Retrospective Alignment

Andreas Haupt
h4upt@stanford.edu
Stanford University
Palo Alto, CA, USA

ABSTRACT

This position paper argues for retrospective alignment, the property of systems to not elicit behaviors from humans that those regret after the fact. We argue for regret data as a technical means of its implementation and a Right to Behavior Log Rectification as a legal mechanism to achieve it. We discuss two alternative viewpoints.

CCS CONCEPTS

• **Social and professional topics** → **Consumer products policy**; • **Applied computing** → *Law*; • **Human-centered computing** → **Social engineering (social sciences)**.

KEYWORDS

alignment, regret, retrospective alignment, buyer's remorse, cool-off period.

Human decisions are increasingly mediated by Artificial Intelligence that, in turn, maximizes objectives based on observations of human behavior. In optimizing AI decisions, several points of misalignment arise. In this paper, we argue that retrospective alignment, the property of systems to not elicit behavior from humans that those later regret, is desirable, and that regrets data is a technical means and a right to modification of behavior logs is a legal avenue for achieving such alignment.

1 RETROSPECTIVE ALIGNMENT

> In a retrospectively aligned system all humans, retrospectively, endorse their past behavior.

Ensuring retrospective alignment is normatively desirable because it helps individuals make choices they will stand by in the long run, thus realizing positive freedom in their decision-making, Kwok [8]. For welfarists, it increases "true" welfare in a system, [2, 3]. It also ensures long-term trust and engagement in systems, as retrospective misalignment will lower trust and lead to humans leaving a system, [5].

Human societies give a plethora of examples of retrospective misalignment. Buyer's remorse after high-pressure door-to-door sales and telemarketing, regrets of extensive fine-print for privacy rules, cold-chicken social media users after excessive use, and failure to cancel unneeded subscription because of dark patterns are all examples where humans in a system regret actions—signing a sales contract, for example—they took in the past, [1, 7]. [1] We contend that such cases of regret-inducing environments cause harm in the humans using a system and hence is undesirable.

Retrospective alignment differs from other alignment paradigms in that it is human-centered. Fundamentally, it is not about representing human preferences [9], but about the behavior that a user shows inside of a system. Retrospective alignment is fundamentally sociotechnical. Retrospective alignment does not directly demand that an AI system ought not exploit or manipulate a user, but rather a requirement on the consistency of users interacting with the system. Retrospective alignment will instrumentally lead to systems that do not exploit and manipulate users, as manipulated users will regret their actions if they have been manipulated.

2 REGRETS DATA

Despite the prevalence of regret in many systems with a machine learning component, there has been little treatment of retrospective alignment in AI. We believe that the main reason for this lack is that evaluating harm in AI-based systems in which goods are given for "free", is harder than in commerce settings. If a buyer feels remorse about a recent purchase, allowing returns of a good both are a reasonable remedy that undoes harm that the user experiences through their unconscionable acquisition. Many digital platforms monetize engagement from third parties rather than direct transactions, making it harder to quantify the harm users experience when they later regret their behaviors.

For training AI-based systems, as opposed to remedies in commerce, such harm analysis is unnecessary. It is sufficient to change how systems are trained to achieve retrospective alignment. Consider, for example, the tool [6] (schematically depicted in Figure 1), a survey tool for collecting regret. Based on a data export of youtube.com and Youtube's Application Programming Interface, the tool allows a user to revisit the decision to watch a particular video and allows the user to express regret or approval for the past consumption. It allows the user to review whole sessions, and shows the data to users.

In principle, tools such as [6] are possible for many user-facing systems that allow for histories of actions a human took (which is many, as such a data export is mandated, *e.g.*, under the Right to Data Portability, Art. 20 of the General Data Protection Regulation), and have APIs that are rich enough to accurately reconstruct a past choice for a user. The only limiting point is how such regret data will be used to improve recommendation or machine learning based systems.

It is worth spelling out why regret data helps retrospective alignment. Consider a recommendation system setting, and the possibly simplest way to interpret regret information: stated regret nullifies a record of consumption, leading to different training data for an algorithm. Given such amended training data, the recommendation algorithm is trained to predict content that a user will consume

[1] That is not to say that retrospective feedback is always achievable or meaningful: Asking for retrospective alignment of a game of *roulette* is nonsensical, as losers will regret having spent their money, and randomness is core to the game. Similarly, irreversible changes to preferences may lead to inconsistent statements of regret, compare [4]. We will focus on cases where statements of regret are meaningful.

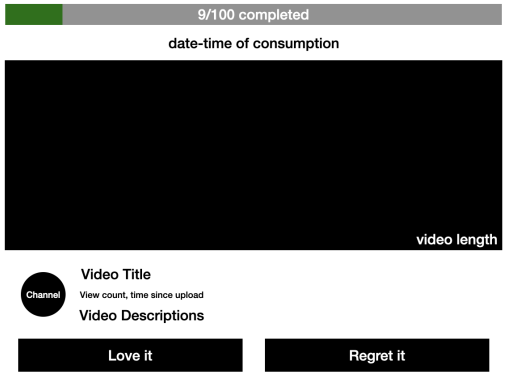

Figure 1: Schematic Representation of the Regret Data collection tool [6].

and not regret, leading to both instantaneous (in the sense of eliciting behavior that proxies for desirable behavior) and retrospective alignment.

Regret data is, however, only useful if incorporated into platforms. Next, we present a legal mechanism incentivizing elicitation of regret data.

3 THE RIGHT TO BEHAVIOR LOG RECTIFICATION

> A user of an online service shall have the right to "rectify" their past behavioral logs.

In this proposal, rectify is in quotation marks as subjective estimates such as regret are, by definition, neither correct nor incorrect. We use the word rectify to show a parallel with the Right to Rectification, Art. 16 of the General Data Protection Regulation. A behavior log is whatever a system stores about user behavior, often consumption or engagement-related behavior.

The Right to Behavior Log Rectification also relates to questions about data subject rights to their algorithmic predictions [10]. The requirement of allowing for behavior log rectifications is strictly lighter than allowing users to change inferences about them (i.e., their full user model), but can powerfully shape what their inferences are, given that so much of user-facing products personalize based on behavior logs [11].

With more and more AI agents deployed, rectifying agents' human behavior logs will become increasingly important. Not only are the stakes higher with AI agents (and likely monetary), but potentially long-term and intimate connections with AI assistants lead to risks of manipulation of users. We, therefore, believe that a right to behavior log rectification, even for commercial AI assistants, helps mitigate some of the most problematic consequences of retrospective misalignment.

4 ALTERNATIVE VIEWPOINTS

We discuss two alternative viewpoints: the comparative effectiveness of other policy instruments and the frivolous use of behavior log rectifications.

4.1 Other Mechanisms Would Be More Effective than Behavior Log Rectification

Many existing regulations implicitly address retrospective misalignment by mitigating regret-inducing decisions through choice frictions, commitment devices, return policies, and price interventions. These policies recognize that individuals often make choices under pressure, with incomplete information, or due to cognitive biases, leading to regret. While current regulations focus on specific domains such as gambling, consumer protection, and addictive goods, they share a common goal: ensuring that individuals have opportunities to make decisions they will stand by over time.

We argue that these policies all benefit from regret data whenever they are applied to machine learning-based settings.

4.1.1 Choice Frictions and Interface Regulations. Regulators have long imposed frictions to curb impulsive decision-making, particularly in high-risk areas like gambling. The UK Gambling Commission mandates cooling-off periods and self-imposed deposit limits, while the EU's Digital Services Act (Regulation (EU) 2022/2065) restricts manipulative design patterns. The U.S. Federal Trade Commission (FTC) also enforces anti-dark-pattern regulations, requiring platforms to simplify opt-outs and disclosures. These measures reduce the likelihood of individuals engaging in behaviors they will regret by giving them structured opportunities to pause and reconsider. We believe that such choice frictions, *e.g.*, timers in reviewing AI Agent behavior, will be helpful and that, ultimately, regret data should be used to evaluate whether interface designs lead to retrospective (mis)alignment.

4.1.2 Commitment Devices. Commitment mechanisms help individuals align short-term actions with long-term goals. Apple's Screen Time allows users to set app usage limits, and gambling self-exclusion programs like the UK's GamStop prevent individuals from accessing betting platforms for predetermined periods. Smoking cessation programs leverage financial incentives and structured plans to help individuals overcome addiction. By providing pre-commitment tools, these interventions empower users to act according to their deeper preferences rather than succumbing to momentary impulses. Importantly, such commitment devices are voluntary affordances to users. Such devices are also helpful for AI assistants: For example, a user could configure an AI agent to not be available between midnight and 6 AM so as not to be tempted to use it. As interfaces, the option of commitment devices can be evaluated using regret data.

4.1.3 Return Policies. Consumer protection laws recognize that individuals often regret purchases made under pressure or without full consideration. The Federal Trade Commission's Cooling-Off Rule (16 CFR Part 429), for example, grants U.S. consumers a three-day period to cancel certain in-person sales, while the EU's Consumer Rights Directive (Directive 2011/83/EU) provides a 14-day no-questions-asked return window for online and off-premises

purchases. These policies allow consumers to reverse decisions they later regret, offering a safeguard against pressure-driven or ill-informed transactions. Returns are, in the case where a purchase happened, a particular form of regret data, and can be fed into a machine learning-based system to minimize the likelihood of such (costly) events.

4.2 Retrospective Alignment Lead to Manipulation by Users

One of the main concerns in achieving retrospective alignment of human systems via *return rights*, for example, the Federal Trade Commission's Cooling-Off Rule and the EU's 14-day return window, is that consumers could abuse their right of return by, *e.g.*, using a bought product for an occasion and sending it back, hence showing strategic behavior as opposed to an acknowledgment of a mistake. While in systems in which transactions happen (also those that AI Agents may soon intermediate) are important, in many systems with a machine learning component, no such *frivolous behavior log manipulations* are plausible.[2]

5 NEXT STEPS

Retrospective alignment is both desirable and achievable with regret data and a Right to Behavior Log Rectification. A promising next step in advocating for retrospective alignment is tooling, that is, the construction of regret data collection tools for more modalities (search and browsing, chatbot interactions, and music listening, among others).

REFERENCES

[1] Amanda Agan, Diag Davenport, Jens Ludwig, and Sendhil Mullainathan. 2023. *Automating automaticity: How the context of human choice affects the extent of algorithmic bias.* Technical Report. National Bureau of Economic Research. https://doi.org/10.3386/w30981

[2] B. Douglas Bernheim and Antonio Rangel. 2009. Beyond revealed preference: Choice-theoretic foundations for behavioral welfare economics. *The Quarterly Journal of Economics* 124, 1 (Feb. 2009), 51–104. https://doi.org/10.1162/qjec.2009.124.1.51 tex.eprint: https://academic.oup.com/qje/article-pdf/124/1/51/5340707/124-1-51.pdf.

[3] B Douglas Bernheim and Antonio Rangel. 2024. Beyond Revealed Preference: Choice-Theoretic Foundations for Behavioral Welfare Economics. (2024).

[4] Micah Carroll, Dylan Hadfield-Menell, Stuart Russell, and Anca Dragan. 2021. Estimating and penalizing preference shift in recommender systems. In *Proceedings of the 15th ACM conference on recommender systems (RecSys '21)*. Association for Computing Machinery, New York, NY, USA, 661–667. https://doi.org/10.1145/3460231.3478849 Number of pages: 7 Place: Amsterdam, Netherlands.

[5] Sarah H. Cen, Andrew Ilyas, and Aleksander Madry. 2024. User strategization and trustworthy algorithms. In *Proceedings of the 25th ACM conference on economics and computation (Ec '24)*. Association for Computing Machinery, New York, NY, USA, 202. https://doi.org/10.1145/3670865.3673545 Number of pages: 1 Place: New Haven, CT, USA.

[6] Andreas Haupt and Mihaela Curmei. 2024. Regret data collection tool. https://doi.org/10.5281/zenodo.13770054

[7] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2023. The Challenge of Understanding What Users Want: Inconsistent Preferences and Engagement Optimization. http://arxiv.org/abs/2202.11776 arXiv:2202.11776 [cs].

[8] Kelvin Hiu Fai Kwok. 2025. An Autonomy Theory of Consumer Protection Law. *SSRN Electronic Journal* (2025). https://doi.org/10.2139/ssrn.5109269

[9] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh conference on neural information processing systems*. https://openreview.net/forum?id=HPuSIXJaa9

[10] Sandra Wachter and Brent Mittelstadt. 2019. A Right to Reasonable Inferences. *Columbia Business Law Review* (May 2019), 494–620 Pages. https://doi.org/10.7916/CBLR.V2019I2.3424 Artwork Size: 494–620 Pages Publisher: Columbia Business Law Review.

[11] Shoshana Zuboff. 2019. *The age of surveillance capitalism: the fight for a human future at the new frontier of power* (first edition ed.). PublicAffairs, New York.

[2] That said, there must be limitations to how often a user can change their mind on behavior, *e.g.*, through a rate limit for how often feedback can be changed. We believe this not to be a big challenge as it is similar to the implementation of the Right to Data Portability, Art. 20 of the General Data Protection Regulation.