e.g., by strategically feeding misinformation to the system in order to change its response or behavior. Quiet neighborhood dwellers negatively affected by Waze's traffic redirection have fought back by reporting road closures and heavy traffic on their streets —to have Waze redirect users out of their neighborhoods. Some Pokémon users in rural areas spoof their locations to urban areas. While explicitly against Pokémon Go's rules, many evade detection. Other users report to OpenStreetMaps —used by Pokémon Go— formerly unreported or false footpaths, swimming pools and parks, resulting in higher rates of Pokémon spawn in their vicinity. Uber drivers have colluded to induce *surge* pricing and temporarily increase their revenue by simultaneously turning off their apps, making the system believe that there are more passengers than drivers, then turning the app back on to take advantage of the surge pricing in the area.

While the long-term effectiveness of these techniques is unclear, they inspire the type of responses that a more principled approach may provide. In fact, these responses essentially constitute adversarial machine learning, seeking to bias system responses in favor of the "adversary". The idea of turning adversarial machine learning around to attack a system for the benefit of the user is already prevalent in the literature around PETs (e.g.[7][8]). It is in fact in the spirit

[1] https://www.theatlantic.com/technology/archive/2018/03/mapping-apps-and-the-price-of-anarchy/555551/

[2] www.latimes.com/local/california/la-me-lopez-echo-park-traffic-20180404-story.html

[3] See [1], although we disagree with this paper's premise that optimization systems will lead to 'optimal' outcomes, with experimentation as its only potential externality.

[4] www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[5] https://nymag.com/selectall/2018/04/malcolm-harris-on-glitch-capitalism-and-ai-logic.html

[6] https://medium.com/@mrtz/how-big-data-is-unfair-9aa544d739de

[7] PETS'12: McDonald et al. "Use fewer instances of the letter "i": Toward writing style anonymization."

[8] PETS'17: Giovanni Cherubin et al. "Website Fingerprinting Defenses at the Application Layer."

of PETs that we attend to the optimization problem, i.e., we explore ideas for technologies that enable people to recognize and respond to the negative affects of optimization systems.

2. POTS

Optimization systems infer, induce and shape events in the real world to fulfill objective functions. *Protective optimization technologies* (POTs) reconfigure these events as a response to the effects of optimization on a group of users or local environment. POTs analyze how events (or lack thereof) affect users and environments, then manipulate these events to influence system outcomes, e.g., by altering the optimization constraints and poisoning system inputs.

To design a POT, we first need to understand the optimization system. What are its user and environmental inputs (U, E) and how do they affect the capture of events? Which are the outcomes $O = F(U, E)$ that may bring undesirable outcomes for subpopulations or environments? Once we have a characterization of the system, as given by $F(U, E)$, we identify those who benefit from the system and those placed at a disadvantage. Our first intuition is to define disadvantage as those people and environments that reside in local minima over the benefit distribution function. Then, we construct an *alternative* benefit function, $B(X, E', Value)$: $(x, e) \rightarrow value$ that includes both users and non users ($U \subset X$) and an environment $E' \subseteq E$.

A POT benefit function B may attend to different goals. B may attempt to *"correct" imbalances* optimization systems create, i.e., by improving a systems' outcomes for populations put at disadvantage. In this case, B ideally balances the system for those at local minima without creating new local minima. Conversely, B may also strategically attempt to reverse system outcomes as a form of *protest*, highlighting the inequalities these systems engender. This further hints at the subversive potential of POTs. POT designers may concoct a B to *contest the authority* of optimization systems, challenging the underlying objective functions these systems optimize to and their very *raison d'être*. To do that, B may attempt to sabotage or boycott the system, either for everyone or for a impactful minority that are more likely to affect change, leveraging the power asymmetries B precisely intends to erode.

Once we select B, we must choose the techniques that implement it. These techniques involve changes to the inputs that users have control over and alterations to constraints over the objective function to reconfigure event capture (i.e., the system's mechanism of detection, prediction, and response to events). Lastly, we deploy and assess the impact of the POT both in terms of local and global effects on users and environments as intended by B and tweak it as necessary.

We note that POTs may elicit a counterresponse from the optimization systems they target. The latter may either attempt to neutralize POTs or expel those deploying them from the system. Anticipating these responses may require POT designers to aim for *undetectability*, e.g., by identifying minimum alterations to inputs and constraints, or optimizing constraints to prevent detection.

3. DISCUSSION

As with PETs, POTs come with moral dilemmas. Some of these are comparable to concerns raised with respect to the use of obfuscation in PETs, although the latter focuses on the protection of individual profiles (privacy) and not the protection of populations and environments (from optimization). In their seminal work on obfuscation, Brunton and Nissenbaum highlight three ethical issues: dishonesty, wasted resources and polluted databases [4]. We evaluate these in the context of POTs.

Since optimization systems are not about knowledge, one could argue using POTs cannot be judged as dishonesty but as inserting unsolicited feedback into the cybernetic loop to get optimization systems to recognize and respond to their externalities. POTs are likely to come at a greater cost to the service providers, andmay give rise to negative externalities that simply impact different subpopulations and environments. In fact, all of the issues that we discussed as harmful effects of optimization systems can be replicated in POTs: they may have an antisocial objective function, have serious side effects, benefit a few and so on. Seen that way, it is possible to argue that if optimization is the problem, then more optimization is not likely to solve the problem and may even come to exacerbate it.

Nevertheless, one could make arguments for POTs. First, optimization history is also one of counter-optimization as evident in the case of search engine optimization or spammers. POTs can be built to ensure that counter-optimization is not only available to a privileged few. The many stories in the news about people applying these techniques to push back on the universal ambitions of optimization systems demonstrate that it provides agency. Ensuring that these acts of protest or unsolicited feedback are more than just inspirational actions will require well designed, well evaluated and stealthy POTs, a topic of research that we think the PETs community is well suited to embark upon.

Finally, an ethical argument can be made for POTs similar to that of obfuscation. In [4], the authors argue that, given the negative outcome of surveillance capitalism, these arguments can be boiled down to (a) the aims of an obfuscation tool being laudable and (b) no alternative path to change with lesser costs [to society] existing. In the LBS examples above, companies accelerate the way in which space gets negotiated in a way invisible to the inhabitants of the effected environments [3]. If so, while short of a revolution, there is a strong argument to be made for POTs that aspire to reintroduce those inhabitants into the negotiations of how their environments are organized.

4. REFERENCES

[1] S. Bird, S. Barocas, K. Crawford, F. Diaz, and H. Wallach, "Exploring or exploiting? social and ethical implications of autonomous experimentation in ai. ssrn scholarly paper id 2846909," *Social Science Research Network, Rochester, NY.*, vol. 2846909, 2016.

[2] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, "Concrete problems in AI safety," *arXiv preprint arXiv:1606.06565*, 2016.

[3] D. Phillips, M. Curry, and D. Lyon, "Privacy and the phenetic urge," *Surveillance as social sorting: Privacy, risk and digital discrimination*, pp. 137–152, 2003.

[4] F. Brunton and H. Nissenbaum, *Obfuscation: A user's guide for privacy and protest.* Mit Press, 2015.