

Closed-Loop View of the Regulation of AI: Equal Impact across Repeated Interactions

Quan Zhou*, Ramen Ghosh[†], Robert Shorten*, and Jakub Mareček[‡]

* Imperial College London

[†] Atlantic Technological University

[‡] Czech Technical University in Prague

Abstract—There has been much recent interest in the regulation of AI. We argue for a view based on civil-rights legislation, built on the notions of equal treatment and equal impact. In a closed-loop view of the AI system and its users, the equal treatment concerns one pass through the loop. Equal impact, in our view, concerns the long-run average behaviour across repeated interactions. In order to establish the existence of the average and its properties, one needs to study the ergodic properties of the closed-loop and, in particular, its unique stationary measure.

I. INTRODUCTION

There has been considerable interest in the regulation of artificial intelligence (AI), recently. It is increasingly recognised that so-called high-risk applications of AI, such as in human resources, retail banking, or within public schools, be it admissions or assessment, cannot be served by black-box AI systems with no human control [Bringas Colmenarejo et al., 2022], predominantly due to concerns for protected human rights. A great many reports and research have revealed the danger of AI systems violating fairness in predicting which areas need patrolling [Courtland, 2018], criminal-risk assessment [Angwin et al., 2016], discriminatory behavior in advertising and recruiting algorithms for people with disabilities [Nugent and Scott-Parker, 2021], [Guo et al., 2020], search engine reinforcing racism [Noble, 2018]; and the threat of breaching privacy [Nguyen et al., 2021], [Sun et al., 2020]. To cope with the challenges of AI, leading technology companies have issued AI principles of their own and developed software tools geared towards fairness and explainability of AI, such as AIF360 [Bellamy et al., 2018] of IBM, SHAP [Lundberg and Lee, 2017] of Microsoft. In a broader context, it is not clear [Dobbe et al., 2021], however, how to phrase even the desiderata for the regulation of AI.

Here, we suggest that the desiderata could be the same as in the Civil Rights Act of 1964 and much of the subsequent civil-right legislation world-wide: equal treatment and equal impact. At the same time, we point out that these desiderata could be in conflict [Binns, 2020], [Zhao and Gordon, 2019]. The Ricci v. DeStefano, 557 U.S. 557 (2009) labour law case has demonstrated the practical differences between them, where the city of New Haven has declined to promote city firefighters based on the same test, which, shows a disproportionate pass rate for a certain race, as to the fear of valiating Title VII of the Civil Right Act of 1964 [McGinley, 2011]. The use of the

same test conducts the principle of equal treatment, while the disparate pass rates and possibly contrasting promotion results do not comply with equal impact.

Let us illustrate the conflict with another example of a system that performs credit-risk estimation in a consumer-credit company. In the US, this is regulated by the Equal Credit Opportunity Act of 1974, but the example applies equally well to other countries. Imagine a situation where the credit decision is uniform: everyone who has not defaulted on any loan is approved a credit up to \$50000. Anyone else is declined credit. This is clearly the most “equal treatment” possible, in the spirit of non-discrimination “on the basis of race, color, religion, national origin, sex, marital status, age, receipt of public assistance”, as mandated by the Equal Credit Opportunity Act. At the same time, if one subgroup (defined by whichever protected attribute, e.g., race or the receipt of public assistance) is having a lower-than-average income, its default rate on the \$50000 loan may be higher than that of the other subgroups. Over time, the subgroup with lower-than-average income will be regularly declined credit as a result of these defaults, in violation of the “equal impact”. On the other hand, if the credit limit is, e.g., set at three times the annual salary, the subgroup with lower-than-average income will be offered lower credit limits, in violation of the “equal treatment”. The differentiated credit limits may make it possible for the same subgroup to repay the loans successfully, though, to develop a credit history, and eventually lead to a positive and “equal impact”.¹ See the penultimate section of this paper for further details of the application.

Our original contribution then stems from the reinterpreting of the meaning of equal treatment and equal impact within a closed-loop view of the AI system. There, an AI system produces information, which is communicated to the users, who respond to the information. The aggregate actions of the users are observed and serve as an input to further uses of the AI system. Equal treatment concerns a single run of this closed-loop, while equal impact concerns long-run properties of this closed-loop.

¹While the Equal Credit Opportunity Act mandates that one must accurately describe the factors actually scored by a creditor, it does not suggest which of the above is preferable. Specifically, it says “if creditors know they must explain their decisions ... they [will] effectively be discouraged from discriminatory practices”.

The closed-loop view of the AI system addresses several important shortcomings of the presently proposed systems:

- it very clearly distinguishes equal impact from equal treatment;
- it allows for a stochastic response of the users to the information produced by the AI system, rather than assuming it is deterministic;
- it explicitly models the “concept drift” and retraining of the AI system over time, inherent in practical AI systems, but ignored by most analyses of AI systems.

In terms of technical results, we formalise the notions above, present one condition that is necessary for the equal impact of an AI system, and illustrate the notions on a credit-risk use case.

II. RELATED WORK

A. Regulation of AI

While there is a long history of research on the interface of AI and law [Bench-Capon et al., 2012], [Narayanan, 2018], [Berente et al., 2021, e.g.], much recent interest [Smuha, 2021b], [Petit and De Cooman, 2021, e.g.] has been sparked by the plans to introduce AI regulation within the legal system. By investigating the self-regulation of leading AI companies from both the USA and Europe, [de Laat, 2021] appeal for future practices and governmental regulation. Arguably, the European Commission regulates AI already: Article 22.1 of the General Data Protection Regulation (GDPR) is sometimes interpreted as prohibiting fully automated decisions with legal effect or “similarly significant effect”. There is much discussion regarding the AI Act [Veale and Borgesius, 2021] and regulatory landscape [Bringas Colmenarejo et al., 2022], [Vokinger and Gasser, 2021] in the Europe Union, and the potential extensions of the regulatory framework in the USA [Chae, 2020]. The EU Artificial Intelligence Regulation Proposal, suggests use of “feedback loops” that perform the detection of biased outputs and the repeated introduction of appropriate methods of bias mitigation.²

Within the recent discussions, a fair amount of attention focuses on the question of defining AI [Schuett et al., 2019] – or whether one should like to regulate the use of any algorithm [Schuett, 2019], [Ellul, 2022] – and defining high-risk uses of AI. One would also like to distinguish [Smuha, 2021a] between the harm of the individual and the society. Further, in high-risk applications of AI, the automated decision-making AI systems are bound to be fair while formalisation of fairness definitions has been a long-standing debate. From the prospectives of fair outcomes, group fairness, such as demographic parity [Calder et al., 2009], equal opportunity [Hardt et al., 2016], requests people from protected groups to be given the same

²Article 15 of this Proposal emphasises that “High-risk AI systems that continue to learn after being placed on the market or put into service shall be developed in such a way to ensure that possibly biased outputs due to outputs used as an input for future operations (‘feedback loops’) are duly addressed with appropriate mitigation measures.”

treatment as others, while individual fairness requests “similar people to be treated similarly” [Petersen et al., 2021], [Dwork et al., 2012]. On the other hand, casual fairness [Chiappa, 2019], [Kusner et al., 2017] asks for a fair decision process, such that protected attributes are not direct causes of decisions, or only through certain causal paths. Some recent works have extended to defining fairness in specific contexts, using users’ feedback [Wen et al., 2021], [D’Amour et al., 2020], [Awasthi et al., 2020].

In contrast, we distinguish between the treatment within a single interaction with the AI system and the impact of repeated interactions with the AI system. Further, we propose a closed-loop framework that repeatedly increases fairness, using aggregated feedback or users’ responses.

B. Control Theory

Our approach is rooted in the closed-loop view of feedback control, but with several important differences.

Classic control often focuses on regulating a single system. The system achieves the required behaviour most efficiently given the restrictions imposed by the challenge and the available resources. Even in areas where large-scale coupled systems are studied, the behaviour of all system components is analyzed and developed. On the other hand, in artificial intelligence, it is not the behaviour of individual users that is of interest. Rather than that, the variable of interest is the aggregate impact of the acts of a large number of users. Examples of this kind of analysis include demand management for shared resources such as water and electricity, and the provision of medical care. De-synchronization alleviates the supply strain, and collective effects quantify the supply’s quality. On the other hand, limits on the needed level of service for persons vary according to the application area.

Second, classical control, in general, is concerned with the control of systems with fixed dimensions. On the other hand, artificial intelligence often regulates and affects the behaviour of large-scale populations. Even the system’s dimensions may be unpredictable and variable in such settings, emphasizing the critical requirement for scale-free management of extremely large-scale systems. Except in the case of passive control design, scale-free control for large systems is a largely unexplored issue in the classical control field.

Thirdly, in classical control, the controlled system’s mathematical description does not change in response to control signals. This underlying concept is challenging to realize in artificial intelligence. By and large, models can only approximate the dynamics of the actual systems. This is not an issue as long as there is an appreciation for the possibility of reality and model deviating from one other. However, models in artificial intelligence are not easily derived from first principles; instead, they are empirical, i.e., based on data gathered from measurements of existing processes. Additionally, controlled studies cannot gather empirical data across a variety of operating points but must be obtained directly from the system.

An effort to enhance the processes above, for example, by sending information to the users involved, establishes a

feedback loop that did not exist earlier. This change in the underlying process may invalidate the empirical model since there were no data available to represent the dynamic influence of such feedback during the model’s development. Frequently, offered solutions ignore this feedback loop. This latter aspect necessitates a far more extensive examination of prediction and optimisation under feedback than has hitherto been the case.

Fourth, data sets are often gathered in a closed-loop fashion like Figure 1. That is, public data sets often contain information about decision-makers. Developing models of large-scale feedback systems is a crucial hurdle to development in applying certain control methods to artificial intelligence. In dealing with such impacts, artificial intelligence researchers may have a lot to learn from economic and control theory.

Finally, and perhaps most significantly, a fundamental distinction between classical control and our approach is the need to investigate the influence of control signals on the statistical features of the populations under control. Given that we are often dealing with service delivery, these statistical features should be stationary and predictable, necessitating ergodic control design.

C. Control of Multi-user Dynamical Systems

Perhaps the closest to our work within control theory are multi-user dynamical systems over networks. There, the principal concern is the design of distributed protocols that provide consensus or synchronisation of states of all users [Blondel et al., 2005], [Nedic and Ozdaglar, 2009]. (The states might indicate vehicle directions or locations, estimations of sensor readings in a sensor network, oscillation frequencies, and each user’s trust opinion, among other things.) To achieve synchronised behaviour in multi-user systems, all systems must agree on the values of these quantities.

Studying their interactions and collective behaviours under the effect of the information flow permitted by the communication network is critical for networked cooperative dynamical systems. This communication network may be seen as a graph with directed edges or connections corresponding to the information travelling between the systems. The systems are portrayed as nodes on the graph and are sometimes referred to as users. In communication networks, information flows exclusively between the graph’s close neighbours. However, if a network is linked, this locally sent information eventually reaches every user in the graph.

In cooperative control systems based on graphs, there are fascinating interactions between the dynamics of the individual users and the communication graph’s topology. The graph topology may severely constrain the performance of the users’ control rules. To be precise, in cooperative control on graphs, all control protocols must be distributed so that each user’s control rule is limited to knowledge about its near neighbours in the network topology. If sufficient attention is not taken while constructing the local user control rules, the dynamics of the individual users may be stable, but the graph’s networked systems may display undesired behaviours. Due to

the communication constraints imposed by graph topologies, complex and fascinating behaviours are seen in multi-user systems on graphs that are not found in single-user, centralised, or decentralised feedback control systems.

The ideas of distributed cooperative control are used in [Lewis et al., 2013] to construct optimal and adaptive control systems for multi-user dynamics on graphs. The requirement complicates these designs that all control and parameter tweaking methods must be dispersed in the network to rely on just their near neighbours.

[Lewis et al., 2013] analysed discrete-time systems and demonstrate that an additional condition between the local user dynamics and the graph topology must be met to ensure global synchronization when the local optimum design is used. Global optimization of collective group movements is more challenging than locally optimizing each user’s motion. A typical issue in optimum decentralized control is that global optimization problems often demand knowledge from all users, which distributed controllers cannot access since they can only utilize information from closest neighbours. Further, they demonstrate, globally optimum distributed form controls may not exist on a particular graph. To achieve globally optimum performance when employing distributed protocols that rely only on local user information in the graph, the global performance index must be chosen to depend on graph features, notably the graph Laplacian matrix. They also establish distinct global optimality for which distributed control solutions are always possible on sufficiently linked networks. There, they examine multi-user graphical games and demonstrate that a Nash equilibrium results when each user optimizes its local performance index. For more results on these direction we refer [Shamma, 2008], [Wang et al., 2017], [Wang et al., 2021], [Yu et al., 2017], [Chen et al., 2019].

III. A CLOSED-LOOP VIEW OF AI SYSTEMS

Let us consider a closed-loop model based on the following constraints:

- Users get information from the AI System, but are not required to take action based on the AI System’s outputs. It will be convenient to encode user’s reaction to the output probabilistically.
- The AI System does not necessarily monitor individual user’s actions (“profiling”), but rather some aggregate or otherwise filtered version.
- The users do not communicate with one another, or only in response to information broadcast by the central authority.

Ultimately, the repeated uses of an AI system can be seen as the closed-loop of Figure 1. The AI System produces some outputs $\pi(k)$ at time k , e.g., lending decisions in financial services, matches in a two-sided market, or suggestions in a decision-support system. The output is taken up by N users of the system, who have some states x_i , $i \in [N]$ internal to them, where $[N] = 1, \dots, N$.

The users take some action, which can be modelled as a probability function of the output and the private state, over

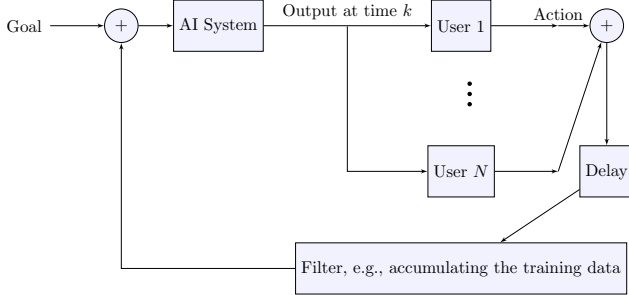


Fig. 1. A closed-loop model of an AI system and its interactions with the users: the AI system provides some outputs, e.g., scorecards in credit scoring, matches in a matching market, or suggestions in a decision-support system. Users observe the outputs and take action in response. With some delay, their actions in response to the outputs are utilized in retraining the AI System.

the certain user-specific sets of actions. The action $y_i(k)$ of user i at time k is then a random variable. In the remainder, we will assume $y_i(k)$ are scalars, but generalisations are easy to obtain. The aggregate of the actions $y(k) = \sum_{i=1}^N y_i(k)$ at time k is then also a random variable. The AI System may not have access to either $x_i(k)$, $y_i(k)$, but perhaps only $y(k)$ or some filtered version. The filter may accumulate the data, for instance, before filtering out anomalies.

IV. EQUAL TREATMENT

Equal treatment very clearly examines the AI system's treatment of its users and the influence on the microscopic qualities over the short run.

Definition 1 (Equal Treatment). For each user i , we require that

- i) the system provides the same information $\pi(k)$ to all users i ,
- ii) that there exists a constant $\bar{\pi}$ such that

$$y_i(j) = \bar{\pi}, \quad (1)$$

where this constant is independent of initial conditions.

Definition 2 (Equal Treatment Conditioned on Non-Protected Attributes). For each user i within a class that is defined by non-protected attributes, we require that

- i) the system provides the same information $\pi(k)$ to all users within the class;
- ii) that there exists a constant $\bar{\pi}$ such that

$$y_i(j) = \bar{\pi}, \quad (2)$$

where this constant is independent of initial conditions.

Notice that there is a sufficiently large overlap of the classes that are defined by non-protected attributes such that the definition reduces to the unconditional equal treatment.

V. EQUAL IMPACT

Equal impact very clearly examines the AI system's influence on the user population's microscopic qualities over the long run. One may desire, for example, that each user obtains a fair portion of the resource on average over time, or, at a far more fundamental level, that the average allocation of the resource to each user over time is a stable number that is predictable and independent of beginning circumstances.

To model equal impact, we construct requirements that ensure ergodicity: the presence of a single invariant measure to which the system is statistically drawn regardless of the starting circumstances.

Definition 3 (Equal Impact). For each user i , we require that

- i) there exists a constant \bar{r}_i such that

$$\lim_{k \rightarrow \infty} \frac{1}{k+1} \sum_{j=0}^k y_i(j) = \bar{r}_i, \quad (3)$$

where this latter limit is independent of initial conditions;

- ii) all the \bar{r}_i coincide.

Definition 4 (Equal Impact Conditioned on Non-Protected Attributes). For each user i within a class that is defined by non-protected attributes, there exists a constant \bar{r}_i such that

$$\lim_{k \rightarrow \infty} \frac{1}{k+1} \sum_{j=0}^k y_i(j) = \bar{r}_i, \quad (4)$$

where this latter limit is independent of the initial conditions. Furthermore, we require that all the \bar{r}_i coincide.

VI. GUARANTEE PROPERTIES

Proving that there is a unique invariant measure is not necessarily an easy undertaking. Even well-known AI systems do not always result in feedback systems that exhibit equal impact.

Under the assumptions of continuity of the closed-loop model, the work on *iterated function systems* [Elton, 1987], [Barnsley et al., 1989], [Diaconis and Freedman, 1999], which are a class of stochastic dynamical systems arising from the multi-user interactions, makes it possible to obtain strong stability guarantees for such stochastic systems under the assumptions of continuity of the closed-loop model. The following are shown in the work [Fioravanti et al., 2019]:

- Even if regulation is accomplished by controlling the behaviour of ensembles of users, feedback control with integral action has the potential to disrupt the closed-loop system's ergodic features. This discovery is significant because ergodic behaviour is necessary for supporting economic contracts and ensuring the existence of attributes such as fairness. Thus, from a practical standpoint, the finding is one of the system's critical features and is not only theoretically interesting.
- A few particular instances are given to demonstrate the loss of ergodicity in seemingly innocuous situations.

- For particular population types and filters, stable control action always results in ergodic behaviour. It was particularly shown for linear and non-linear systems with both real-valued and finite-set actions.
- Finally, a minor contribution was made to demonstrate how the results from the study of iterated function systems might be used in designing controllers for specific types of dynamic systems.

In this paper, we have to relax the continuity assumptions, however. Indeed, the classification problems involve discrete sets such as the “credit denied” or “credit approved”, which cannot be easily modelled with continuous functions. So in this case, stochastic, user-specific response to the feedback signal $\pi(k) \in \Pi$ can be modelled by user-specific and signal-specific probability distributions over the certain user-specific sets of actions

$$\mathbb{A}_i = \{a_1, \dots, a_L\} \subset \mathbb{R}^{n_i}, \quad (5)$$

where \mathbb{R}^{n_i} can be seen as the space of i^{th} user’s private state space x_i . Assume that the set of possible resource demands of user i is \mathbb{D}_i , where in the case that \mathbb{D}_i is finite we denote

$$\mathbb{D}_i := \{d_{i,1}, d_{i,2}, \dots, d_{i,m_i}\}. \quad (6)$$

In the general case, we assume there are $\tau_i \in \mathbb{N}$ state transition maps

$$w_{ij} : \mathbb{R}^{n_i} \rightarrow \mathbb{R}^{n_i}, j = 1, \dots, \tau_i$$

for user i and output maps

$$w'_{i\ell} : \mathbb{R}^{n_i} \rightarrow \mathbb{D}_i, \ell = 1, \dots, \kappa_i, \kappa_i \in \mathbb{N},$$

for each user i . The evolution of the states and the corresponding demands then satisfy:

$$x_i(k+1) = w_{ij}(x_i(k)) \mid j = 1, \dots, \tau_i, \quad (7a)$$

$$y_i(k) = w'_{i\ell}(x_i(k)) \mid \ell = 1, \dots, \kappa_i, \quad (7b)$$

where the choice of user i ’s response at time k is governed by probability functions

$$p_{ij} : \Pi \rightarrow [0, 1], j = 1, \dots, \tau_i \quad (8a)$$

$$p'_{i\ell} : \Pi \rightarrow [0, 1], \ell = 1, \dots, \kappa_i, \quad (8b)$$

respectively. Specifically, for each user i , for all $k \in \mathbb{N}$ and for all signal π we have that:

$$\mathbb{P}(x_i(k+1) = w_{ij}(x_i(k))) = p_{ij}(\pi(k)), \quad (9a)$$

$$\mathbb{P}(y_i(k) = w'_{i\ell}(x_i(k))) = p'_{i\ell}(\pi(k)), \quad (9b)$$

$$\sum_{j=1}^{\tau_i} p_{ij}(\pi) = \sum_{\ell=1}^{\kappa_i} p'_{i\ell}(\pi) = 1. \quad (9c)$$

Then, one can prove that when the graph $G = (X, E)$ is strongly connected, there exists an invariant measure for the feedback loop. If in addition, the adjacency matrix of the graph is primitive, then the invariant measure is attractive and the system is uniquely ergodic.

For linear systems, this is a direct consequence of (Werner, 2004) and the observation that the necessary contractivity

properties follow from the internal asymptotic stability of controller and filter. For non-linear systems, similar results can be obtained using [Marecek et al., pear, Theorem 2]. See also [Ghosh et al., 2021] and the Supplementary information.

VII. NUMERICAL ILLUSTRATIONS

Credit scoring refers to the process of lenders, usually financial institutions, measuring the creditworthiness of a person or a small business, usually derived from its historical default. In USA, Equal Credit Opportunity Act (ECOA) and the part of the law that defines its authority and scope, known as Regulation B, require statements of specific reasons for adverse credit decisions, where it would be difficult, yet impossible to comply if complex algorithms or “black-box” models are used. Instead, scorecards are commonly adopted in practice, due to their good explainability, while alternatively, counterfactual explanations [Dutta et al., 2022], [Verma et al., 2020] work as an explainer of “black-box” models to guide an applicant on the easiest improvement that could change the model outcome. Table I displays a simple scorecard.

Factor	Code	Description	Score
History	-	× Average Default Rate	-8.17
Income	0	≤ \$15K	0
	1	> \$15K	+5.77

TABLE I

A SIMPLE SCORECARD FOR EXISTING USERS. FOR EXAMPLE, A USER WITH ANNUAL INCOME \$50K AND AN AVERAGE DEFAULT RATE 0.1 WOULD BE GIVEN A SCORE OF $-8.17 \times 0.1 + 5.77 = 4.953$.

Although Table I might seem fair at first sight, income is a factor closely related to protected attributes, e.g., race. Figure 2 displays the 2020 annual income distribution of households by race, including “BLACK ALONE” (blue), “WHITE ALONE” (pink) and “ASIAN ALONE” (green), in the USA, sourced from *Table A-2. Households by Total Money Income, Race, and Hispanic Origin of Householder: 1967 to 2020* (Table A-2), from US Census Bureau³. The green bar on the index “over 200” implied that a larger share (almost 20%) of “ASIAN ALONE” households makes more than \$200K in 2020. On the other hand, the income of most “BLACK ALONE” households is less than \$75K. This figure casts doubt on the equal treatment using the scorecard in Table I, because races with generally lower incomes would receive a lower credit score. If a lender tries to maintain similar credit distributions across different races, the results may not be as expected in the long run, as low-income households might end up defaulting or even not be able to apply for another mortgage ever after, thus hurting their long-term credit history.

Our notion of equal impact in the context of credit scoring would equalise the long-term average default rate across races or across individuals, such that low-income households can keep better credit history. Recall Figure 1 from the perspective of credit scoring. Given the goal of equal impact, at each time step, the income $z_i(k)$ is internal to the user (user),

³See <https://www.census.gov/data/tables/2021/demo/income-poverty/p60-273.html>

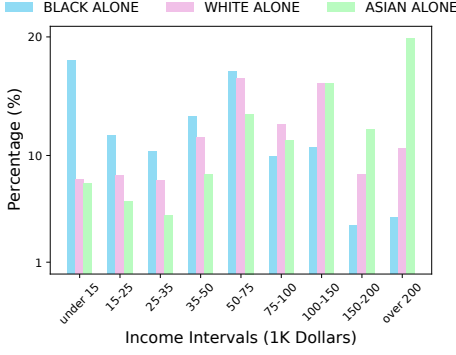


Fig. 2. The 2020 annual income distribution of “BLACK ALONE”, “WHITE ALONE” and “ASIAN ALONE” households in USA, with three races distinguished by colours. Data are sourced from Table A-2 of the Current Population Survey (CPS) of US Census Bureau.

while her income code $\mathbb{1}_{\geq 15} z_i(k)$ is visible to a lender, where $\mathbb{1}_{\geq 15}$ is an indicator function that maps the input to one if ≥ 15 is satisfied and all other values to zero. The lender would use the AI system, i.e., logistic regression in our case, to build a scorecard and reveal a credit decision $\pi(k, i)$ (e.g., approval or denial of a mortgage transaction) to user i at time k . Note that the scorecard only gives a credit score, but, based on a cut-off score, the lender is able to reach a credit decision. Confidential to client i , her state $x_i(k)$ at time k is determined by her income and, in turn, influences the repayment action. Its repayment action $y_i(k) \in \{0, 1\}$ is modelled as a Gaussian conditional independence model [Tang et al., 2021], [Leitao and Ortiz-Gracia, 2020], [Rutkowski and Tarca, 2015]. Afterwards, the filter calculates the average default rates of each user, using historical repayment actions $y_i(k)$ for $i \in 0, \dots, k$. The average default rates, along with the income code of users, would be used as training data for the AI system, and further, new credit decisions $\pi(k+1, i), i \in [N]$ are made again using logistic regression.

For the numerical experiments, we use the real-world data from Table A-2, which gives the number of households and income distribution by year and race. We consider a period from 2002 to 2020, with a year being a time step, because in 2002 the Annual Social and Economic Supplement (ASEC) of the Current Population Survey (CPS) started to allow households to report their race from more diverse options. Let \mathcal{S} be a set that includes 3 races: “BLACK ALONE”, “WHITE ALONE” and “ASIAN ALONE”. In the beginning of 2002 (time 0), we generate $N = 1000$ users (households), whose races are sampled from \mathcal{S} with a distribution of $[0.1235, 0.8406, 0.0359]$. Notice that the distribution is the ratio of the number of households of the three races in 2002 in Table A-2. The generated user set is then divided into 3 subsets according to race, denoted by \mathcal{N}_s , for $s \in \mathcal{S}$. Further, following the income distribution of the year $2002 + k$ and race s , we sample the income $z_i(k)$ of user $i \in \mathcal{N}_s$ at time k .

For simplicity, let $\pi_i(k, i) = 1$ denote that user i is

offered a 3.5-times-income mortgage at time k . Assuming that the annual mortgage rate and the basic living cost are 2.16% per annum and \$10K, we use the Gaussian conditional independence model [Rutkowski and Tarca, 2015] to generate the repayment actions. Suppose that the state $x_i(k)$ measures the portion of income left after deduction of living cost and mortgage interest:

$$x_i(k) = \frac{z_i(k) - 10 - 3.5 \times 2.16\% \times z_i(k)}{z_i(k)}. \quad (10)$$

The binary repayment action $y_i(k)$ (1 for repaid) is defined by (11).

$$\begin{cases} y_i(k) = 0 & \text{for } x_i(k) \leq 0 \\ & \text{or } \pi(k, i) = 0, \\ y_i(k) \sim \text{Bernoulli}(F(5 \times x_i(k))) & \text{otherwise,} \end{cases} \quad (11)$$

where user i would not make a repayment if no mortgage is offered or if her income cannot cover the basic living cost plus mortgage interest. Otherwise, the repayment action follows a Bernoulli distribution with $\Pr(y_i(k) = 1) = F(5 \times x_i(k))$, where $F(\cdot)$ is the cumulative distribution function of the standard normal distribution.

Furthermore, we define default as a mortgage offered but not repaid, i.e., $y_i(k) = 0 | \pi(k, i) = 1$. We introduce the average default rate $\text{ADR}_i(k)$ for user i and the race-wise version $\text{ADR}_s(k)$ for race s at time k , as defined in (12):

$$\begin{aligned} \text{ADR}_i(k) &:= \Pr(y_i(k) = 0 | \pi(k, i) = 1) \\ &= 1 - \sum_{j=0}^k \frac{y_i(j)}{\pi(k, i)} \\ \text{ADR}_s(k) &:= \Pr(y_i(k) = 0 | \pi(k, i) = 1, i \in \mathcal{N}_s) \\ &= 1 - \frac{1}{|\mathcal{N}_s|} \sum_{i \in \mathcal{N}_s} \sum_{j=0}^k \frac{y_i(j)}{\pi(k, i)}, \end{aligned} \quad (12)$$

where $|\mathcal{N}_s|$ denotes the number of users of race s . With the goal of equal impact, we wish to equalise the outcome of credit scoring among individuals in the long run, such that

$$\lim_{k \rightarrow \infty} \text{ADR}_i(k) = \bar{r}_i, \quad \lim_{k \rightarrow \infty} \text{ADR}_s(k) = \bar{r}_s, \quad (13)$$

and that all \bar{r}_i coincide and all \bar{r}_s coincide.

For the year of 2002-2003 (time 0 & 1), no scorecard is used and we assume all users are given the approval of the mortgage, e.g., $\pi(k, i) := 1$, for $i \in [N]$ and $k = \{0, 1\}$. Thus, we obtain the initialization of average default rates, i.e., $\text{ADR}_i(0), \text{ADR}_s(0)$ and $\text{ADR}_i(1), \text{ADR}_s(1)$. Afterwards, for time $k \geq 2$, a scorecard is built, whose parameters are trained from a logistic model, with independent variables being $\mathbb{1}_{\geq 15} z_i(k)$, $\text{ADR}_i(k-1)$ and the dependent variable being $\ln \frac{y_i(k)}{1-y_i(k)}$. Although, the scorecard $\pi(k)$ can vary in time steps, we use the same cut-off score 0.4 to decide each user’s credit decision (0 for denial and 1 for approval). Using our notation, the example of Table I would be rewritten as

$$\begin{aligned} -8.17 \times \text{ADR}_i(k-1) + 5.77 \times \mathbb{1}_{\geq 15} z_i(k) &= 4.953 \\ \Rightarrow 4.953 > 0.4 &\Rightarrow \pi(k, i) = 1. \end{aligned}$$

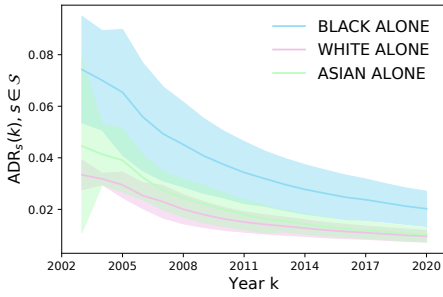


Fig. 3. Solid curves depict the mean value of time series $\{ADR_s(k)\}_{k \in [N]}$, across five trials, with race information distinguished by colour. Error shades display mean \pm one standard deviation.

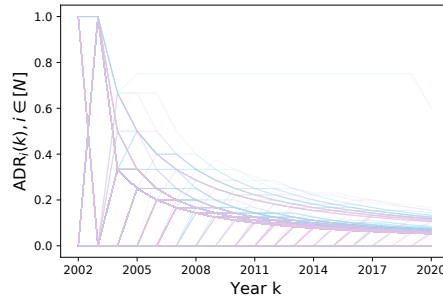


Fig. 4. The time series $\{ADR_i(k)\}_{k \in [N]}$ for all users from five trials (5×1000 curves), with their race information distinguished by colour.

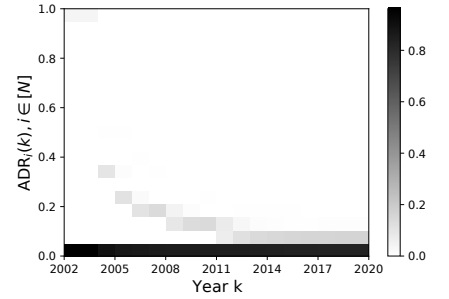


Fig. 5. The density of $ADR_i(k)$ at different time steps, with the race information ignored. Darker colours denote higher density.

We define a trial as the simulation of generating 1000 users ($N = 1000$) and repeating the closed-loop for the period 2002-2020. In our numerical experiments, five trials are conducted, with each trial using a new batch of 1000 users. For consistency with Figure 2, the races “BLACK ALONE”, “WHITE ALONE”, and “ASIAN ALONE” are represented by blue, pink, and green colours, respectively.

In Figure 3, we show the race-wise performance in five trials. Given a certain race s , the sequence of $\{ADR_s(k)\}_{k \in [N]}$ for one trial forms a time series. Across all five trials, the mean value and \pm one standard deviation could be calculated from the five time series. We denote the mean value of the time series across five trials by a solid curve and \pm one standard deviation by error shades, with the corresponding race distinguished by colour.

In Figures 4 and 5, we show the user-wise performance in five trials. Similarly, given a certain user i , the sequence of $\{ADR_i(k)\}_{k \in [N]}$ for one trial is a time series. From the five trials, and all users in $[N]$, 1000×5 time series. In Figure 4, the 1000×5 time series are visualised directly, with their races distinguished by colours. In Figure 5, the race information of the users are erased, as we intend to present the distribution of the 1000×5 time series by grey shades. Note that darker shades denote higher density of $ADR_i(k)$ at the certain time step.

Recalling the goal of equal impact in (13), we would like to see these time series converge (weakly to the same distribution). From Figure 3-5, we do observe that all time series, aggregated by race or not, are dwindling to a similar level.

VIII. CONCLUSIONS

We have presented a novel, closed-loop view of the impact of AI systems. On the example in consumer-credit approvals, we showcase, that equal impact is possible while preserving equal treatment conditional on a non-protected attribute of income. An important question for further work is how to impose constraints on the equality of impact [Celis et al., 2019]. Another important question asks whether the coupling arguments of Hairer et al. [Hairer et al., 2011] could make it possible to

show certain contrapositive statements, suggesting when such guarantees are impossible to provide.

REFERENCES

- [Angeli, 2002] Angeli, D. (2002). A Lyapunov approach to incremental stability properties. *IEEE Transactions on Automatic Control*, 47(3):410–421.
- [Angwin et al., 2016] Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016). Machine bias. In *Ethics of Data and Analytics*, pages 254–264. Auerbach Publications.
- [Awasthi et al., 2020] Awasthi, P., Cortes, C., Mansour, Y., and Mohri, M. (2020). Beyond individual and group fairness. *arXiv preprint arXiv:2008.09490*, abs/2008.09490.
- [Barnsley et al., 1989] Barnsley, M. F., Elton, J. H., and Hardin, D. P. (1989). Recurrent iterated function systems. *Constructive approximation*, 5(1):3–31.
- [Bellamy et al., 2018] Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., and Zhang, Y. (2018). AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias.
- [Bench-Capon et al., 2012] Bench-Capon, T., Araszkievicz, M., Ashley, K., Atkinson, K., Bex, F., Borges, F., Bourcier, D., Bourguine, P., Conrad, J. G., Francesconi, E., et al. (2012). A history of ai and law in 50 papers: 25 years of the international conference on ai and law. *Artificial Intelligence and Law*, 20(3):215–319.
- [Berente et al., 2021] Berente, N., Gu, B., Recker, J., and Santhanam, R. (2021). Managing artificial intelligence. *MIS quarterly*, 45(3):1433–1450.
- [Binns, 2020] Binns, R. (2020). On the apparent conflict between individual and group fairness. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 514–524.
- [Blondel et al., 2005] Blondel, V. D., Hendrickx, J. M., Olshevsky, A., and Tsitsiklis, J. N. (2005). Convergence in multiagent coordination, consensus, and flocking. In *Proceedings of the 44th IEEE Conference on Decision and Control*, pages 2996–3000. IEEE.
- [Bringas Colmenarejo et al., 2022] Bringas Colmenarejo, A., Nannini, L., Rieger, A., Scott, K. M., Zhao, X., Patro, G. K., Kasneci, G., and Kinder-Kurlanda, K. (2022). Fairness in agreement with european values: An interdisciplinary perspective on ai regulation. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 107–118.
- [Calder et al., 2009] Calder, B. J., Malthouse, E. C., and Schaedel, U. (2009). An experimental study of the relationship between online engagement and advertising effectiveness. *Journal of interactive marketing*, 23(4):321–331.
- [Celis et al., 2019] Celis, L. E., Huang, L., Keswani, V., and Vishnoi, N. K. (2019). Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 319–328.
- [Chae, 2020] Chae, Y. (2020). Us ai regulation guide: legislative overview and practical considerations. *The Journal of Robotics, Artificial Intelligence & Law*, 3.

- [Chen et al., 2019] Chen, F., Ren, W., et al. (2019). On the control of multi-agent systems: A survey. *Foundations and Trends® in Systems and Control*, 6(4):339–499.
- [Chiappa, 2019] Chiappa, S. (2019). Path-specific counterfactual fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7801–7808.
- [Courtland, 2018] Courtland, R. (2018). The bias detectives. *Nature*, 558(7710):357–360.
- [D’Amour et al., 2020] D’Amour, A., Srinivasan, H., Atwood, J., Baljekar, P., Sculley, D., and Halpern, Y. (2020). Fairness is not static: deeper understanding of long term fairness via simulation studies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 525–534.
- [de Laat, 2021] de Laat, P. B. (2021). Companies committed to responsible ai: From principles towards implementation and regulation? *Philosophy & technology*, 34(4):1135–1193.
- [Diaconis and Freedman, 1999] Diaconis, P. and Freedman, D. (1999). Iterated random functions. *SIAM Review*, 41(1):45–76.
- [Dobbe et al., 2021] Dobbe, R., Gilbert, T. K., and Mintz, Y. (2021). Hard choices in artificial intelligence. *Artificial Intelligence*, 300:103555.
- [Dutta et al., 2022] Dutta, S., Long, J., Mishra, S., Tilli, C., and Magazzeni, D. (2022). Robust counterfactual explanations for tree-based ensembles. In *International Conference on Machine Learning*, pages 5742–5756. PMLR.
- [Dwork et al., 2012] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226.
- [Ellul, 2022] Ellul, J. (2022). Should we regulate artificial intelligence or some uses of software? *Discover Artificial Intelligence*, 2(1):1–6.
- [Elton, 1987] Elton, J. H. (1987). An ergodic theorem for iterated maps. *Ergodic Theory and Dynamical Systems*, 7(04):481–488.
- [Fioravanti et al., 2019] Fioravanti, A. R., Marecek, J., Shorten, R. N., Souza, M., and Wirth, F. (2019). On the ergodic control of ensembles. *Automatica*, 108:108483.
- [Ghosh et al., 2021] Ghosh, R., Kungurtsev, V., Marecek, J., and Shorten, R. N. (2021). On the ergodic control of ensembles in the presence of non-linear filters. *arXiv preprint arXiv:2112.06767*.
- [Guo et al., 2020] Guo, A., Kamar, E., Vaughan, J. W., Wallach, H., and Morris, M. R. (2020). Toward fairness in ai for people with disabilities sbg@ a research roadmap. *ACM SIGACCESS Accessibility and Computing*, (125):1–1.
- [Hairer et al., 2011] Hairer, M., Mattingly, J. C., and Scheutzow, M. (2011). Asymptotic coupling and a general form of Harris’ theorem with applications to stochastic delay equations. *Probability theory and related fields*, 149(1-2):223–259.
- [Hardt et al., 2016] Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323.
- [Kusner et al., 2017] Kusner, M. J., Loftus, J., Russell, C., and Silva, R. (2017). Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pages 4066–4076.
- [Leitao and Ortiz-Gracia, 2020] Leitao, Á. and Ortiz-Gracia, L. (2020). Model-free computation of risk contributions in credit portfolios. *Applied Mathematics and Computation*, 382:125351.
- [Lewis et al., 2013] Lewis, F. L., Zhang, H., Hengster-Movric, K., and Das, A. (2013). *Cooperative control of multi-agent systems: optimal and adaptive design approaches*. Springer Science & Business Media.
- [Lundberg and Lee, 2017] Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc.
- [Marecek et al., pear] Marecek, J., Roubalik, M., Ghosh, R., Shorten, R. N., and Wirth, F. (to appear). Predictability and fairness in load aggregation and operations of virtual power plants. *Automatica*. arXiv preprint arXiv:2110.03001.
- [McGinley, 2011] McGinley, A. C. (2011). Ricci v. destefano: Diluting disparate impact and redefining disparate treatment. *Nev. LJ*, 12:626.
- [Narayanan, 2018] Narayanan, A. (2018). Translation tutorial: 21 fairness definitions and their politics. In *Proc. Conf. Fairness Accountability Transp., New York, USA*, volume 1170, page 3.
- [Nedic and Ozdaglar, 2009] Nedic, A. and Ozdaglar, A. (2009). Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 54(1):48–61.
- [Nguyen et al., 2021] Nguyen, V.-L., Lin, P.-C., Cheng, B.-C., Hwang, R.-H., and Lin, Y.-D. (2021). Security and privacy for 6g: A survey on prospective technologies and challenges. *IEEE Communications Surveys & Tutorials*, 23(4):2384–2428.
- [Noble, 2018] Noble, S. U. (2018). Algorithms of oppression. In *Algorithms of Oppression*. New York University Press.
- [Nugent and Scott-Parker, 2021] Nugent, S. and Scott-Parker, S. (2021). Recruitment ai has a disability problem: anticipating and mitigating unfair automated hiring decisions.
- [Petersen et al., 2021] Petersen, F., Mukherjee, D., Sun, Y., and Yurochkin, M. (2021). Post-processing for individual fairness. *Advances in Neural Information Processing Systems*, 34:25944–25955.
- [Petit and De Cooman, 2021] Petit, N. and De Cooman, J. (2021). *Models of Law and Regulation for AI*. Routledge.
- [Rutkowski and Tarca, 2015] Rutkowski, M. and Tarca, S. (2015). Regulatory capital modeling for credit risk. *International Journal of Theoretical and Applied Finance*, 18(05):1550034.
- [Schuett, 2019] Schuett, J. (2019). Defining the scope of ai regulations. *arXiv preprint arXiv:1909.01095*.
- [Schuett et al., 2019] Schuett, J. et al. (2019). A legal definition of ai. *arXiv preprint arXiv:1909.01095*.
- [Shamma, 2008] Shamma, J. (2008). *Cooperative control of distributed multi-agent systems*. John Wiley & Sons.
- [Smuha, 2021a] Smuha, N. A. (2021a). Beyond the individual: governing AI’s societal harm *Internet Policy Review*, 10(3).
- [Smuha, 2021b] Smuha, N. A. (2021b). From a ‘race to AI’ to a ‘race to AI regulation’: regulatory competition for artificial intelligence. *Law, Innovation and Technology*, 13(1):57–84.
- [Sun et al., 2020] Sun, Y., Liu, J., Wang, J., Cao, Y., and Kato, N. (2020). When machine learning meets privacy in 6g: A survey. *IEEE Communications Surveys & Tutorials*, 22(4):2694–2724.
- [Tang et al., 2021] Tang, H., Pal, A., Wang, T.-Y., Qiao, L.-F., Gao, J., and Jin, X.-M. (2021). Quantum computation for pricing the collateralized debt obligations. *Quantum Engineering*, 3(4):e84.
- [Veale and Borgesius, 2021] Veale, M. and Borgesius, F. Z. (2021). Demystifying the Draft EU Artificial Intelligence Act-Analysing the good, the bad, and the unclear elements of the proposed approach. *Computer Law Review International*, 22(4):97–112.
- [Verma et al., 2020] Verma, S., Dickerson, J., and Hines, K. (2020). Counterfactual explanations for machine learning: A review. *arXiv preprint arXiv:2010.10596*.
- [Vokinger and Gasser, 2021] Vokinger, K. N. and Gasser, U. (2021). Regulating ai in medicine in the united states and europe. *Nature machine intelligence*, 3(9):738–739.
- [Wang et al., 2021] Wang, C., Zuo, Z., Wang, J., and Ding, Z. (2021). *Robust Cooperative Control of Multi-Agent Systems: A Prediction and Observation Prospective*. CRC Press.
- [Wang et al., 2017] Wang, Y., Garcia, E., Casbeer, D., and Zhang, F. (2017). Cooperative control of multi-agent systems: Theory and applications. .
- [Wen et al., 2021] Wen, M., Bastani, O., and Topcu, U. (2021). Algorithms for fairness in sequential decision making. In *International Conference on Artificial Intelligence and Statistics*, pages 1144–1152. PMLR.
- [Werner, 2004] Werner, I. (2004). Ergodic theorem for contractive markov systems. *Nonlinearity*, 17(6):2303.
- [Yu et al., 2017] Yu, W., Wen, G., Chen, G., and Cao, J. (2017). *Distributed cooperative control of multi-agent systems*. John Wiley & Sons.
- [Zhao and Gordon, 2019] Zhao, H. and Gordon, G. (2019). Inherent trade-offs in learning fair representations. *Advances in neural information processing systems*, 32.

APPENDIX

A Markov system (see Figure 6) is a family $(X_{i(e)}, w_e, p_e)_{e \in E}$ where E consisting of edges of a finite directed (multi) graph (V, E, i, t) with $V = \{1, 2, \dots, N\}$ are vertices and $N = 1$ is also possible, $i : E \rightarrow V$ indicates the initial vertex of each edge and $t : E \rightarrow V$ indicates the terminal vertex of each edge, X_1, \dots, X_N is a partition of the metric space (X, d) into non-empty Borel subsets, $(w_e)_{e \in E}$ is a family of Borel-measurable maps on the metric space such that

$$w(X_{i(e)}) \subseteq X_{t(e)} \text{ for all } e \in E,$$

and $(p_e)_{e \in E}$ is a family of Borel measurable maps on X with the property $p_e(x) \geq 0$ for all $e \in E$ and $\sum_{e \in E} p_e(x) = 1$ for all $x \in X$. A Markov system is called irreducible or aperiodic if its directed graph is irreducible or aperiodic. A Markov system is called contractive with contraction factor a if its probability functions satisfy the following average contractivity condition, for all $x, y \in X_i, i = 1, 2, \dots, N$,

$$\sum_{e \in E} p_e(x) d(w_e(x), w_e(y)) \leq ad(x, y).$$

The Markov system defined above determines a Markov operator P on the space of bounded Borel measurable functions on X , which is denoted by $\mathcal{L}^0(X)$,

$$Pf(x) = \sum_{e \in E} p_e f \circ w_e \text{ for all } f \in \mathcal{L}^0(X),$$

and the adjoint of P is denoted by P^* acts on the space of Borel probability measures $\mathcal{M}_p(X)$ as

$$P^*\nu(f) = \int P(f) d\nu \text{ for all } \nu \in \mathcal{M}_p(X).$$

A Borel probability measure μ is said to be an invariant probability measure for the Markov system if it is a stationary distribution of the associated Markov process i.e.

$$P^*\mu = \mu.$$

A Borel probability measure μ is called attractive for the contractive Markov system iff

$$\lim_{n \rightarrow \infty} (P^*)^n \nu \rightarrow \mu \text{ for all } \nu \in \mathcal{M}_p(X).$$

Incremental stability is a well-established concept to describe the asymptotic property of differences between any two solutions. One can utilise the concept of incremental input-to-state stability, which is defined as follows:

Definition 5. A function $\gamma : \mathbb{R}^+ \rightarrow \mathbb{R}^+$ is said to be of class \mathcal{K} if it is continuous, increasing and $\gamma(0) = 0$. It is of class \mathcal{K}_∞ if, in addition, it is proper, i.e., unbounded.

Definition 6. A continuous function $\beta : \mathbb{R}^+ \times \mathbb{R}^+ \rightarrow \mathbb{R}^+$ is said to be of class \mathcal{KL} , if for all fixed t the function $\beta(\cdot, t)$ is of class \mathcal{K} and for all fixed s , the function $\beta(s, \cdot)$ is non-increasing and tends to zero as $t \rightarrow \infty$.

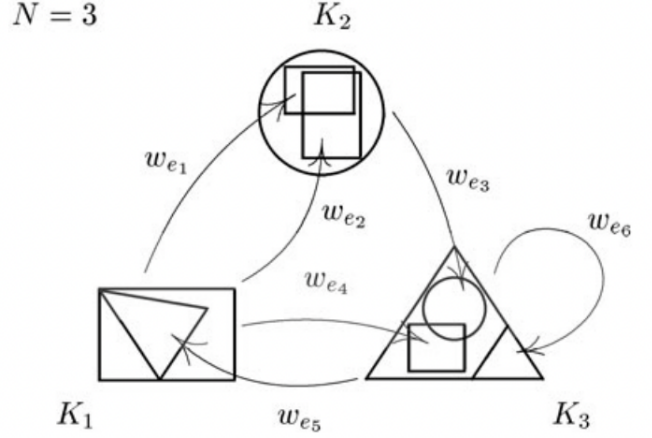


Fig. 6. A Markov system [Werner, 2004]

Definition 7 (Incremental ISS, [Angeli, 2002]). Let \mathcal{U} denote the set of all input functions $u : \mathbb{Z}_{\geq k_0} \rightarrow \mathbb{R}^d$. Suppose $F : \mathbb{R}^d \times \mathbb{R}^n \rightarrow \mathbb{R}^n$ is continuous, then the discrete-time non-linear dynamical system

$$x(k+1) = F(x(k), u(k)), \quad (14)$$

is called (globally) *incrementally input-to-state-stable* (incrementally ISS), if there exist $\beta \in \mathcal{KL}$ and $\gamma \in \mathcal{K}$ such that for any pair of inputs $u_1, u_2 \in \mathcal{U}$ and any pair of initial condition $\xi_1, \xi_2 \in \mathbb{R}^n$:

$$\|x(k, \xi_1, u_1) - x(k, \xi_2, u_2)\| \leq \beta(\|\xi_1 - \xi_2\|, k) + \gamma(\|u_1 - u_2\|_\infty), \quad \forall k \in \mathbb{N}.$$