Classification of public feedback system to know problem faced by common group of people research paper

Mohammed Bilal
Computer Science and Engineering
Alva's Institute Of Engineering
Mangalore, India
Nmmohdblal5664@gmail.com

Mohammed Fayiz
Computer Science and Engineering
Alva's Institute Of Engineering
Mangalore, India
mohammedfahiz07@gmail.com

Omkar Panchakattimath
Computer Science and Engineering
Alvas's Institute Of Engineering
Mangalore, India
omkarpanchakattimath1@gmail.com

Pavan Kumar V
Computer Science and Engineering
Alva's Institute Of Engineering
Mangalore, India
Tanupavan888@gmail.com

Abstract **Forecast based dynamic frameworks are turning out to be progressively predominant in different spaces. Past investigations have exhibited that such frameworks are powerless against out of control input circles, e.g., when police are more than once sent back to similar areas no matter what the real pace of crime, which worsen existing predispositions. Practically speaking, the computerized choices have dynamic input impacts on the actual framework that can sustain over the long haul, making it hard for foolhardy plan decisions to control the framework's development. While specialists began proposing longer-term answers for forestall unfriendly results (like predisposition towards specific gatherings), these mediations generally rely upon impromptu demonstrating presumptions and a thorough hypothetical comprehension of the criticism elements in ML-based dynamic frameworks is right now absent. In this paper, we utilize the language of dynamical frameworks hypothesis, a part of applied science that arrangements with the examination of the interconnection of situation with dynamic ways of behaving, to thoroughly characterize the various sorts of criticism circles in the ML- based dynamic pipeline. By checking on existing insightful work, we show that this order covers numerous models examined in the algorithmic decency local area, subsequently giving a binding together and principled structure to concentrate on criticism circles. By subjective investigation, and through a reproduction illustration of recommender frameworks, we show which explicit sorts of ML predispositions are impacted by each kind of input circle. We find that the presence of criticism circles in the ML-based dynamic pipeline can propagate, support, or even lessen ML predispositions.**

I. INTRODUCTION

Many of today's automated processes depend on machine learning (ML) calculations to illuminate choices that significantly affect individuals' lives. For example, they are utilized to assess whether an individual ought to be owned up to a specific school [42], be conceded a credit [25], or treated as high gamble of recidivism [6]. The benefit of these ML-based dynamic frameworks is their versatility, i.e., the capacity to deal with countless choices in an effective way. Notwithstanding, specialists have found proof that these calculations frequently intensify existing predispositions that underlie human choices [18, 27, 41] and even present new ones [1, 8, 16, 62]. To take care of this issue, a new line of examination in algorithmic decency began exploring arrangements that can moderate these predispositions by implementing a few measurements of individual or gathering reasonableness [4, 9, 15, 30, 50]. Albeit these endeavors end up finding success temporarily, they frequently don't perform similarly well in the long haul, i.e., after numerous rounds of the dynamic cycle [46, 65].1 The basic explanation appears to lie in the way that the moderating arrangements are intended for fixed frameworks [13, 52], while the actual framework progressively advances over the long haul. All the more explicitly, the framework changes over the long haul on the grounds that the result (the choice) takes care of back as contribution to the actual framework, subsequently making what specialists allude to as a "input circle". The outcome is that inclinations are propagated (or even built up) because of the presence of the criticism circle, notwithstanding implementing the moderation procedures. Despite the fact that scientists as of late begun concentrating on the drawn out impacts of successive dynamic calculations (e.g., [17, 35, 47], see [75] for a new overview), the proposed reproduction put together arrangements are attracted with respect to impromptu models which forestall a correlation of their fundamental suppositions and a profound understanding of the driving elements, i.e., what causes the criticism circles and what parts of the framework are involved. Subsequently, until now, we miss the mark on exhaustive grouping and hypothetical comprehension of these criticism circles, and how they connect with the intensification of various kinds of inclination.

This is an important initial step to significantly impact the examination point of view from growing shallow arrangements pointed toward recognizing and fixing existing predispositions to an all the more long haul situated view that endeavors to expect and forestall inclinations. Contrasted with past work in the field, in this paper, we don't endeavor to give a recreation based answer for the presence of criticism circles, rather we fill this hypothetical hole by giving a proper definition and a thorough scientific categorization of input circles in the ML-based dynamic pipeline, and by connecting them to the predispositions they influence. To do as such, we initially explain the distinction between open-circle and shut circle (or input circle) frameworks by acquiring the language and the instruments from dynamical frameworks hypothesis, the discipline that spotlights on the examination of frameworks with dynamical way of behaving (and their interconnection). Then, we apply this framework hypothetical structure to the dynamic pipeline, which is made out of various sub-frameworks: the people's examining cycle, the people's qualities addressing the choice applicable build, the noticed elements and results, the ML model, and a ultimate conclusion. A ultimate choice can input into any of these sub-frameworks, in this manner shaping various sorts of criticism circles. This, thusly, implies that a definitive impact in general pipeline and the enhancement of predispositions rely upon what sorts of criticism circles are at the same time present in the framework. The first contribution of this paper (see Sec. 2) is to projected the ML-based dynamic pipeline into a framework hypothetical system that underscores the various parts. Our subsequent commitment (see Sec. 3) comprises in giving an order of the various kinds of criticism circles, which we call testing, individual, element, result, and ML model input circle contingent upon what part is impacted. Moreover, we present the thought of "ill-disposed criticism circles," which address unique instances of input circles in which a ultimate choice feeds once again into the framework as a result of some essential activity of the impacted individual(s). As a third commitment (see Sec. 4), we give an outline of the various sorts of predisposition that can be impacted by every one of the five input circles we present. As a fourth and last commitment (see Sec. 5), we show the capability of our grouping structure with regards to news recommender frameworks. In particular, we show that various kinds of criticism circles can influence unmistakable pieces of the dynamic pipeline, bringing about framework elements that produce different types of predisposition.

II. LITERATURE SURVEY

A. Viability of Criticism in Further developing Understudies Learning and Professionalizing Instructing

Creator Md Mamoon-Al-Bashir, AHEA

Year: 2016

There is an extraordinary significance of criticism in further developing growth opportunity for the under studies. This has likewise tremendous impact in professionalizing showing in the advanced education level.Not withstanding, criticism is considered as a troublesome issue in this field. A large portion of the teachers are as yet going on with the practice type of input. This type of criticisms frequently unfit to fulfill the understudies in further developing their opportunity for growth. It is about time for the teachers to reevaluate about the input giving cycle. They ought to keep away from conventional approach to demonstrating input towards the understudies. This paper accompanies some advanced and innovation based approach to demonstrating criticism which can ultimately help understudies in further developing understudy learning experience. This can likewise help in professionalizing the educating of speakers in higher instruction. The overarching difficulties in the conventional arrangement of criticism are diverse. Teachers frequently end up caught in the unbending nature of conventional structures, which, more frequently than not, miss the mark regarding tending to the nuanced learning prerequisites of understudies. The one-size-fits-all approach neglects to fulfill the different necessities of a cutting edge understudy body, reducing the effect of input on their learning process.

B. Client Criticism Data Framework for Quality Improvement

Creator: Ke Wang

Year: 2014

This examination tended to the essential requirements for a successful client input data framework and the data set innovation to foster the framework. It talked about the framework idea what's more, design of a regular client input data framework, data set administration innovation, programming stream diagrams, program capabilities and capacities. A proposed client criticism data framework comprises of client information inputs, data set the board framework and results to different offices in an association. With a userfriendly interface, data set administration framework fills in as a data the board tool.It can handle the criticism from clients and utilize the data in navigation. Accordingly, client input can be expeditiously and really used to make remedial activities by assembling and administration offices. The possibility of persistent improvement can be really did to meet and surpass clients' fulfillments and assumptions. Evans what's more, Lindsay (1993) showed that "quality starts with the shopper." Coshliller requests what's more, consistent mechanical changes have opened new and profoundly aggressive business sectors.

III. RELATED WORK

Alongside giving a typical framework hypothetical structure for the developing writing on decency in consecutive direction (currently examined in Sec. 1), our point is to prepare for a productive joint effort with various networks dealing with, e.g., dissemination shifts, ill-disposed AI, control hypothesis, and ideal vehicle. In what follows, we elaborate more on these examination bearings and on future interdisciplinary methodologies. Dispersion Movements. Many works have researched ML under various kinds of conveyance shifts over the long haul. The issue of idea float is comprehensively characterized as a shift of the objective dispersion over the long run [26, 70]. This is a fairly expansive definition, which incorporates circulation shifts because of exogenous impacts, e.g., a pandemic or a monetary emergency.

Nonetheless, such moves can be erratic and don't expect that criticism circles are available overall. As of late, endogenous dispersion shifts, i.e., target appropriation shifts brought about by the conveyed expectation model, have been examined all the more completely. The idea of performative expectations recognizes the way that ML-based dynamic frameworks can influence the result they attempt to foresee [57]. The thought of performative dependability, which is characterized as an indicator that isn't just aligned against verifiable information yet in addition against future results that are delivered by acting in light of the expectation, is a potential arrangement that accomplishes a steady point for retraining [57]. This steady point implies that a model remaining parts the very same assuming it is retrained on future results. Performative expectation is an umbrella term for a circumstance where ML-based choices cause a change in the result dissemination. In any case, this conveyance shift can happen through a criticism circle we presented in Segment 3.1.11 As we displayed in Area 5, these input circles have various properties and suggestions. For instance, changing a stage client's viewpoint (through a singular input circle) is totally different from changing the person's acknowledged result (through a result criticism circle). In all cases, the suggestion changes the singular's utilization. In the previous case, this is brought about by formed inclinations. Conversely, the choice important individual credits stay unaltered in the last option case. More examination is expected to concentrate on the impacts of the particular criticism circles we arrange on the idea of performative power [28], which just thinks about moving result conveyances in its more broad comprehension as in the performative forecast writing. Antagonistic Machine Learning. AdversariaMLstudiesattacksonMLalgorithmsandhowtheycanbedefended[37,69]. The thought is that ill-disposed assaults are executed by an aggressor who means to impact some piece of the ML pipeline, while the engineer of the ML calculation defeats the assailant's goal. Conversely, criticism circles don't happen because of noxious outer control yet are an immediate outcome of the elements in successive dynamic frameworks. However, the results of certain ill-disposed assaults are firmly connected with the criticism circles we order in this paper. For instance, information harming assaults are related with ML model criticism circles in that they alter the information utilized for preparing. Applying measures intended to counter ill-disposed assaults to manage criticism circles in successive dynamic frameworks addresses a fascinating road for future examination for instance, vigorous learning through information sub-testing [39] or managed enhancement [45] to counter ML model criticism circles. First outcomes have shown that this turns out to be more confounded if the reasonableness of the dynamic frameworks is a worry [71]. Control Hypothesis and Ideal Vehicle. Control Hypothesis gives the instruments to drive dynamical frameworks towards a state with an ideal way of behaving. These objectives are arrived at by planning a regulator with the necessary remedial way of behaving. Our structure gives a premise to decipher the chief as a regulator that can be deliberately intended to commonly accomplish wanted exhibitions and inclination relief, empowering the utilization of instruments from Control Frameworks hypothesis, for example, Monetary Model Prescient Control [20] or Ideal Control [44]. Utilizing these devices, one could consolidate reasonableness ensures as requirements and execution ensures, e.g., commitment on a web-based stage, as the goal capability, potentially in a Lament minimization design [7]. For the instance of ill-disposed criticism circles, one can contemplate the chief and the outer climate as the two players of a lose situation where Strong Control [77] procedures track down their regular articulation. For instance, alluding to the choice cycle model in Segment 3.1.1, one can display the ill-disposed moves made by the up-and-comers as an unsettling influence for the chief in accomplishing the best competitors determination. While taking a gander at predisposition moderation procedures at the gathering level, devices from Ideal Vehicle can likewise prove to be useful [11]. Ideal Vehicle permits evaluating the infringement of reasonableness limitations as the distance between the gathering's ongoing conveyance and an ideal one [12]. This device would permit the plan of a regulator (leader) that drives the underlying circulation towards an optimal one, satisfying reasonableness limitations.

IV. THE ML-BASED DYNAMIC PIPELINE IN THE EDGE OF DYNAMICAL

Frameworks Hypothesis Representing the way that ML-based dynamic frameworks are generally not static however advance over the long haul, we follow Dobbe et al. [19] in utilizing the language of dynamical frameworks hypothesis to portray them. A dynamical framework is an interaction that relates a bunch of info signs to a bunch of result signals. A sign is a variable or amount of interest that might differ over the long haul. Subsequently, a calculation is an illustration of a dynamical framework that gets discernible highlights as A Characterization of Criticism Circles and Their Connection to Predispositions in Computerized Dynamic Frameworks input signals and creates forecasts or choices as result signals. Dynamical frameworks hypothesis is worried about the numerical demonstrating of dynamical frameworks with the target of understanding as well as controlling essential properties, for example, whether the framework arrives at an anticipated working point or displays oscillatory ways of behaving. It is normal to address dynamical frameworks in block outlines, where blocks signify frameworks and bolts mean signs, as a method for giving a significant level graphical portrayal of a certifiable framework.

Block graphs are especially helpful to comprehend and concentrate on the interconnection of various (sub-)frameworks, which are made to shape bigger frameworks. A series interconnection happens when the result of a framework (or calculation) is the contribution for another.

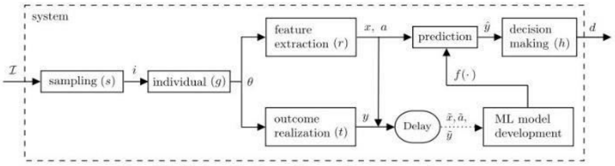


Fig. 1. The ML-based decision-making pipeline as an open-loop system.

A equal interconnection happens when similar info enters two frameworks whose results are then consolidated. In a criticism interconnection, the result of a framework is infused back as a contribution to (at least one) of its parts, making an input circle. Series and equal interconnections lead to open-circle frameworks, though input interconnections lead to shut circle frameworks - see Fig. 4 in Supplement B for a visual portrayal.

The prototypical ML-based dynamic pipeline can likewise be addressed as a block graph. We start by portraying its open-circle parts, displayed in Fig. 1, preceding portraying conceivable criticism interconnections in Area 3. Toward the start of the pipeline, an individual i is tested from the world (i.e., the climate) I, which addresses a sign entering in the testing capability block s : I → i. Let i be the singular's personality - i.e., its record in the populace, which [34] call likely space (PS) - and let g : i → θ be a capability that profits the singular's credits. All the more exactly, θ signifies the build that is pertinent for the expectation - what [24] call develop space (CS). The highlights x, removed through the capability r : θ → x, a, and the result y (additionally called mark or target), acknowledged through the capability t : θ → y, are flawed intermediaries that can be estimated - what [24] call noticed space (operating system). For example, y can address whether an individual reimburses a conceded credit and x is a bunch of elements (for instance, the FICO rating, as generally utilized in the US) that are utilized by the chief to anticipate the reimbursement likelihood yˆ to choose whether to concede the credit or not. For each tested individual, a ultimate conclusion d is educated by the expectation yˆ, which is delivered in view of the noticed elements x to rough y utilizing a learned capability f : x → yˆ.

When the result is noticed, i.e., after one time-unit of deferral, the previous time's element mark pair (x,˜ y˜) can wind up as an example in the dataset (X, Y) that is utilized to (re)train and (re)evaluate a ML model (more subtleties on the ML model improvement process are examined in Supplement C). In completely computerized dynamic frameworks, the choice rule h is exclusively founded on the forecast (h : yˆ → d), typically appearing as a straightforward limit rule, e.g., d = 1 if and provided that yˆ ≥ y¯. The image a shows the touchy trait of the individual (e.g., race or orientation) and might perhaps at the same time.

V. FEEDBACK LOOPS IN THE ML-BASED DECISION-MAKING PIPELINE

Criticism Circles IN THE ML-BASED Dynamic PIPELINE Rather than ML, in the field of dynamical frameworks hypothesis, criticism circles are not generally seen as an unwanted component of a framework. Bunches of the accentuation of dynamical frameworks hypothesis is on relating properties of the open-circle framework, i.e., the framework without a criticism circle, to those of the shut circle framework, i.e., the framework with an input circle. In this paper, we influence shut circle framework properties to characterize criticism components in ML-based navigation frameworks. Curiously, shut circle frameworks might display beneficial properties contrasted with their open-circle partners. In this segment, we complete the particular of the ML pipeline as a dynamical framework by thinking about the criticism interconnections that could be available. We initially characterize different kinds of input circles relying upon the part of the ML pipeline impacted by the result of the framework (i.e., a ultimate choice of the chief). Then, we present the idea of ill-disposed input circles. Then, at that point, we portray how various kinds of criticism circles can exist together. At last, we explain a phrasing regarding positive and negative criticism circles.

A. Feedback Loops

In some genuine settings, the choice taken toward the finish of the ML pipeline might criticism into a portion of its blocks. Each block in the ML pipeline (with the exception of the forecast block, as this normally essentially comprises of applying f to another information model x) can be impacted by the choice, each shaping an alternate kind of criticism circle, as portrayed in Fig. 2. In what follows, we group these criticism circles to give a jargon and a few models. To approve this phrasing, we checked on a sum of 24 late pertinent papers that examine issues of criticism circles with regards to ML-based dynamic frameworks - a considerable lot of which especially center around decency viewpoints. These papers are recorded in Table 1 (we portray the writing search process in more detail in Supplement A). We underline that the grouping of the five A Classification of Feedback Loops and Their

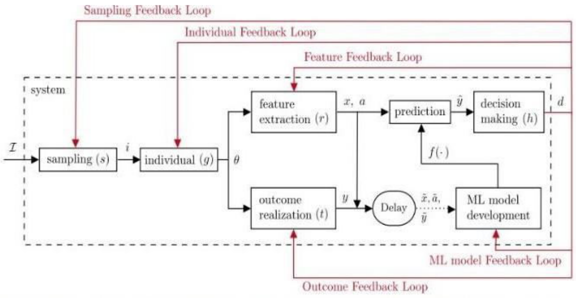
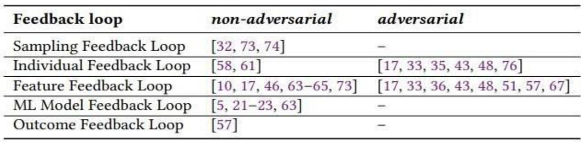


Fig. 2. The ML-based decision-making pipeline as a closed-loop system in which different feedback loops can emerge.

Table 1. Overview of feedback loops in the algorithmic fairness literature



Feedback loop	*non-adversarial*	*adversarial*
Sampling Feedback Loop	[32, 73, 74]	–
Individual Feedback Loop	[58, 61]	[17, 33, 35, 43, 48, 76]
Feature Feedback Loop	[10, 17, 46, 63–65, 73]	[17, 33, 36, 43, 48, 51, 57, 67]
ML Model Feedback Loop	[5, 21–23, 63]	–
Outcome Feedback Loop	[57]	–

criticism circles addressed in Fig. 2 is finished concerning the models and use cases we distinguished in the present status of the writing on fair powerful dynamic frameworks. Regardless of covering existing writing, this input circle arrangement can undoubtedly be stretched out to catch more nuanced sorts of feedback.

Sampling Feedback Loop: The principal kind of input circle we present is the one that contains the impacts of the choice on the inspecting of the person from the populace. This impacts the standard for dependability of various gatherings and adjusts their portrayal. Consider the accompanying illustration of a school confirmation situation examined in [53]. In the first place, let the all out populace be divided into two gatherings A and B. The populace goes through a determination cycle in which a foundation, the leader, plans a strategy that maps every person to a likelihood of being chosen, potentially relying upon the gathering character a and on discernible traits x that bear data about capability, e.g., GPA, SAT, or suggestion letters. As indicated by the creators of [53], the determination interaction at time t could change the capability profiles of one or the other gathering at time $t + 1$ through a self-choice cycle acting through sifting the pool of people accessible at the following emphasis. At the end of the day, with the presence of an inspecting input circle, people having a place with a gathering that had gotten lower confirmation rates at the past emphasis may be deterred from applying as up-and-comers at the following cycle, in this manner influencing the application rates from the two gatherings (and at last the choice rates). Note that, this input circle could prompt one of the two gatherings vanishing from the competitor pool. To grasp this, consider a comparative model connected with discourse acknowledgment items like Amazon's Alexa what's more, Google Home, which have been displayed to have emphasize inclination against non-local speakers [31], with local speakers encountering a lot better than non-local speakers. This distinction can prompt an inspecting input circle, where non-local speakers stop utilizing such items. This might be difficult to distinguish on the grounds that the discourse acknowledgment model, starting there on, just gets info and preparing information from local speakers, possibly bringing about a model that is even more slanted towards the excess clients, i.e., the local speakers. Without mediation, the model turns out to be even less exact for non-local speakers, which builds up the underlying client experience [32]. Extra instances of the examining criticism circle can likewise be found in [73, 74].

Individual Feedback: Loop One more conceivable impact of the choice demonstrations straightforwardly on the singular's attributes θ, i.e., through the capability g. An illustration of this kind of criticism circle can be tracked down in the clients' responses to customized proposals. As examined in [58, 61], a client's perspective on, e.g., a specific policy centered issue, is impacted by the news stories got. In this way, the choice of the recommender framework to advance a particular sort of satisfied shifts the assessment of the people that get such a suggestion. Extra instances of the singular criticism circle are examined with regards to ill-disposed input circles (see Sec. 3.2)

Feature Feedback Loop: The third kind of input circle is moderately near the past one. Nonetheless, in difference to the singular input circle, the choice affects the noticeable qualities of the individual instead of on the real ones, i.e., on x as opposed to θ. One of the most widely recognized instances of this element input circle can be found in credit loaning situations in which a moneylender chooses whether or not to endorse an advance application in light of the candidate's FICO rating, which is deciphered as a quantifiable and detectable intermediary for the singular's capacity of taking care of a conceded credit [46]. For any certain choice, we notice an element input circle: in the event that the advance is reimbursed, the FICO rating increments; in any case, on the off chance that the candidate defaults, the FICO assessment diminishes. Note that, in this model, the input circle happens provided that the choice is positive, and it likewise requires data on the genuine result y. Notwithstanding, none of these circumstances is completely essential for an element criticism circle to occur.4 One more model is comprised by satisfied recommender frameworks where the time a client sees some happy is part of the perception caught in the element x [10, 63]. In any case, the time unequivocally relies upon what the recommender framework has recently proposed, subsequently shutting a component criticism circle. This happens independent of whether this proposal influences the singular's advantages, i.e., even without any a singular input circle. Extra instances of the component criticism circle can likewise be found in [17, 64, 65, 73]. Moreover, correspondingly to the singular input circle, additionally for the component criticism circle, there exists an ill-disposed partner (see Sec. 3.2).

ML Model Feedback Loop: In the ML model criticism circle, a ultimate choice d influences the ML model by changing the preparation or the approval informational collections (X, Y) that will be utilized for future expectations. Run of the mill models in this class are known as ML-based decision-production with restricted [22] or fractional input [5] and the explanation is that ML models are retrained utilizing recently accessible information. ML model criticism circles portray the situation when the information that becomes recently accessible over the long haul relies upon the choice taken. For instance, recruiting calculations just find out about the abilities of the up-and-comers who were employed, credit loaning calculations just get reimbursing likelihood data from individuals who got the credit, and prescient policing calculations just register wrongdoing in watched areas. In every one of these situations, the choice will make a door to the pair (x, y), which will be added to the current informational index (X, Y) just when the choice is positive $(d = 1)$. Notice that, while the retraining of the model doesn't rely upon the choice (i.e., if the include mark pair (x, y) is added to the current informational collection autonomously of d), there is no ML model input circle. Utilizing the language of dynamical frameworks hypothesis, this case is just seen as an open-circle framework with memory where the state variable (X, Y) advances as per the inward elements, yet autonomously of the result variable (the choice d). Extra instances of the ML model criticism circle can likewise be found in [21, 23, 63].

Outcome Feedback Loop: At long last, in the result criticism circle, the choice d influences the result y before it is understood and eventually noticed. Notice that this noticed result then should be reused in some structure all together to close the circle. Specifically, it possibly frames a circle in the event that the result is utilized, e.g., as a component of the preparation or approval information while retraining the model5 . To perceive how a result criticism circle can emerge, rethink the credit loaning situation: on the off chance that an individual is anticipated at high gamble of default, the credit may be conceded, yet at a higher loan cost. Notwithstanding, the choice to implement a higher loan fee further builds the possibilities that the client defaults [57]. Rather than the model gave in Segment 3.1.3, here we expect that the moneylender's choice d meaningfully affects the acknowledgment of the result y, i.e., regardless of whether the advance is repaid, instead of on the elements (FICO rating).

VI. ADVERSARIAL FEEDBACK LOOPS

A portion of the recently portrayed criticism circles can appear as what we call ill-disposed input circles, which can emerge if the choice d is interwoven with an ill-disposed response to it. While there is no visual distinction with regard to the block-outline portrayal, ill-disposed criticism circles contrast from their non-antagonistic partner in that the choice triggers the response of the individual(s) exposed to the dynamic cycle, which then, at that point, influences the ML-pipeline. By and by, in ill-disposed criticism circles, people exposed to the dynamic cycle respond decisively to the past choices by making moves that increment their possibilities getting ideal choices. Notice that this is a significant qualification for the plan of measures that expect to control the framework's elements, e.g., through predisposition moderation procedures (as we will talk about in Segment 6 in more detail). For example, think about the consideration designation issue examined in [17]. Here, the chief has restricted (inadequate) assets to thoroughly assess N various areas, and thusly they need to choose where to (powerfully) designate the consideration. As the creators contend, the episode pace of each of the N destinations answers progressively (and adversarially) to the past assignment, i.e., it increments where there was definitely no control, as well as the other way around it diminishes relatively to the measure of assessment. Basically, this model depicts the instance of an ill-disposed individual input circle, in light of the fact that the choice at last influences the episode rate, i.e., θ. To give another model, consider a school that distributes the choice rule for its confirmation strategy. Planned understudies can decisively put resources into their own capabilities to meet the necessities. If this activity genuinely changes the arrangement level of the understudy [48], then it is again an ill-disposed individual input circle. Notwithstanding, it is likewise conceivable that main the discernible elements of the individual are changed [33], e.g., assuming the understudies put resources into SAT test arrangement without changing their genuine capability for the school. Then, we are confronting an ill-disposed include criticism circle. Essentially, on the off chance that an individual is applying for a credit, it very well may be helpful to open various credit lines to further develop their detectable elements [57]. This activity isn't really changing the singular's capacity of taking care of the credit, however it is just a method for gaming the dynamic strategy, in this way we have an ill-disposed highlight criticism circle. Extra instances of ill-disposed individual and element criticism circles can be found in [33, 35, 43, 76] and [17, 36, 43, 51, 67], separately. Nonetheless, we underline that recognizing the individual isn't simple 100% of the time also, the component ill-disposed criticism circles, in light of the fact that large numbers of these works expect that the choice influences the capability θ of the people, however generally they mean that it just influences its noticeable highlights x.

VII. COEXISTENCE OF FEEDBACK LOOPS

As found in the past areas, different criticism circles can exist together inside a similar application space. For example, the recommender frameworks for an internet based stage can influence the assessment of the clients θ (individual input circle) or just their portrayal in the component space x (highlight criticism circle). School affirmation strategies can actuate understudies to work on their capability (antagonistic individual input circle) or simply their portrayal x (ill-disposed include criticism circle). On the other hand, they can likewise prompt different standards for dependability across gatherings (testing criticism circle). Loaning choices can influence a singular's reliability θ (individual criticism circle), FICO assessment x (include input circle), acknowledged result y (i.e., whether the allowed credit is taken care of, addressing a result criticism circle), or even the information utilized for the ML model turn of events (X, Y) (bringing about a ML model input circle) or the test of people applying for a credit in any case (causing an examining criticism circle). Every one of the five grouped input circles address some causal impact of a ultimate conclusion on one more part of the ML-based dynamic pipeline. Accordingly, which type(s) of criticism loop(s) (co)exists exclusively relies upon the setting explicit suppositions in regards to the fundamental causal impacts of the choice. The chance of the conjunction of various mixes of criticism circles brings about coupled conduct and, surprisingly, more perplexing elements.

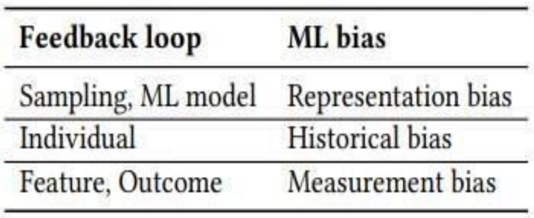
VIII. POSITIVE/NEGATIVE FEEDBACK LOOPS AND RELATION TO STABILITY

In many disciplines, including the ML people group, an extensive accentuation is put on characterizing criticism circles as one or the other positive or negative [38, 54, 60]. This is many times joined by some uncertainty in the meaning of these ideas. In frameworks hypothesis, a positive criticism circle (otherwise called building up) enhances the impact of contributions on the yields, while a negative input circle (otherwise called adjusting) constricts it. In different spaces, the idea of a positive/negative input circle is some of the time related with attractive/unwanted results, paying little heed to how it acts to intensify/weaken inputs. For instance, the criticism circle that increments recidivism because of detained people's diminished admittance to back is alluded to as a negative criticism circle in [75, p. 2]. This equivocalness is tricky, particularly taking into account that in frameworks hypothesis the ideal objective is frequently to make the result an anticipated capability of the info and free from other exogenous yet unavoidable sources of info (considered as unsettling influences). Therefore, appropriately planned negative criticism circles are considered best, while positive input circles are frequently thought to be dangerous. Notwithstanding, frameworks hypothesis frequently puts more accentuation on the soundness of the shut circle framework instead of grouping criticism circles as sure or negative. A steady framework joins to an anticipated harmony point, while an unsteady framework either sways or develops past limits. It is natural to connect positive criticism with precariousness and negative input with security, in any case, this instinct isn't general [2, 72]. From one viewpoint, positive criticism is ensured to prompt unsteadiness just in the extraordinary class of straight frameworks. The presence of non-linearity (e.g., immersion or on the other hand hysteresis) can balance out a positive criticism circle, which is deliberately presented as a rule (e.g., the plan of signal speakers). Then again, negative criticism doesn't ensure strength (even in straight frameworks). Additionally, a similar framework could be in one or the other positive or negative criticism relying upon the working system (e.g., the recurrence of the information signal). Hence, in this paper, we shift the concentration from grouping criticism circles as certain/negative to inquiring whether the shut circle framework meets (or not) to a (alluring) state. As we will find in the models in Area 5, criticism circles frequently drive the ML-based choice framework to stable balance focuses over the long haul.

XI. FEEDBACK LOOPS AND ALGORITHMIC BIASES

Having the option to reason about what caused particular kinds of predisposition is of staggering commonsense significance to stay away from or balance them in the long haul. In any case, ML-based dynamic frameworks can bring about socially unfortunate results over the long run. Many works guarantee that those predispositions can be propagated or even supported because of input circles [3, 13, 14, 40, 49, 50, 55, 56, 68]. Notwithstanding, an unmistakable comprehension of the causal impacts of input circles on algorithmic predispositions is at present missing. We fill this hole by interfacing the grouping of criticism circles (which we presented in Segment 3.1) to algorithmic predispositions and make sense of in more detail which sorts of predisposition they influence. Table 2 gives an outline of the associations we lay out. Be that as it may, the term 'inclination' can have various implications and be utilized reciprocally with equivalent words for various kinds of predisposition. To guarantee consistency, we take on the ideas and phrasing presented by Suresh and Guttag [66]. Portrayal Inclination. As indicated by [66], there are various subtleties of portrayal predisposition: Portrayal inclination can emerge (I) on the off chance that the characterized target populace doesn't mirror the utilization populace, (ii) assuming the objective populace contains. A Classification of Feedback Loops and Their Relation to Biases in Automated Decision-Making Systems underrepresented gatherings, and (iii) in the event that the tested gathering of people isn't illustrative of the objective populace. All three renditions address some distinction between the utilized dataset (X, Y) and the populace I. Inspecting input circles can influence portrayal inclination. Inspecting criticism circles influence the testing capability s that yields a bunch of people on which a ML-based dynamic framework acts. A testing criticism circle changes the example of people for whom an expectation and, at last, a choice is made (i.e., the individuals who have an opportunity to be chosen). In this manner, it can bring about portrayal predisposition, which depicts what is going on in which s undersamples some piece of the populace. Thus, the accessible information isn't illustrative of I and, hence, the ML model probably does not sum up well for the burdened gathering [66]. ML model input circles can likewise influence portrayal predisposition. The ML model criticism circle changes the example of

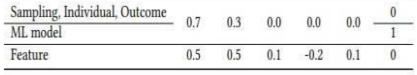
Table 2. Feedback loops and the ML biases they affect



Feedback loop	ML bias
Sampling, ML model	Representation bias
Individual	Historical bias
Feature, Outcome	Measurement bias

people whose acknowledged result becomes noticeable, i.e., those that are chosen and can hence be added as a new highlight mark pair (x, y) to the example (X, Y) - see Fig. 5 in Supplement C for a perception of this interaction. Along these lines, it can influence portrayal predisposition, which originates from a change in the preparation information distributions.6 Authentic Predisposition. Individual input circles can influence authentic predisposition. Individual criticism circles follow up on the build space (CS) of an individual, i.e., the inborn properties of an individual θ change and not just the noticed intermediaries x, y, which are estimated in the noticed space (operating system) [24]. This can bring about verifiable inclination (additionally called "life predisposition" [34]), which portrays shameful acts that manifest in imbalance between bunches in the CS. As choices can change people's properties θ, which can appear in modified future elements x, it turns out to be more hard to treat people reasonably since the choice really transformed them. This implies that the world is precisely addressed by the information (i.e., the estimation capabilities r and t are satisfactory), however the condition of the world (i.e., a person's inborn choice important properties θ) is the aftereffect of uncalled for medicines in past choice rounds [66]. For instance, not thinking about counterfactual choices for people (i.e., accepting that people would have developed indistinguishably over the long run, regardless of whether they had been alloted various choices) can drive the choice framework to a state in which people are burdened exclusively in light of an unfortunate occasion previously, regardless of whether their properties are totally quantifiable. Estimation Predisposition. Result input circles and element criticism circles can influence estimation predisposition [24, 50, 55, 66, 68]. These two criticism circles follow up on the estimation functions$r$ and t and in this manner influence a person's perceptible properties x, a, y. The result criticism circle changes the acknowledgment of the result (y). Interestingly, the component criticism circle changes the recognizable qualities that are taken care of into the expectation model (x and, possibly, a), i.e., the highlights for future choices. Hence, the two sorts of input circles can influence estimation inclination: the highlights x and names y are generally only intermediaries as they attempt to gauge an innate property of a person, which could address a develop that isn't straightforwardly quantifiable or even recognizable (θ) [66]. Estimation predisposition portrays the progress among CS and Operating system [24]. In this manner, it portrays what is happening in which those intermediaries less intently estimated the expected trait beyond a shadow of a doubt.

Table 3. Initial conditions for the different experiments. The acronyms stand for group 1 (G1), group 2 (G2), population size (n), training sample size (n_{train}), distribution mean (μ) and standard deviation (σ), distribution of the feature realization (r), distribution of the outcome realization during the simulation (t) and for the initial training set (t_{train}). For all experiments, we set the following parameters: $n = 1000$, $n_{\text{train},G1} = n_{\text{train},G2} = 500$, $\sigma_{\theta,G2} = 0.15$, $\sigma_{t,G1} = 0.1$, $\mu_{r,G1} = 0$, $\mu_{t,G1} = 0$, $\sigma_{\theta,G1} = 0.15$, $\mu_{t,G2} = 0$, $\sigma_{t,G2} = 0.1$, $\mu_{t_{\text{train}}} = 0$. In the table, we describe the parameters that vary from one experiment to another.



Feedback loop	$\mu_{\theta,G1}$	$\mu_{\theta,G2}$	$\sigma_{r,G1}$	$\mu_{r,G2}$	$\sigma_{r,G2}$	$\sigma_{t_{\text{train}}}$
Sampling, Individual, Outcome	0.7	0.3	0.0	0.0	0.0	0
ML model						1
Feature	0.5	0.5	0.1	-0.2	0.1	0

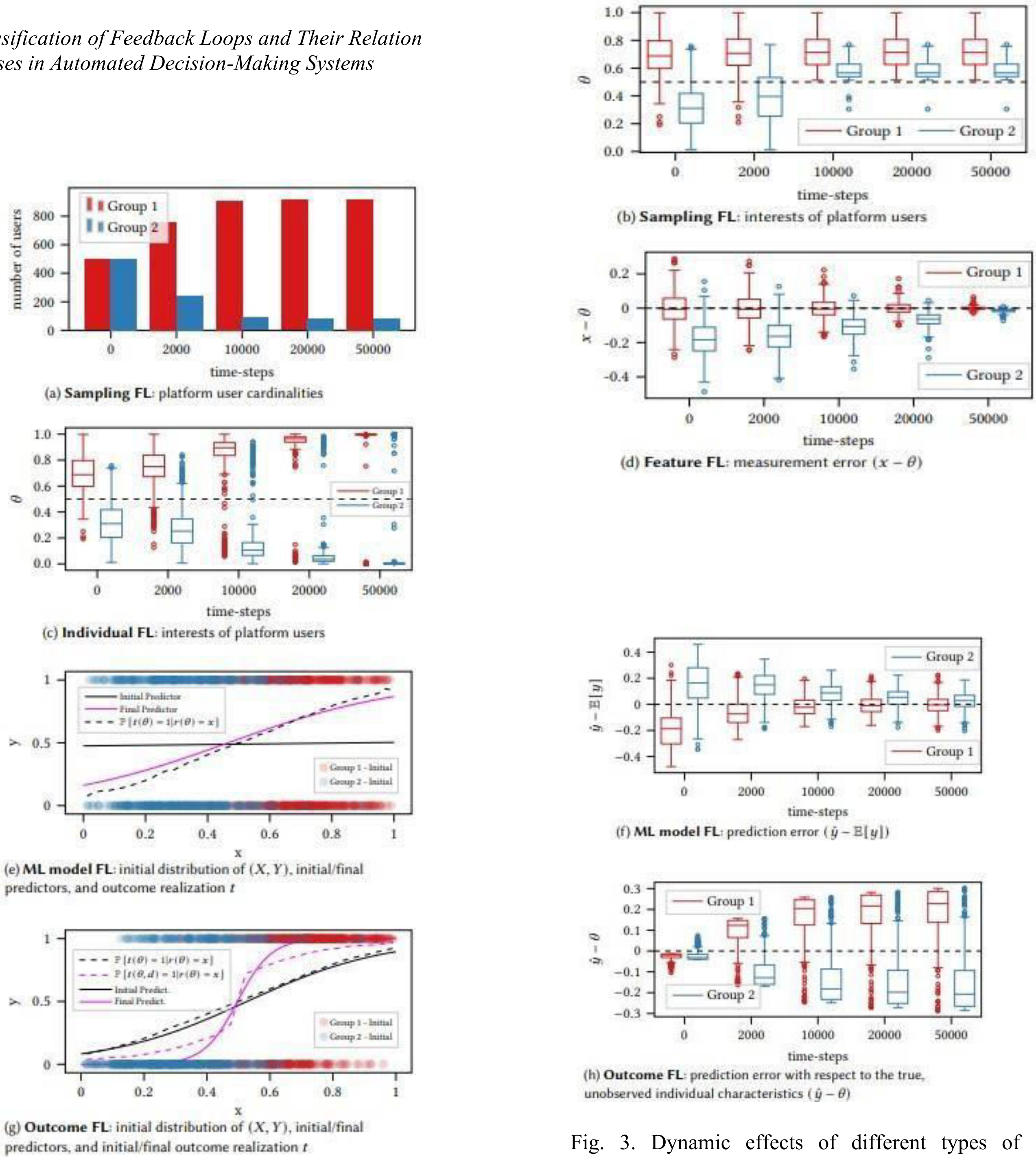
people or gatherings, and that intends that r or t (or both) are not fitting to catch the significant construct.7 For model, involving captures as an intermediary for the gamble of carrying out a wrongdoing (similar to the case in the recidivism risk expectation device COMPAS [1]) is risky assuming there are bunches that are considerably more liable to be captured for specific wrongdoings.

To exhibit the capability of our arrangement of input circles and their connection to the various predispositions, we present a bringing together contextual investigation on recommender frameworks (RS).8 We consider the instance of an internet based stage where the RS is utilized to give content the clients are keen on. For straightforwardness, we think about only one important thing (e.g., a particular video) what's more, signify a client's advantage in this thing with $\theta \in [0, 1]$, where bigger θ relates to higher interest. The understood result y signifies whether a client shows interest (e.g., taps on the pertinent thing being referred to), $y = 1$, or not, $y = 0$. The stage utilizes a RS to foresee a client's advantage $\hat{y} = f\ (x)$, where the component $x \in [0, 1]$ addresses the client's past clicking conduct on the stage. For this straightforward model, x is the level of suggested significant things that the client has tapped on previously and consequently fills in as an intermediary of the client's advantage in the important thing. The capability $f\ : [0, 1] \rightarrow [0, 1]$ is learned through a calculated relapse (LR) calculation (which is fitted to a sigmoid capability) prepared on information (X, Y), which comprises of an assortment of element mark matches (x, y).

To conclude whether the significant thing ought to be displayed as one of the top proposals ($d = 1$) or not ($d = 0$), the accompanying limit rule is utilized: $d = 1$ if $\hat{y} > 0.5$, what's more, $d = 0$ in any case. After every proposal round, y is noticed, (x, y) is added to the current dataset (X, Y), and the RS is retrained. We think about two gatherings of clients $a \in \{\text{G1, G2}\}$. For effortlessness, a isn't utilized as a contribution for the RS.

We currently give one guide to each sort of criticism circle depicted in Segment 3.1 to show how they are related with various predispositions. The underlying circumstances well defined for every one of these reenactment models are portrayed in Table 3 and the underlying θ dissemination is displayed in Fig. 6 in Reference section D. Notice that the mean of θ is higher for bunch G1overall.

Inspecting Input Circle. To start with, we take a gander at a unique case in which $d = 0$ compares to not getting any recom mendation, prompting clients leaving the stage. All things being equal, while getting $d = 1$, clients stay on the stage. At first, around half of the dynamic clients on the stage are from bunch 1 and half from bunch 2: $nG1 = 496$ and $nG2 = 504$. Each time somebody leaves the stage, another client replaces them. To emulate clients' homophily, the new client is drawn from



(a) **Sampling FL**: platform user cardinalities

(b) **Sampling FL**: interests of platform users

(c) **Individual FL**: interests of platform users

(d) **Feature FL**: measurement error $(x-\theta)$

(e) **ML model FL**: initial distribution of (X, Y), initial/final predictors, and outcome realization t

(f) **ML model FL**: prediction error $(\hat{y}-\mathbb{E}[y])$

(g) **Outcome FL**: initial distribution of (X, Y), initial/final predictors, and initial/final outcome realization t

(h) **Outcome FL**: prediction error with respect to the true, unobserved individual characteristics $(\hat{y}-\theta)$

Fig. 3. Dynamic effects of different types of feedback loops (FL) acting on an RS pipeline for an online platform. Circles in the box plots denote outliers.

bunch 1 with likelihood $p = nG1n$(else, from bunch 2), i.e., the higher the level of clients from bunch 1 in the stage, the higher the likelihood the new client has a place with bunch 1. As should be visible in Fig. 3a, this peculiarity leads to the decrease of $nG2$ from 504 to 89 people after 10,000 time-steps. This dispersion perseveres in future time-steps, recommending that it is a (locally) stable balance point of the dynamical framework. Bunch 2 is underrepresented on the stage in the long haul with only 8.9% of the stage clients. This relates to subtlety (ii) of the portrayal inclination as depicted in Segment 4. Be that as it may, simultaneously, subtlety (iii) of the portrayal inclination is available for the two gatherings: since just those given $d = 1$ stay on the stage, the example of dynamic clients turns out to be less delegate over the long haul, i.e., just intrigued clients (those with high qualities for θ) remain on the stage (see Fig. 3b).

Notice that it is hard to order the testing criticism circle as sure or negative for this situation, as there is no underlying portrayal inclination against Gathering 2 that gets enhanced by the circle. The subsequent one-sided balance point is basically a property of the shut circle elements. Individual Criticism Circle. An illustration of a singular input circle is the point at which the suggested content impacts the client's viewpoint θ, which we model by allowing the new assessment to be a raised blend of the past one and the suggested content. Fig. 3c shows that this outcomes in a polarization of interests on the stage. Specifically, clients with high starting interest (i.e., $\theta > 0.5$) are bound to be suggested the thing and, thus, their θ further increments after some time, as well as the other way around for clients with low starting interest.9 Because of the underlying distinction in the dispersions of θ (see Table 3), authentic predisposition increments. In particular, results show greater gathering level differences with an extremely high θ for by and large. The consistent state esteem arrived at by the directions in Fig. 3c addresses a one-sided stable balance point of the shut circle ML framework in which the conclusions are energized. Include Criticism Circle. Fig. 3d shows the consequence of a model in which the substance proposal takes care of once again into the element extraction block x (as opposed to following up on the genuine assessment θ, similar to the case in the past model), subsequently framing a component criticism circle. Contrasted with the wide range of various models, there is no distinction in the mean of the underlying θ circulation across bunches G1 and G2.

turn into the ones bound to get positive choices. Here, we measure the forecast mistake as $y\hat{} - \mathrm{E}[y]$, in any case, without any estimation mistake in the result acknowledgment ($t(\theta) = y$), θ is roughly identical to $\mathrm{E}[y]$ with the exception of some commotion, which is unimportant for the typical over a gathering of people. As should be visible in Fig. 3f, the expectation mistake rapidly moves toward 0 for G1, however the LR calculation keeps on performing inadequately for G2 in the short to medium term. In the long haul, because of the commotion in the perception of x 10 it at last methodologies 0 additionally for G2. Retraining the ML model over the long run lessens the portrayal inclination, subtlety (ii) of the portrayal predisposition as depicted in Area 4. In any case, it is because of the ML model criticism circle that the example (X, Y) turns out to be more delegate of bunch 1 after really hardly any time-ventures while taking significantly longer to lessen portrayal inclination for bunch 2. Result Input Circle. At last, we consider a model involving similar starting circumstances as in the examining and individual criticism circles, yet this time the RS's choice influences the result acknowledgment t. In particular, the likelihood of the acknowledged result to be $y = 1$ increments/diminishes by 20% for positive/adverse choices, individually. This implies that the acknowledged results $t(\theta, d)$ are more limit than they would be assuming there were no result input circle (see run lines in Fig. 3g). Regardless of beginning with an unprejudiced ML model, over the long haul, the retrained ML model approximates $t(\theta,$ $d)$, i.e., the underlying indicator is a lot compliment sigmoid capability contrasted with the last indicator. To be specific, the result criticism circle presents an estimation predisposition on the acknowledged result y for the two gatherings G1 and G2. Accordingly, as is noticeable in Fig. 3h, the expectation mistake $y\hat{} - \theta$ veers from 0 (as $y\hat{}$ predicts the acknowledged result y and not θ) until it comes to a stable balance point after around 10,000 time-ventures (at roughly 0.2 and - 0.2 for G1 and G2). From the point of view of stage clients, a result criticism circle can bring about a circumstance in which one continues to get proposals due to having tapped on comparative substance previously, in spite of not being keen on it.

RESULT

The result of ML-based dynamic frameworks, i.e., the choice, frequently influences different pieces of the actual framework, making a supposed criticism circle. However, ML assessment methods generally preclude possibly significant fleeting dynam ics [13, 46, 52] and considering input circles is urgent to stay away from potentially negative results [17, 46, 65, 74]. In this work, we expand on dynamical frameworks hypothesis to give an overall system that reveals insight into the various kinds of input circles that can happen all through the ML pipeline. We distinguish five particular sorts of criticism circles, some of which can be delegated "ill-disposed" at whatever point the choice feeds once more into the framework as a result of some key activity of the impacted individual(s). Besides, we partner the various sorts of criticism circles with the comparing predispositions they influence, and we By thoroughly breaking down the ML pipeline, we accept that our system is a fundamental primer step towards (I) understanding the specific job of the input circles and (ii) moving the examination center from foolhardy arrangements that mean to distinguish and address existing predispositions to a more forward-looking methodology that tries to expect and forestall inclinations in the long haul. To begin with, giving a thorough characterization of criticism circles will prepare for an efficient survey of existing works in the ML writing and it will permit placing their outcomes into the viewpoint of their suspicions (e.g., which sorts of criticism circles are thought of and which are not). Second, with the assistance of extra apparatuses, e.g., dynamical frameworks and control hypothesis, it will be feasible to completely take advantage of the capability of our structure in the deliberate plan of criticism circles, and for the improvement of powerful long haul shamefulness relief techniques demonstrate these elements utilizing a recommender framework model.

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompany a successful completion of any task would be incomplete without the mention of people who made it possible, success is the epitome of hard work and perseverance, but steadfast of all is encouraging guidance. So, with gratitude we acknowledge all those whose guidance and encouragement served as beacon of light and crowned the effort with success. We thank our project guide Mr. Mahesh Kini M, Associate Professor, in Department of Computer Science & Engineering, who has been our source of inspiration. He has been especially enthusiastic in giving his valuable guidance and critical reviews. The selection of this project work as well as the timely completion is mainly due to the interest and persuasion of our project coordinator Mrs. Vidya, Senior Assistant Professor, Department of Computer Science & Engineering. We will remember her contribution forever. We sincerely thank, Dr. Manjunath Kotari, Professor and Head, Department of Computer Science & Engineering who has been the constant driving force behind the completion of the project. We thank our beloved Principal Dr. Peter Fernandes, for his constant help and support throughout. We are indebted to Management of Alva's Institute of Engineering and Technology, Mijar, Moodbidri for providing an environment which helped us in completing our project. Also, we thank all the teaching and non-teaching staff of Department of Computer Science & Engineering for the help rendered.

REFERENCES

[1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias. ProPublica, May 23, 2016 (2016), 139–159. https://www.propublica. org/article/machine-bias-risk-assessments-in-criminal-sentencing

[2] Karl Johan Åström and Richard M Murray. 2021. Feedback systems: an introduction for scientists and engineers. Princeton university press.

[3] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. Fairness and Machine Learning. fairmlbook.org. http://www.fairmlbook.org

[4] Joachim Baumann, Anikó Hannák, and Christoph Heitz. 2022. Enforcing Group Fairness in Algorithmic Decision Making: Utility Maximization Under Sufficiency. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22). Association for Computing Machinery, New York, NY, USA, 2315–2326. https://doi.org/10.1145/3531146.3534645

[5] Yahav Bechavod, Katrina Ligett, Aaron Roth, Bo Waggoner, and Zhiwei Steven Wu. 2019. Equal opportunity in online classification with partial feedback. Advances in Neural Information Processing Systems 32, NeurIPS (2019).

[6] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2021. Fairness in Criminal Justice Risk Assessments: The State of the Art. Sociological Methods & Research 50, 1 (2021), 3–44. https://doi.org/10.1177/0049124118782533

[7] Avrim Blum and Yishay Monsour. 2007. Learning, regret minimization, and equilibria. (2007).

[8] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research, Vol. 81), Sorelle A Friedler and Christo Wilson (Eds.). PMLR, 77–91. https://proceedings.mlr.press/v81/buolamwini18a.html

[9] Simon Caton and Christian Haas. 2020. Fairness in Machine Learning: A Survey. (2020). http://arxiv.org/abs/2010.04053

[10] Allison J. B. Chaney, Brandon M. Stewart, and Barbara E. Engelhardt. 2018. How algorithmic confounding in recommendation systems increases homogeneity and decreases utility. (2018), 224–232. https://doi.org/10.1145/3240323.3240370

[11] Yongxin Chen, Tryphon T. Georgiou, and Michele Pavon. 2021. Optimal Transport in Systems and Control. Annual Review of Control, Robotics, and Autonomous Systems 4, 1 (2021), 89–113.

https://doi.org/10.1146/annurev-control-070220-100858

[12] Silvia Chiappa, Ray Jiang, Tom Stepleton, Aldo Pacchiano, Heinrich Jiang, and John Aslanides. 2020. A General Approach to Fairness with Optimal Transport. The 34th AAAI Conference on Artificial Intelligence (2020). https://ojs.aaai.org/index.php/AAAI/article/view/5771

[13] Alexandra Chouldechova and Aaron Roth. 2018. The Frontiers of Fairness in Machine Learning. (2018), 1–13. http://arxiv.org/abs/1810.08810

[14] Alexandra Chouldechova and Aaron Roth. 2020. A Snapshot of the Frontiers of Fairness in Machine Learning. Commun. ACM 63, 5 (4 2020), 82–89. https://doi.org/10.1145/3376898

[15] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic Decision Making and the Cost of Fairness. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '17). Association for Computing Machinery, New York, NY, USA, 797–806. https://doi.org/10.1145/3097983.3098095

[16] Kate Crawford. 2016. Artificial intelligence's white guy problem. The New York Times 25, 06 (2016).

[17] Alexander D'Amour, Hansa Srinivasan, James Atwood, Pallavi Baljekar, D. Sculley, and Yoni Halpern. 2020. Fairness is not static: Deeper understanding of long term fairness via simulation studies. FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (2020), 525–534. https://doi.org/10.1145/3351095.3372878

[18] Robyn M Dawes, David Faust, and Paul E Meehl. 1989. Clinical Versus Actuarial Judgment. Science 243, 4899 (1989), 1668–1674. https://doi.org/10.1126/science.2648573

[19] Roel Dobbe, Sarah Dean, Thomas Gilbert, and Nitin Kohli. 2018. A broader view on bias in automated decision-making: Reflecting on epistemology and dynamics. arXiv preprint arXiv:1807.00553 (2018).

[20] Matthew Ellis, Helen Durand, and Panagiotis D. Christofides. 2014. A tutorial review of economic model predictive control methods. Journal of Process Control 24, 8 (2014). http://dx.doi.org/10.1016/j.jprocont.2014.03.01

[21] Hadi Elzayn, Michael Kearns, Shahin Jabbari, Seth Neel, Zachary Schutzman, Christopher Jung, and Aaron Roth. 2019. Fair algorithms for learning in allocation problems. FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency (2019), 170–179. https://doi.org/10.1145/3287560.3287571

[22] Danielle Ensign, Sorelle A Friedler, Scott Neville, Carlos Scheidegger, Suresh Venkatasubramanian.

[23] Danielle Ensign, Sorelle A Friedler, Scott Neville, Carlos Scheidegger, Suresh Venkatasubramanian, and Christo Wilson. 2018. Runaway Feedback Loops in Predictive Policing. In Proceedings of Machine Learning Research, Vol. 81. 1–12. https://github.com/algofairness/

[24] Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2016. On the (im)possibility of fairness. https://arxiv.org/abs/1609.07236

[25] Andreas Fuster, Paul Goldsmith-Pinkham, Tarun Ramadorai, and Ansgar Walther. 2022. Predictably Unequal? The Effects of Machine Learning on Credit Markets. Journal of Finance 77, 1 (2022), 5–47. https://doi.org/10.1111/jofi.13090

[26] João Gama, Indrundefined Žliobaitundefined, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. 2014. A Survey on Concept Drift Adaptation. ACM Comput. Surv. 46, 4 (3 2014). https://doi.org/10.1145/2523813

[27] W M Grove, D H Zald, B S Lebow, B E Snitz, and C Nelson. 2000. Clinical versus mechanical prediction: a meta-analysis.

Psychological assessment 12, 1 (3 2000), 19–30.

[28] Moritz Hardt, Meena Jagadeesan, and Celestine Mendler-Dünner. 2022. Performative Power. In Advances in Neural Information Processing Systems, Alice H Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (Eds.). https://doi.org/10.48550/ARXIV.2203.17232

[29] Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. 2016. Strategic Classification. In Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science (ITCS '16). Association for Computing Machinery, New York, NY, USA, 111–122. https://doi.org/10.1145/2840728.2840730

[30] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of opportunity in supervised learning. In Advances in Neural Information Processing Systems (NIPS'16). Curran Associates Inc., Red Hook, NY, USA, 3323–3331.

[31] Drew Harwell. 2018. Amazon's Alexa and Google Home show accent bias, with Chinese and Spanish hardest to understand. https://www.scmp. com/magazines/post-magazine/long-reads/article/2156455/amazons-alexa-and-google-home-show-accent-bias

[32] Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. 2018. Fairness Without Demographics in Repeated Loss Minimization. In Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 80), Jennifer Dy and Andreas Krause (Eds.). PMLR, 1929–1938. https://proceedings.mlr.press/v80/hashimoto18a.html

[33] Hoda Heidari, Vedant Nanda, and Krishna P. Gummadi. 2019. On the Long-term Impact of Algorithmic Decision Policies: Effort unfairness and feature segregation through social learning. 36th International Conference on Machine Learning, ICML 2019 2019-June (2019), 4787–4796.

[34] Corinna Hertweck, Christoph Heitz, and Michele Loi. 2021. On the Moral Justification of Statistical Parity. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21). Association for Computing Machinery, New York, NY, USA, 747–757. https: //doi.org/10.1145/3442188.3445936

[35] Lily Hu and Yiling Chen. 2018. A short-term intervention for long-term fairness in the labor market. The Web Conference 2018 - Proceedings of the World Wide Web Conference, WWW 2018 2 (2018), 1389–1398. https://doi.org/10.1145/3178876.3186044

[36] Lily Hu, Nicole Immorlica, and Jennifer Wortman Vaughan. 2019. The Disparate Effects of Strategic Manipulation. In Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19). Association for Computing Machinery, New York, NY, USA, 259–268. https://doi.org/10. 1145/3287560.3287597

[37] Ling Huang, Anthony D Joseph, Blaine Nelson, Benjamin I P Rubinstein, and J D Tygar. 2011. Adversarial Machine Learning. In Proceedings of the 4th ACM Workshop on Security and Artificial Intelligence (AISec '11). Association for Computing Machinery, New York, NY, USA, 43–58. https://doi.org/10.1145/2046684.2046692

[38] Sterman John D. 2000. Business Dynamics: Systems Thinking and Modeling for a Complex World. McGraw-Hill Education.

[39] Michael Kearns and Ming Li. 1993. Learning in the Presence of Malicious Errors. SIAM J. Comput. 22, 4 (1993), 807–837. https://doi.org/10.1137/0222052

[40] Michael Kearns and Aaron Roth. 2019. The Ethical Algorithm: The Science of Socially Aware Algorithm Design. Oxford University Press, Inc., USA.

[41] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2017. Human Decisions and Machine Predictions. Technical Report

23180. National Bureau of Economic Research. https://doi.org/10.3386/w23180

[42] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan. 2018. Algorithmic Fairness. AEA Papers and Proceedings 108 (5 2018), 22–27. https://doi.org/10.1257/pandp.20181018

[43] Jon Kleinberg and Manish Raghavan. 2020. How Do Classifiers Induce Agents to Invest Effort Strategically? ACM Transactions on Economics and Computation 8, 4 (2020). https://doi.org/10.1145/3417742

[44] Frank L Lewis, Draguna Vrabie, and Vassilis L Syrmos. 2012. Optimal control. John Wiley & Sons.

[45] Chang Liu, Bo Li, Yevgeniy Vorobeychik, and Alina Oprea. 2017. Robust Linear Regression Against Training Data Poisoning. In Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security (AISec '17). Association for Computing Machinery, New York, NY, USA, 91–102. https://doi.org/10.1145/3128572.3140447

[46] Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. 2018. Delayed Impact of Fair Machine Learning. In Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 80), Jennifer Dy and Andreas Krause (Eds.). PMLR, 3150–3158. https://proceedings.mlr.press/v80/liu18c.html

[47] Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. 2018. Delayed Impact of Fair Machine Learning. In Proceedings of the 35th International Conference on Machine Learning. https://proceedings.mlr.press/v80/liu18c.html

[48] Lydia T. Liu, Adam Tauman Kalai, Ashia Wilson, Christian Borgs, Nika Haghtalab, and Jennifer Chayes. 2020. The disparate equilibria of algorithmic decision making when individuals invest rationally. FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (2020), 381–391. https://doi.org/10.1145/3351095.3372861

[49] Kristian Lum and William Isaac. 2016. To predict and serve? Significance 13, 5 (2016), 14–19. https://doi.org/10.1111/j.1740-9713.2016.00960.x

[50] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. ACM Comput. Surv. 54, 6 (7 2021). https://doi.org/10.1145/3457607

[51] Smitha Milli, John Miller, Anca D. Dragan, and Moritz Hardt. 2019. The social cost of strategic classification. FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency (2019), 230–239. https://doi.org/10.1145/3287560.3287576

[52] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. 2021. Algorithmic Fairness: Choices, Assumptions, and Definitions. Annual Review of Statistics and Its Application 8, 1 (3 2021), 141–163. https://doi.org/10.1146/annurev-statistics-042720-125902

[53] Hussein Mouzannar, Mesrob I. Ohannessian, and Nathan Srebro. 2019. From fair decision making to social equality. FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency (2019), 359–368. https://doi.org/10.1145/3287560.3287599

[54] Nataša Obermajer, Ravikumar Muthuswamy, Jamie Lesnock, Robert P Edwards, and Pawel Kalinski. 1983. Positive feedback between PGE$_2$ and COX2 redirects the differentiation of human dendritic cells toward stable myeloid-derived suppressor cells. Immunobiology 119, 20 (1983). https://doi.org/10.1182/blood-2011-07-365825

[55] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kıcıman. 2019. Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries. Frontiers in Big Data

2 (7 2019). https://doi.org/10.3389/fdata.2019.00013

[56] Cathy O'neil. 2017. Weapons of math destruction: How big data increases inequality and threatens democracy. Crown.

[57] Juan C. Perdomo, Tijana Zrnic, Celestine Mendler-Dunner, and Moritz Hardt. 2020. Performative prediction. 37th International Conference on Machine Learning, ICML 2020 PartF16814 (2020), 7555–7565.

[58] Nicola Perra and Luis E C Rocha. 2019. Modelling opinion dynamics in the age of algorithmic personalisation. Scientific reports 9, 1 (2019), 1–11.

[59] Dana Pessach and Erez Shmueli. 2022. A Review on Fairness in Machine Learning. ACM Comput. Surv. 55, 3 (2 2022). https://doi.org/10.1145/3494672

[60] Arkalgud Ramaprasad. 1983. On the definition of feedback. Journal of the Society for General Systems Research 28, 1 (1983). https://doi.org/10.1002/ bs.3830280103

[61] Wilbert Samuel Rossi, Jan Willem Polderman, and Paolo Frasca. 2021. The closed loop between opinion formation and personalised recommendations. IEEE Transactions on Control of Network Systems (2021), 1. https://doi.org/10.1109/TCNS.2021.3105616

[62] Tom Simonite. 2015. Probing the dark side of google's ad-targeting system. MIT Technology Review (2015).

[63] Ayan Sinha, David F Gleich, and Karthik Ramani. 2016. Deconvolving Feedback Loops in Recommender Systems. In Advances in Neural Information Processing Systems, D Lee, M Sugiyama, U Luxburg, I Guyon, and R Garnett (Eds.), Vol. 29. Curran Associates, Inc. https://proceedings.neurips.cc/ paper/2016/file/962e56a8a0b0420d87272a682bfd1e53-Paper.pdf

[64] Yi Sun. 2022. Algorithmic Fairness in Sequential Decision Making. Ph. D. Dissertation.

[65] Yi Sun, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. 2022. The Backfire Effects of Fairness Constraints. ICML 2022 Workshop on Responsible Decision Making in Dynamic Environments (2022). https://responsibledecisionmaking.github.io/assets/pdf/papers/44.pdf

[66] Harini Suresh and John Guttag. 2021. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. In Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '21). Association for Computing Machinery, NewYork, NY, USA. https://doi.org/10.1145/3465416.3483305

[67] Stratis Tsirtsis, Behzad Tabibian, Moein Khajehnejad, Adish Singla, Bernhard Schölkopf, and Manuel Gomez-Rodriguez. 2019. Optimal Decision Making Under Strategic Behavior. (2019). http://arxiv.org/abs/1905.09239

[68] Benjamin van Giffen, Dennis Herhausen, and Tobias Fahse. 2022. Overcoming the pitfalls and perils of algorithms: A classification of machine learning biases and mitigation methods. Journal of Business Research 144 (2022), 93–106. https://doi.org/10.1016/j.jbusres.2022.01.076

[69] Yevgeniy Vorobeychik and Murat Kantarcioglu. 2018. Adversarial Machine Learning. Springer International Publishing, Cham. https://doi.org/10. 1007/978-3-031-01580-9

[70] Geoffrey I Webb, Roy Hyde, Hong Cao, Hai Long Nguyen, and Francois Petitjean. 2016. Characterizing concept drift. Data Mining and Knowledge Discovery 30, 4 (2016), 964–994. https://doi.org/10.1007/s10618-015-0448-4

[71] Han Xu, Xiaorui Liu, Yaxin Li, Anil Jain, and Jiliang Tang. 2021. To be Robust or to be Fair: Towards Fairness in Adversarial Training. In Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139), Marina Meila and Tong Zhang (Eds.). PMLR, 11492–11501.

https://proceedings.mlr.press/v139/xu21b.html

[72] Bernard P Zeigler, Tag Gon Kim, and Herbert Praehofer. 2000. Theory of modeling and simulation. Academic press.