Quality & Quantity
https://doi.org/10.1007/s11135-023-01755-z

Reinforcement learning for sequential decision making in population research

Nina Deliu[1,2]

Accepted: 13 September 2023

Abstract
Reinforcement learning (RL) algorithms have been long recognized as powerful tools for optimal sequential decision making. The framework is concerned with a decision maker, the agent, that learns how to behave in an unknown environment by making decisions and seeing their associated outcome. The goal of the RL agent is to infer, through repeated experience, an optimal decision-making policy, i.e., a sequence of action rules that would lead to the highest, typically long-term, expected utility. Today, a wide range of domains, from economics to education and healthcare, have embraced the use of RL to address specific problems. To illustrate, we used an RL-based algorithm to design a text-messaging system that delivers personalized real-time behavioural recommendations to promote physical activity and manage depression. Motivated by the recent call of the UNECE for government-wide actions to adapt to population ageing, in this work, we argue that the RL framework may provide a set of compelling strategies for supporting population research and informing population policies. After introducing the RL framework, we discuss its potential in three population-study applications: international migration, public health, and fertility.

Keywords Reinforcement learning · Multi-armed bandits · Optimal decision making · Population policies · Public health

1 Introduction

Decision making is a recurrent, ubiquitous activity that individuals, organizations, and governments face in their everyday life. In applied demography, the subject plays a critical role in problems such as decisions to marry, have a child, migrate, or retire (Kintner and Pol 1996). The process can be more or less rational or irrational and can be based on explicit or tacit information and beliefs (Gilboa 2009). Essentially, it can be regarded as a problem-solving activity yielding a solution deemed to be optimal or at least satisfactory.

✉ Nina Deliu
nina.deliu@uniroma1.it

[1] Dipartimento di Metodi e Modelli per l'Economia, il Territorio e la Finanza (MEMOTEF), Sapienza Università di Roma, Rome, Italy

[2] MRC - Biostatistics Unit, University of Cambridge, Cambridge, UK

Published online: 02 November 2023

As an illustrative guiding example, consider the context of international student mobility. From the point of view of a government, there is a growing interest in exploring or finding new ways to attract foreign students as a potential source of highly skilled workers. To this end, they may consider a set of interventions (such as financial support, housing, or paid work experiences) and decide *which policy* to implement to maximize the average retention rate, i.e., the fraction of incoming students who stay in the country after graduation. From the point of view of a student, the decision-making problem may be about *whether to move or not* so as to maximize their future income, for example.

Although decisions may well involve evidence-based practices (Movsisyan et al. 2021) or irrational beliefs, the scientific community widely recognizes the potential of following a *personalized data-driven* approach (see e.g., Kosorok and Laber 2019; Liu et al. 2017). This approach is also advocated in the recent EU-funded *HumMingBird* project,[1] which aims to leverage real-time information from mobile phones, social media, and remote sensing services in the context of migration problems. The idea is that decisions should be guided by the set of individual characteristics of a unit, rather than by intuition or average evidence alone. In the above example, if a government acts in a setting with limited resources and has to limit the decision to at most one option for each potential student, the task is to decide not only *which intervention* to offer, but also *whether or not* to offer it and *to which student*. Similarly, the decision of the student on whether to move or not will account for their *current state*, i.e., their individual and contextual information, such as socio-demographics and preference-related aspects.

Furthermore, unlike traditional protocols that regard the decision-making problem as a single-stage classification or prediction problem, it is now widely agreed that formulating it as a *sequential* decision problem—when appropriate—can better reflect the dynamics of individual (e.g., students) behavior and their evolving interaction. In fact, it is unrealistic to believe that governmental policies for attracting international students will have the same effect over time. Practical examples of such protocols can be found in personalized medicine (Chakraborty and Moodie 2013; Deliu and Chakraborty 2022; Tsiatis et al. 2019)—where interventions are dynamically tailored to the uniquely evolving health status of each patient—or recommender systems (Afsar et al. 2022), which adapt news, entertainment, or shopping items to individual customer preferences. Notably, these examples have extensively embraced the use of *reinforcement learning* (RL) to solve complex sequential decision-making tasks.

Reinforcement learning is an area of machine learning (ML) concerned with understanding how agents (humans, animals, machines) might learn to improve their decisions through repeated experience. More formally, it aims at identifying optimal decision rules (i.e., *policies*) in sequential decision-making problems under *uncertainty* (Bertsekas 2019; Sutton and Barto 2018). An optimal RL policy is the one that maximizes the expected long-term utility (as in classical economic models and decision theory; Fishburn 1970; Gilboa 2009), assuming this is likely to outweigh the associated short-term costs. Even if an individual may act primarily with a short-term horizon, they do understand that saving money is essential to build wealth and having a secure financial future.

The general RL framework is formalized through a continuous interaction between a *learning agent* (i.e., the decision maker) and the *environment* it belongs to and wants to learn about. At each interaction stage, the agent observes some representation of the environment's *state* or *context*, and on that basis selects an *action* or *arm*, i.e., makes a decision.

[1] https://cordis.europa.eu/article/id/444860-a-data-driven-approach-to-the-migration-challenge.

The impact of the chosen action is evaluated through a *reward* (or outcome) provided by the environment. Based on the reward received, the agent learns, by *trial-and-error*, on how to take better actions in the future to maximize the cumulative reward over time.

RL has existed for decades and has been widely studied, with tremendous theoretical achievements in efficiency, generalization, and representation. Nowadays, it is increasingly being applied in different domains such as robotics, business management, finance, and healthcare, just to name a few (Chakraborty and Moodie 2013; Charpentier et al. 2021; Mnih et al. 2015). More recently, its value has been leveraged in the mobile-health domain (Figueroa et al. 2022, 2021; Steinhubl et al. 2015). To illustrate, we used it to design a mobile-health app for delivering behavioral interventions in the form of text messages to promote physical activity in university students (Figueroa et al. 2022). The RL strategy was designed to evaluate, on a daily basis, which type of text message, and at what time, is more likely to maximize the number of steps walked the next day by a specific individual. The study is now being implemented in a clinical population with diabetes and depression (Aguilera et al. 2020).

Despite significant progress and an increase in the number of success stories, the considerations of RL-guided decision making in population research are still very limited. Compared to other ML methods, which have received much more attention within demography (see e.g., Carammia et al. 2022; Nigri et al. 2022), RL can be particularly suited to inform population policies in a context of global population changes (Billari 2022; Vollset et al. 2020) thanks to the following features:

(i) RL explicitly tackles the problem of making decisions and learning policies in an *uncertain* and *changing* environment;
(ii) RL focuses on *long-term goals*; furthermore, it is able to handle long and complex sequential decision-making tasks with sampled, delayed, non-stationary, and exhaustive outcomes;
(iii) RL can be implemented *online* and learn as new data are acquired, without requiring massive amounts of representative historical data. Note that most of ML techniques learn *offline* from a fixed dataset only;
(iv) RL approaches can easily leverage existing *population models* and simulations–such as agent-based models (Klabunde and Willekens 2016)–to extrapolate and integrate the impact on population dynamics when predicting future possibilities.

In light of this, a number of decision-making problems in demographic research (such as fertility and migration) could be formalized and studied through a decision-theoretical framework and ultimately solved with RL solutions.

This work re-echos the observation that the main models and techniques used for policy making have practical and theoretical limitations (Banha et al. 2022). In Hallsworth (2011), a number of major challenging points are discussed; notably, they point to a dominant policy-making model that is outdated and does not accurately reflect reality. Nevertheless, the vision of "modernized" policy making introduced in the *Modernising Government* White Paper in 1999 (Bullock et al. 2001; Cabinet 1999) illustrates nine features of modern policy making, which include, among others:

Forward looking: taking a long-term view, based on statistical trends and informed predictions, of the likely impact of policy;
Evaluation: building systemic evaluation of early outcomes into policy processes;
Learn lessons: learning from experience of what works and what does not;

> *Innovative and creative*: questioning established ways of dealing with things and encouraging new ideas.

A policy-making process should thus be a nonmyopic evidence-based approach that learns from experience and systematically evaluates and incorporates early outcomes into the decision making. Additionally, the process should accommodate innovative technologies and ideas.

We believe that there is ample scope to raise awareness of the RL potential in this context. In fact, RL is based on a continuous evaluation of past actions to learn how to improve future decisions to optimize a long-term goal. In this work, we argue that leveraging 1) mathematical decision-making frameworks to formulate decisions and 2) modern RL solutions into the policy-making process could support decisions and enhance their effectiveness and efficiency. This is particularly the case for processes with high degrees of uncertainty and complexity, such as those related to migration aspects, which may relate to exogenous conditions that are consequences of period circumstances.

It is aim of this work to make demographers and policy-makers more aware of the potential of innovative approaches to decision making, as well as to the interplay between policy-making and demography, recently emphasized by the *United Nations Economic Commission for Europe* (UNECE 2022):

> Population ageing has social and economic implications for which societies need to prepare and to which they need to adapt. This requires a coordinated, whole-of-government and whole-of-society effort. [...] A comprehensive situation analysis, including demographic projections, helps identify priorities and directions for ageing-related policies overall but also to identify the relevance of demographic change for different sectors.

We note that this interplay can be broadly contextualized within a number of domains–going from economics to social aspects–where population estimates play a key role. As extensively discussed (Ahn et al. 2005; Buettner 2022; United Nations Department of Economic and Social Affairs 2022), population projections are widely used in various policy-making processes and for development planning. Demographic forecasts have created awareness of population ageing and assisted numerous public policies such as changes in pension or birth control. Although many of these policies have made demographers active participants, these are mostly regarded as a part of other public goals and not as an instrument of population policy.

In this work, after a formal introduction of the RL framework in Sect. 2, our focus will be on examples that used RL, and specific RL subclasses, for learning policies in population problems. Section 3 will discuss the applicability of RL in three demographic areas: international migration, public health, and fertility. Some challenges that may limit the adoption of RL in real world will be presented in Sect. 4.

2 Overview of the RL framework

2.1 Basic ingredients

In reinforcement learning, differently from other ML methods, data are available in sequential order and learning is performed through many stages. For practicality, consider a discrete time space indexed by $t \in \mathbb{N} = \{0, 1, \dots, \}$. At each time t, the RL framework describes an interaction between an agent and an unknown environment, articulated in the following three key elements:

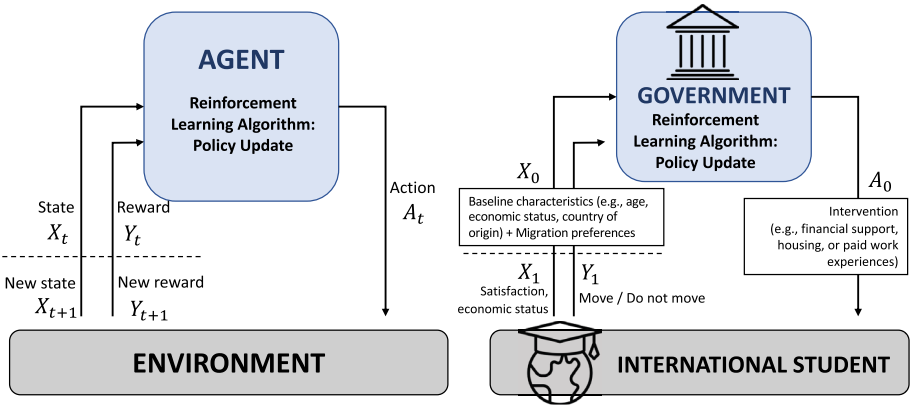

Fig. 1 Schematic of the RL framework through its sequential agent-environment interaction (left) and an illustrative one-stage example in the context of governmental policies for attracting international students (right)

- *State* or *context*, denoted by $X_t \in \mathcal{X}_t$, being the representation of the environment at step t. This includes the set of information (demographic and nondemographic covariates, such as age and country of origin) that may be relevant to make predictions about the consequences of policy alternatives.
- *Action* A_t, taken by the agent from a set of admissible actions $\mathcal{A}_t$, i.e., the set of policy alternatives. When making the choice, the agent weighs the consequences of the alternatives and their likelihood, given the state.
- A (short-term) *reward* $Y_{t+1} \in \mathcal{Y}_{t+1} \subset \mathbb{R}$ provided by the environment in response to the chosen action A_t, in correspondence with an observed state X_t. This closely relates to the concept of utility or welfare, which should be the ultimate criterion for judging whether the whole policy works well or not. In conjunction with the reward, the environment makes a transition into a new state $X_{t+1} \in \mathcal{X}_{t+1}$.

By repeating this process for each $t \in \mathbb{N}$, the result is a *trajectory* $\mathcal{T}$ of states visited, actions pursued, and rewards received:

$$\mathcal{T} \doteq \left\{(X_t, A_t, Y_{t+1})\right\}_{t\in\mathbb{N}}. \tag{1}$$

In an international migration context, this trajectory can be viewed as the *history* of contextual and individual students' information X_t, the type of financial support offered A_t, and the ultimate outcome Y_t of the student (moving or not into the host country). A schematic of the generic agent-environment interaction, along with the specific student-migration example, is reported in Fig. 1. Note that in some settings there may be multiple interventions at different stages and only one terminal reward; in this case, the rewards at all previous stages are taken as 0. In addition, in many settings, the context may also depend on the selected action, that is $X_t = X_{t,a_t}$.

2.2 Mathematical formalization of RL

Define now $\mathbf{X_t} \doteq (X_0, \dots, X_t)$, $\mathbf{A_t} \doteq (A_0, \dots, A_t)$, $\mathbf{Y_t} \doteq (Y_1, \dots, Y_t)$, and similarly $\mathbf{x_t}$, $\mathbf{a_t}$ and $\mathbf{y_t}$, where the upper- and lower-case letters denote random variables and their particular realizations, respectively. Also define the *history* $\mathbf{H_t}$ as all the information available at time t prior to decision A_t, i.e., $\mathbf{H_t} \doteq (\mathbf{A_{t-1}}, \mathbf{X_t}, \mathbf{Y_t})$; similarly $\mathbf{h_t}$. The history $\mathbf{H_t}$ at stage t belongs to the product set $\mathscr{H}_t = \mathscr{X}_0 \times \prod_{\tau=1}^{t} \mathscr{X}_\tau \times \mathscr{A}_{\tau-1} \times \mathscr{Y}_\tau$. Note that, by definition, $\mathbf{H_0} = X_0$. We assume that each longitudinal history is sampled independently according to a distribution $P_{\boldsymbol{\pi}}$, given by

$$P_{\boldsymbol{\pi}} \doteq p_0(x_0) \prod_{t \geq 0} \pi_t(a_t \mid \mathbf{h_t}) p_{t+1}(x_{t+1}, y_{t+1} \mid \mathbf{h_t}, a_t), \tag{2}$$

where:

- p_0 is the probability distribution of the initial state X_0.
- $\boldsymbol{\pi} \doteq \{\pi_t\}_{t \geq 0}$ represents the *exploration policy* and determines the sequence of actions generated throughout the decision-making process. More specifically, π_t maps histories of length t, $\mathbf{h_t}$, to a probability distribution over the action space $\mathscr{A}_t$, i.e., $\pi_t(\cdot \mid \mathbf{h_t})$. The conditioning symbol "|" in $\pi_t(\cdot \mid \mathbf{h_t})$ reminds us that the exploration policy defines a probability distribution over $\mathscr{A}_t$ for each $\mathbf{h_t} \in \mathscr{H}_t$. Sometimes, A_t is uniquely determined by the history $\mathbf{H_t}$, and the policy is simply a function of the form $\pi_t(\mathbf{h_t}) = a_t$. We call it *deterministic policy*, in contrast to *stochastic policies* that determine actions probabilistically.
- $\{p_t\}_{t \geq 1}$ are the unknown *transition probability distributions* and they completely characterize the dynamics of the environment. At each time $t \in \mathbb{N}$, the transition probability p_t assigns to each trajectory $(\mathbf{x_{t-1}}, \mathbf{a_{t-1}}, \mathbf{y_{t-1}}) = (\mathbf{h_{t-1}}, a_{t-1})$ at time $t-1$ a probability measure over $\mathscr{X}_t \times \mathscr{Y}_t$, i.e., $p_t(\cdot, \cdot \mid \mathbf{h_{t-1}}, a_{t-1})$.

At each time t, the transition probability distribution $p_{t+1}(x_{t+1}, y_{t+1} \mid \mathbf{h_t}, a_t)$ gives rise to: (i) the *state-transition probability distribution* $p_{t+1}(x_{t+1} \mid \mathbf{h_t}, a_t, y_{t+1})$, i.e., the probability of moving to state x_{t+1} conditioned on the observed history $\mathbf{h_t}$, the current selected action a_t, and the reward received y_{t+1}; and (ii) the *immediate reward distribution* $r_{t+1}(y_{t+1} \mid \mathbf{h_t}, a_t, x_{t+1})$, which specifies the reward Y_{t+1} after transitioning to x_{t+1} with action a_t. To better incorporate uncertainty, we assume a stochastic reward distribution.

The cumulative discounted sum of immediate rewards from time t onwards is known as *return*, say R_t, and is given by

$$R_t \doteq Y_{t+1} + \gamma Y_{t+2} + \gamma^2 Y_{t+3} + \cdots = \sum_{\tau \geq t} \gamma^{\tau - t} Y_{\tau+1}, \quad t \in \mathbb{N}.$$

The *discount rate* $\gamma \in [0, 1]$ determines the current value of future rewards: a reward received τ time steps in the future is worth only γ^τ times what it would be worth if it were received immediately. If $\gamma = 0$, the agent is *myopic* in being concerned only with maximizing the immediate reward, i.e., $R_t = Y_{t+1}$. If $\gamma = 1$, the return is *undiscounted* and it is well defined (finite) as long as the time horizon is finite, i.e., $T < \infty$ (Sutton and Barto 2018).

The goal in RL is to learn an optimal way of choosing the set of actions or learning an *optimal policy*, so as to maximize the expected future return. Note that the *expected* return is the most common approach to handling decisions under uncertainty (De Lara

and Doyen 2008). Thus, at any time $t \in \mathbb{N}$, the RL problem is to find an optimal policy $\boldsymbol{\pi}_t^* \doteq \{\pi_t^*\}_{\tau \geq t}$ such that

$$\boldsymbol{\pi}_t^* = \arg\max_{\boldsymbol{\pi}_t} \mathbb{E}_{P_{\boldsymbol{\pi}}}[R_t] = \arg\max_{\boldsymbol{\pi}_t} \mathbb{E}_{P_{\boldsymbol{\pi}}}\left[\sum_{\tau \geq t} \gamma^{\tau - t} Y_{\tau+1}\right], \tag{3}$$

where the expectation is meant with respect to the distribution in Eq. (2).

For estimating optimal policies, various methods have been developed so far in the RL literature: see Sutton and Barto (2018) and Sugiyama (2015) for an overview. A traditional approach is through *value functions*, which define a partial ordering over policies, with insightful information on the optimal ones. In fact, optimal policies share the same (optimal) value function. For this reason, efficient estimation of the value function is one of the most important components of almost all RL algorithms. For example, comparing estimated value functions of different candidate policies offers a way to understand which strategy may offer the greatest expected outcome.

There are two types of value functions: (i) *state-value* or simply *value* functions, say $V_t^{\boldsymbol{\pi}}$ representing how good it is for an agent to be in a given state, and (ii) *action-value* functions, say $Q_t^{\boldsymbol{\pi}}$ indicating how good it is for the agent to perform a given action in a given state. These are formally defined as:

$$V_t^{\boldsymbol{\pi}}(\mathbf{h_t}) \doteq \mathbb{E}_{P_{\boldsymbol{\pi}}}\left[R_t \mid \mathbf{H_t} = \mathbf{h_t}\right] = \mathbb{E}_{P_{\boldsymbol{\pi}}}\left[\sum_{\tau \geq t} \gamma^{\tau - t} Y_{\tau+1} \mid \mathbf{H_t} = \mathbf{h_t}\right],$$

$$Q_t^{\boldsymbol{\pi}}(\mathbf{h_t}, a_t) \doteq \mathbb{E}_{P_{\boldsymbol{\pi}}}\left[R_t \mid \mathbf{H_t} = \mathbf{h_t}, A_t = a_t\right] = \mathbb{E}_{P_{\boldsymbol{\pi}}}\left[\sum_{\tau \geq t} \gamma^{\tau - t} Y_{\tau+1} \mid \mathbf{H_t} = \mathbf{h_t}, A_t = a_t\right],$$

$\forall t \in \mathbb{N}$, $\forall \mathbf{h_t} \in \mathscr{H}_t$ and $\forall a_t \in \mathscr{A}_t$. By definition, at stage $t = 0$, $V_0^{\boldsymbol{\pi}}(\mathbf{h_0}) \doteq V_0^{\boldsymbol{\pi}}(x_0)$; while for the terminal stage, if any, the state-value function is 0.

For ease of notation, in this work, we assume that we are at the beginning of a process, that is, at stage $t = 0$, and we focus on state-value functions. The aim is to finding the policy having the greatest value of expected total-discounted reward, $\boldsymbol{\pi}^* = \arg\max_{\boldsymbol{\pi}} V_0^{\boldsymbol{\pi}}(x_0)$, where

$$V_0^{\boldsymbol{\pi}}(x_0) = \mathbb{E}_{P_{\boldsymbol{\pi}}}\left[\sum_{t \geq 0} \gamma^t Y_{t+1} \mid X_0 = x_0\right] \tag{4}$$

and $\gamma \in [0, 1]$. Note that Y_{t+1} is a function of the current history $\mathbf{h_t}$ (or state x_t) and the selected action a_t.

A fundamental property of the value functions used throughout RL is that they satisfy particular recursive relationships, known as *Bellman optimality* equations (Bellman 1957). Denoted with $V_t^{\boldsymbol{\pi}^*}$ the *optimal value function* at time t, which is the one that yields the largest expected return for each history, it holds that:

$$V_t^*(\mathbf{h_t}) = \mathbb{E}\left[Y_{t+1} + \gamma V_{t+1}^*(\mathbf{h_{t+1}}) \mid \mathbf{H_t} = \mathbf{h_t}\right].$$

This property allows for the estimation of (optimal) value functions recursively, from T backwards in time. In finite-horizon *dynamic programming* (DP), this technique is known as *backward induction*, and represents one of the main methods for solving the Bellman equation.

Offline and Online RL. In many decision problems, the *estimation policy* we want to learn about, say **d**, might be different from the *exploration policy* $\boldsymbol{\pi}$ that generated the data. This may happen when we want to estimate an optimal policy without interacting with the environment but using some already collected data (e.g., observational data), for which a certain exploration policy, often unknown, was used. We refer to it as *offline* RL, as opposed to *online* RL, where the agent interacts with the environment to collect samples and iteratively improve the policy. In practice, a learning problem where the policy is only evaluated according to the state of the system at the end of the process, even if decisions and outcomes occurred in a sequential manner, is regarded as an offline learning problem. On the contrary, a problem is said to be online if the policy is evaluated and optimized as rewards are collected and new information is gathered. The multi-armed bandit (MAB) problem is a classic example of an online RL problem.

2.3 The multi-armed bandit class

MAB problems represent some of the simplest expressions of RL problems. The classical form of the MAB problem is as follows. There are multiple actions or arms, say K, each associated with a (possibly different) reward distribution. At each time $t = 0, 1, \ldots$, the learner makes a choice among the K arms and only receives a reward for the action chosen at each time t. As information accumulates, the learner's goal in a MAB problem is to maximize the cumulative reward by trading-off between selecting the best actions so far (*exploitation*) and acquiring new knowledge about the other actions (*exploration*). Within this framework, an optimal solution to Eq. (3), can be derived by solving k 1-dimensional optimization problems instead of the k-dimensional problem as required by dynamic programming. Basically, the goal reduces to the selection of the optimal action A_t^* at each time t, with

$$A_t^* = \arg\max_{a_t \in \mathscr{A}} V_t(x_t) = \arg\max_{a_t \in \mathscr{A}} \mathbb{E}(Y_{t+1} \mid X_t = \mathbf{x}_t, A_t = a_t).$$

Based on how exploration is approached, there are several ways to build policies to solve the MAB problem. For example, one may focus on exploration and never exploit any of the data they have gathered. Pure-exploration policies are completely random; we call them *Random policies*. Pure-exploitation policies would always choose the best possible solution, assuming they already have all the data to exploit and know the underlying truth; we call this an *Oracle*. Clearly, the latter is possible only in theory, and decisions can be equally bad as random policies. Thus, the state-of-the-art MAB policies rely on an efficient balance between exploration and exploitation. We refer to Lattimore and Szepesvári (2020) for an extensive overview and analysis of the matter, while here we report two alternative strategies known as the *Gittins index* and the *Thompson sampling*.

2.3.1 The Gittins index

An alternative approach to DP for solving Eq. (3) is to associate an index to every state or stage and select the arm with the highest index at every stage t. The *Gittins index* (Gittins 1974), originally named *dynamic allocation index*, offers a solution to a very large number of problems (see Chapter 1 of Gittins et al. 2011, for an overview) and represents a key breakthrough for the MAB problem. Formally, the Gittins index theorem states that, for any infinite-horizon discounted RL problem, with finitely many arms and

bounded rewards, the policy obtained by backward induction is optimal if and only if it always selects the arm k with the highest Gittins index at each time t. In correspondence to the initial state $x_{0,k}$, the index of arm k, denoted by $\mathscr{G}_k(x_{0,k})$, is defined as:

$$\mathscr{G}_k(x_{0,k}) \doteq \sup_{\tau \geq 1} \frac{\mathbb{E}\left[\sum_{t=0}^{\tau-1} \gamma^t Y_{t+1}(X_{t,k}) \mid X_{0,k} = x_{0,k}\right]}{\mathbb{E}\left[\sum_{t=0}^{\tau-1} \gamma^t \mid x_{t,k}\right]}, \tag{5}$$

where τ is a past-measurable stopping-time. Intuitively, the index can be interpreted as the maximal value of the ratio of the ex discounted reward to the expected discounted time under policies that choose a stopping time τ.

Importantly, Eq. (5) tells us that the index depends only on the information concerning arm k, greatly reducing the dimensionality of the problem and its solution. The Gittins index policy assigns the value to each arm based on the observed state variables, and suggests as optimal strategy the one with the highest value. It can be computed offline, calculating the table of values of the index for each state and action, or online, calculating the value of the index for the current state, and the corresponding stopping time for the current state. We refer to Chakravorty and Mahajan (2014) for an overview.

2.3.2 Thompson sampling

Rooted in a Bayesian framework, the Thompson sampling algorithm guides the choice of actions in proportion to the posterior probability of producing the maximum expected reward at each time t. For a given action k, the policy π at each time $t+1$ is explicitly defined as:

$$\begin{aligned} \pi_{t+1,k} &= \mathbb{P}\Big(Q^{\boldsymbol{\pi}}_{t+1}(X_{t,k}, A_t = k) \geq Q^{\boldsymbol{\pi}}_{t+1}(X_{t,k'}, A_t = k'), \forall k' \neq k \mid \mathbf{H_t} = \mathbf{h_t}\Big) \\ &= \mathbb{P}\Big(\mathbb{E}[Y_{t+1}(X_{t,k})] \geq \mathbb{E}[Y_{t+1}(X_{t,k'})], \forall k' \neq k \mid \mathbf{H_t} = \mathbf{h_t}\Big), \end{aligned} \tag{6}$$

where the conditioning term $\mathbf{H_t} = \mathbf{h_t}$ reflects the posterior nature of this probability and should not be confused with the conditioning terms of the Q-function.

For some families of reward distributions, it is possible to compute $\pi_{t+1,k}$ either analytically or by quadrature. In any case, it may be computationally and memory intensive, thus, the typical way for implementing the Thompson sampling algorithm does not involve their direct computation, but follows a posterior sampling procedure as detailed in Agrawal and Goyal (2013).

2.3.3 A simulation-based comparison

For the sole purpose of better understanding the advantages of RL in a resource allocation problem, let us now evaluate it in a simplified simulation experiment. Consider again the example of international student mobility. Assume that the decision-making agent is the government or an educational institution, whose task is to select and offer one type of intervention to each potential incoming student. In an RL framework, the student would represent the unknown environment the agent wants to learn about. The agent may also decide for a "`no offer`" option, due, e.g., to limited resources. Specifically, let us consider the following action space:

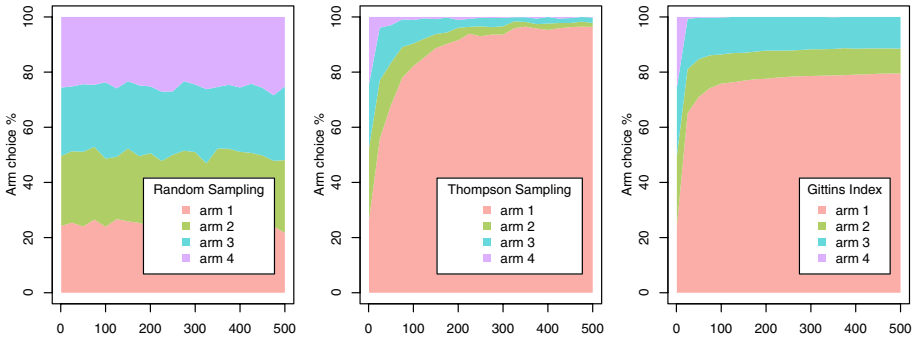

Fig. 2 Arm choice percentage of policies Random sampling (left), Thompson sampling (middle) and Gittins index (right). The Oracle, by definition, always selects the best arm(s), thus it is not shown. Results are averaged over $M = 1000$ independent simulation runs

$$\begin{aligned}\mathscr{A} = \{&\text{arm 1} = \text{“}\texttt{financial support (equivalent to the housing cost)}\text{”},\\ &\text{arm 2} = \text{“}\texttt{housing}\text{”},\\ &\text{arm 3} = \text{“}\texttt{paid work (earning greater than the housing cost)}\text{”},\\ &\text{arm 4} = \text{“}\texttt{no offer}\text{”}\}.\end{aligned}$$

In principle, the agent does not know which of the four interventions is optimal, i.e., is associated with the highest outcome and for whom. In this example, we define the outcome of interest as the binary variable $\mathscr{Y} = \{0, 1\}$, where 1 indicates a student who has decided to move to the host country after receiving the offer and 0 otherwise. We assume a stationary setting, where the outcome distribution does not change over time, and model it as a Bernoulli variable depending on the unknown success parameter p_k of each arm k:

$$Y_k \sim \text{Bernoulli}(p_k), \quad k = 1, 2, 3, 4. \tag{7}$$

Therefore, the agent's goal is to learn sufficient information about the different interventions in terms of their associated outcome or p_k, so as to assign the most promising intervention to most students. This results in maximizing the total number of international students moving to the host country.

In this example, each step t of the RL framework corresponds to the unique student involved in the international mobility process. In other examples (such as the one that will be discussed in Sect. 3.1), it may represent different time points of the same individual. In both cases, effective learning is based on observing the outcomes obtained at previous time points (e.g., of previous students or groups of students, or of the same individual at prior time steps) and using the continuously accrued information to make better decisions in the future. The outcomes depend on the intervention k and may also depend on one or more specific characteristics X of the students, e.g., gender, age, country of origin. In that case we may consider p_k to be a logistic function of the form $p_k(x_t) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_t)}}$, with x_t being the characteristic of the individual t and β_0, β_1 the unknown regression parameters. Here, for simplicity of exposition and without loss of generality, we assume $p_k(x_t) = p_k, \forall t$.

Comparative results in terms of reward and arm choice are reported in Figs. 2 and 3. These are generated on the basis of the illustrative student example model in Eq. (7) with arm success parameters taken as $p_1 = 0.8, p_2 = 0.6, p_3 = 0.6, p_4 = 0.2$. We assume

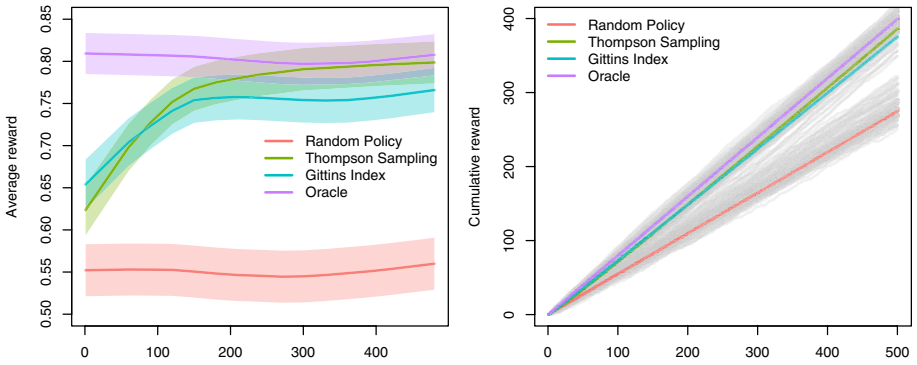

Fig. 3 Average reward with 95% confidence bound (left) and cumulative reward with individual simulation traces (right) of the compared policies. Results are averaged over $M = 1000$ independent simulation runs

that arm 1 = "`financial support`" is the optimal intervention, capable of attracting 80% of potential students, while when arm 4 = "`no offer`" is provided, the student is still open to move in 20% of cases. Arm 2 and arm 3 are assumed to be equivalent in terms of student attractiveness: even if the professional and economic return of arm 3 is greater than the one of arm 2, this comes at the cost of working hours.

We compare the two MAB policies introduced above (*Thompson sampling* and *Gittins index*, for which we use the approximation in Brezzi and Lai (2002) and a discount rate of $\gamma = 0.95$) with a *Random policy* that chooses arms uniformly at random and an *Oracle* that always selects the best intervention. Results are averaged over $M = 1000$ independent simulation runs, with a horizon of $T = 500$, equivalent to the number of students.

As shown in Fig. 2 (middle and right plot), after an initial learning phase in which the two MAB policies explore and learn about the optimal interventions, the choice of the arm quickly skews toward the optimal arm 1 = "`financial support`". It should also be noted that the arm associated with the lowest probability of success (arm 4 = "`no offer`") is quickly dropped. The result naturally translates in an increased probability of attracting international students (as shown in the reward plots in Fig. 3). While a Random policy attracts 55% of the students over time, a MAB-based policy is able to achieve performances similar to an Oracle in the long run. On average, over 500 students, the MAB policies attract between 380 and 385 students, compared to an Oracle with 400 and a Random policy with 275 students, respectively.

3 Applications in population research

In principle, most of the decision-making problems in dynamic population phenomena, ranging from birth control to migration policies, could be formalized through a model of rational decision making under an RL framework. In practice, the literature on the matter is very limited. In what follows, we offer a window to its potential by illustrating two application examples that addressed population problems with RL.

3.1 Individual migration decision making

An inevitable phenomenon of modern globalized societies is migration: a relatively permanent movement of people to new geographical locations, typically for the purpose of employment and improved living conditions. Over the past decades, international labor migration has been increasing worldwide, especially from lower income countries and in border-free systems such as Europe (Boswell 2018). There are a number of positive effects of labor migration. For the hosting country, it allows filling potential gaps in the labour market, boosting the local economy, and increasing the demographic and cultural variety, among other factors. For the country of origin, it lowers the unemployment rate, reducing job rivalry. For the immigrant, it represents a means by which individual workers can enhance their standard of living, skills, and/or career opportunities.[2] However, it can also create challenges such as pressure on public services, racial tensions, and discrimination. Therefore, the need to factor international mobility and migration into strategies and policies is increasingly recognized among nations (OECD 2017).

Notably, "interest in artificial intelligence, machine learning, predictive analytics, and automated decision making is not immune to this tendency" (Molnar and Gill 2018). To some extent, this is already a reality. For example, Canada uses algorithmic decision making in the determination of immigration and asylum (Molnar and Gill 2018); similarly, Switzerland is developing an algorithm to improve the integration of refugees to optimize their overall employment rate (Bansak et al. 2018). However, as far as our knowledge is concerned, existing strategies are not constructed on the basis of a theoretical framework for decision making where RL could be applied. Rather, they are based on a combination of supervised machine learning techniques (used for prediction and classification) and optimization. Thus, we argue that the use of RL to support migration policy making, currently underexplored, could revolutionize the way governments and international organizations seek to manage this international phenomenon. An illustration of its applicability and potential for migration decisions is now provided from the perspective of an individual (the RL agent) who has to decide whether and where to migrate (the actions).

Back in 1997, Berninghaus and Seifert-Vogt (Berninghaus and Seifert-Vogt 1987) embraced the modern paradigm of sequential decision making under uncertainty to model the *individual* migration decision making. Authors regarded the problem of international migration as a dynamic precess under incomplete information in which agents make their decisions so as to maximise their utility. Referring back to the *human capital theory* of Sjaastad (1962), migration is seen as a form of individual behavior: individuals migrate to improve their economic situation compared to staying where they are. The decision to migrate is eventually based on comparing the expected discounted future return abroad with that achievable in their home country. The discount rate reflects the degree of time preference over the time horizon. As the migrant will acquire continuously more knowledge by staying in the chosen country, they might decide to leave that country or to move to a different one in case of unexpectedly unfavourable events or rewards. That is, the migrant must be able to revise their decision at each decision time point in face of incoming information.

Consider an individual migrant who faces a decision-making problem at finitely many decision times $t = 0, 1, \dots, T$ ($T < \infty$). At each time t, the migrant has to make a choice among one of K countries, in addition to the one they are currently staying,

[2] https://www.vazirgroup.com/news/impacts-of-migration-around-the-globe/.

with the action space being $\mathscr{A} = \{0, 1, 2, \ldots, K\}$. Aligned with other migration-related works (Constant and Zimmermann 2012), authors assume a Markov decision model, meaning that agent's decisions can be entirely determined based on the last/current available information (state) only, i.e., $p_{t+1}(\cdot, \cdot \mid \mathbf{H_t}, A_t) = p_{t+1}(\cdot, \cdot \mid X_t, A_t)$, $\forall t \geq 0$. The idea is that it is only this current state, if known, that influences the future migration choice. The trajectory distribution in Eq. (2) is thus exemplified as follows:

$$P_{\boldsymbol{\pi}} \doteq p_0(x_0) \prod_{t \geq 0} \pi_t(a_t \mid x_t) p_{t+1}(x_{t+1} \mid x_t, a_t) r_{t+1}(y_{t+1} \mid x_{t+1}, a_t).$$

It is assumed that a migration decision is influenced by the states of the countries $X_{t,k}$ at time t, with $k = 0, \ldots, K$. These represent a set of selected real-valued socioeconomic variables that characterize the attractiveness of a given location. We may have, for example, $X_{t,k} = (W_k, L_k) \in \mathbb{R}^2_+$, where $W_k > 0$ denotes the wage rates and $L_k > 0$ is the quality of life, $k = 0, \ldots, K$, that is, a state of information about a pecuniary and a non-pecuniary factor, respectively. Additional factors–such as individual preferences for certain countries or migration times–or potential information on how W_k and L_k will evolve during the decision process can be taken into account.

The attractiveness of a location can be measured in terms of its utility (e.g., in a utility maximisation model) or value (e.g., in a value expectancy model). It reflects a behavioural model of migration, based on which a migrant is more likely to move if they expect to be better off elsewhere. Considering a value expectancy model, and given an initial state $X_{0,k} = x_{0,k}$ associated with the country k where an individual is living, the problem is to find an optimal migration strategy $\boldsymbol{\pi}^* \doteq \{\pi_t^*\}_{t \geq 0}$ that maximizes the expression:

$$\mathbb{E}_{P_{\boldsymbol{\pi}}}[R_t] = \mathbb{E}_{P_{\boldsymbol{\pi}}}\left[\sum_{t \geq 0} \gamma^t r_{t+1}(\cdot, A_t) \mid X_{0,k} = x_{0,k}\right]. \tag{8}$$

The concept of a migration strategy reflects that of a general policy, i.e., a sequence of rules $\{\pi_t\}_{t \geq 0}$, where each π_t is a mapping from the state space that characterizes the country an individual lives in at time t to the space of migration options. Basically, it prescribes which country to move for the period $[t, t+1)$ depending on the current state observed $x_{t,j}$. Note that only the states associated with the country in which a migrant has lived are observed; information on the other countries remains unknown to the agent before exploring it. The migrant's decision to whether to move or not will thus be based on the *expectation* they may have on that country (and, of course, the cost).

An important novelty of the approach proposed in Berninghaus and Seifert-Vogt (1987)–compared to other works that incorporate incomplete information in the migratory context (see Molho 1986, for an overview)–is the use of algorithms developed under the theory of stochastic dynamic optimization, more specifically within the MAB framework introduced in Sect. 2.3. An illustrative example using the Gittins index is presented below. So far, the progress in incorporating decision processes into migration models has been increasingly growing (see e.g., the concept of *agent-based models* for describing individual-based dynamics; Klabunde and Willekens 2016), but their solution has been limited to standard optimization techniques, ignoring the potential of RL.

3.1.1 An illustrative example

Consider a guest worker who has to choose between two options $\mathscr{A} = \{0, 1\}$, with $k = 1$ corresponding to moving into a host country, and $k = 0$ corresponding to staying at home. Let us assume that if the worker stays at home ($k = 0$) he is completely informed of his net returns in terms of both the wage rate $W_0 = w_0$ and the quality of life $L_0 = l_0$, whereas the state values of the host country ($k = 1$) are only partially known before they are experienced. We recall that capital letters denote random variables and small letters their observed or realized value. Specifically, the following schematical sequence of information acquisition can be utilized:

- $t = 0$: the decision process starts with the artificial state

$$X_{0,k} = \begin{cases} (w_0, l_0) : \text{“complete information about country } k\text{”} & k = 0, \\ (w_1, L_1) : \text{“incomplete information about country } k\text{”} & k = 1; \end{cases}$$

- $t = 1$: if the migrant has moved to country $k = 1$, after one period, the state of information changes to

$$X_{1,k} = (w_k, l_k), \quad k = 0, 1,$$

that is, the migrant obtains full information exactly after one period living in the host country $k = 1$.

At time $t = 0$, the migrant has no certainties about the non-pecuniary return, but they are informed about the pecuniary factor. This may be the case of a worker who signs a contract with a fixed wage rate before migrating. To formalize this assumption we suppose that the prior (before migration) non-pecuniary return for the worker is a random variable L_1 where

$$L_1 = \begin{cases} l_1 & \text{with probability } \frac{1}{2} \\ -l_1 & \text{with probability } \frac{1}{2}, \end{cases}$$

and $l_1 > 0$. At each time *t,* the decision process is evaluated according to the value of each country's state, as defined in Eq. (8). The reward $Y : X \times A \to \mathbb{R}$, that is, the utility the migrant can extract from living in a country k, is conceptualized as a function of W_j and L_j, and it is supposed to be a Markov process where:

$$Y_{t+1}(X_{t,0}, 0) = w_0 + l_0, \quad t = 0, 1, \ldots, T \tag{9}$$

for the home country, while for the host country we could express it as

$$\begin{aligned} Y_1(X_{0,1}, 1) &= w_1 - k_1, \quad (X_{0,1} = w_1) \\ Y_2(X_{1,1}, 1) &= w_1 + L_1, \quad (X_{1,1} = (w_1, L_1)), \end{aligned} \tag{10}$$

where k_1 denotes the migration cost of moving into country $k = 1$. Note that other formulations for the reward function, and well as for the state definition, are possible; we point to Berninghaus and Seifert-Vogt (1987) for more details. We also emphasize that the representation in Eqs. (9) and (10) gives a simplified example to make the Gittins index easily tractable and understandable. They could be formulated according to more accurate domain theories, with the inclusion of the main determinants of quality of life l_k of a country k, as well as quality of life indicators related to the countries of interest.

Once the migrant is fully informed about the relevant variables in the two countries they can decide whether to stay in that country or move back. For illustrative purposes we assume:

$$\mathrm{P}\big(Y_{t+1}(X_{t,1}, 1) = w_1 \pm l_1 \mid Y_t(X_{t-1,1}, 1) = w_1 \pm l_1\big) = 1, \quad \forall t \geq 2.$$

In other words, the state $w_1 \pm l_1$ is an absorbing state of the Markov process.

For this specific example, as shown in McCall and McCall (1984), the Gittins indices, following the definition in Eq. (5), are given by:

$$\begin{aligned}
\mathscr{G}_0(x_{\tau,0}) &= w_0 + l_0, \quad \tau \geq 0, \\
\mathscr{G}_1(x_{\tau,1}) &= w_1 \pm l_1, \quad \tau \geq 1, \\
\mathscr{G}_1(x_{0,1}) &= \begin{cases} \frac{(w_1 - k_1) + \frac{1}{2}\frac{\gamma}{1-\gamma}(w_1 + l_1)}{1 - \frac{1}{2}\gamma}, & k_1 < \frac{l_1}{1-\gamma}, \\ (w_1 - k_1) + \frac{1}{2}\frac{\gamma w_1}{1-\gamma}, & k_1 > \frac{l_1}{1-\gamma}. \end{cases}
\end{aligned}$$

By these indices, it is possible now to derive the migration history of the migrant in our stylized example. Suppose $\mathscr{G}_0(x_{0,0}) \geq \mathscr{G}_1(x_{0,1})$; then migration will not take place at all. If $\mathscr{G}_0(x_{0,0}) < \mathscr{G}_1(x_{0,1})$, the worker will move for at least one period into the host country, and after one period living there, the optimal decision will be to stay in the host country forever if $\mathscr{G}_1(x_{1,1}) \geq \mathscr{G}_0(x_{1,0})$, and return to the home country otherwise. As expected, the decision depends not only on wages and moving costs, but–when l_1 is "large" relative to k_1–on its non-pecuniary attributes. While in principle the illustrated model suggests that individuals can freely migrate at any time t, we acknowledge that constraints to migration (such as visa requirements or family circumstances) may be in place. One way to express this constraint may be through the cost component k_1, which could be formulated as a function of individual characteristics. More generally, if other dimensions relevant to the migration decision are to be considered, they may be entered into the reward function through the quality of life or a new additional component. We refer to McCall and McCall (1984) for other extended models of migration.

3.2 Public health: management of infectious diseases

In early 2020, the COVID-19 pandemic engulfed the world. Since its debut, it has altered the course of the global economy and devastated human communities (Horton 2021), with dramatic harms on health, on population, and on the society as a whole. Before vaccines and effective medical treatments were made available, global governments faced the emergency with testing, contact tracing, and lockdowns to ensure social distancing and to mitigate unnecessary travel. However, it became quickly obvious that these restrictive measures would not have been economically and socially sustainable in the long run. As a consequence, many countries sought to judiciously relax restrictions on travel and social distancing, with the risk of allowing the spread of asymptomatic or presymptomatic infected individuals. Certainly, extensive active testing might have been an ideal solution, but, at that moment, resources were scarce and often inaccurate.

In the context of international travelling, most of the nations adopted border screening protocols, typically based on traveller's country of origin and differentiating between high-, mid-, and low-risk countries. As reported in Bastani et al. (2021), most of the European nations defined the risk entirely based on population-level epidemiological metrics such as positivity

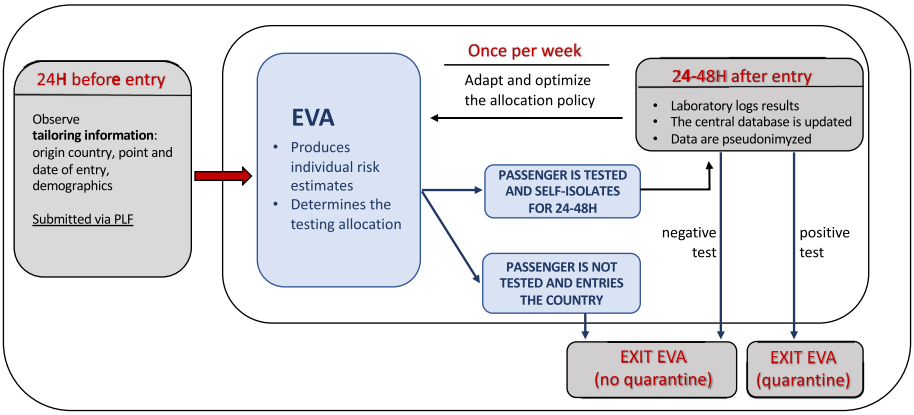

Fig. 4 Schematic of the RL-based *EVA* system adopted by the Greek government to limit the influx of travellers infected with severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)

rates, which may be biased by under-reporting or reporting delays, among other factors. An exception was Greece, engaged in the nationwide (across all Greek borders) deployment of *EVA* (Bastani et al. 2021): a real-time RL-based system for personalized COVID-19 screening. The goal was dual: i) to collect real-time data to monitor the epidemic progress and to inform better boarder policies; and ii) to combine the accumulated data on previously tested passengers with the new individual data so as to determine whether they should be screened or not at the Greek border. In contrast to country-wide protocols, *EVA* adopted a personalized data-driven approach to determine the high-risk individuals to be allocated to the scarce (screening) resource adaptively. In fact, the prevalence of COVID-19 was constantly evolving as the pandemic ebbed and flowed, with changing risks across subgroups. Hence, the ideal policy should adapt to the dynamic nature of the pandemic. In Fig. 4, we illustrate the operational flow of *EVA*.

Formally, denote with $r_t(\mathbf{x})$ the unknown underlying risk for a passenger with characteristics $\mathbf{x}$ on day t, where $r_t(\mathbf{x})$ is defined as the probability of a random passenger with characteristics $\mathbf{x}$ to test positive on on day t. Let $N_t(\mathbf{x}, e)$ be the number of travellers with characteristics $\mathbf{x}$ arriving at entry point e on day t, with $e \in \mathscr{E} = \{1, \dots, 40\}$. Finally, let $B_t(e)$ denote the e-th entry-specific testing budget determined exogenously by the Secretary General of Public Health. The goal of *EVA* was to determine the number passengers $n_t(\mathbf{x}, e)$ with characteristics $\mathbf{x}$, arriving at entry e on day t to screen in order to maximize the expected number of infections caught at the border over the time horizon T (summer of 2020). That is, the goal is to maximize

$$\max_{n(\cdot)} \mathbb{E}\left[\sum_{t=1}^{T}\sum_{e\in\mathscr{E}}\sum_{\mathbf{x}\in\mathscr{X}} n_t(\mathbf{x}, e) r_t(\mathbf{x})\right], \tag{11}$$

subject to constraints on the entry-specific testing budget and capacity,

$$\sum_{\mathbf{x}\in\mathscr{X}} n_t(\mathbf{x}, e) \le B_t(e), \quad \forall e \in \mathscr{E}$$
$$n_t(\mathbf{x}, e) \le N_t(\mathbf{x}, e), \quad \forall e \in \mathscr{E}, \forall \mathbf{x} \in \mathscr{X}.$$

Note that the true traveler risk $r_t(\mathbf{x})$ is unknown; it must be estimated using accumulated data from previous time steps, resulting in a typical *exploration-exploitation* dilemma characterizing MABs (Lattimore and Szepesvári 2020). The chosen MAB algorithm was the optimistic Gittins index: each arm (testing vs. not testing) was associated with a deterministic index that represented its "risk score", incorporating both its estimated prevalence and uncertainty. Unlike the classical stochastic bandit literature, the problem tackled here suffers from nonstationary rewards (risks change over time), batched decision making (decisions are made at the start of the day), delayed feedback (test results are returned only after 48 h), and specific constraints (entry-specific budgets and arrivals), which had to be accounted for when designing the algorithm.

Overall, this work represents a successful example of the potential of RL and large-scale real-time data for safeguarding public health. However, while the COVID-19 pandemic shone a light on the importance of flexible data-driven adaptive experimentation and/or deployments for ethical and efficient decision making, it has also highlighted that in public-health contexts such practices are very limited. We refer to Weltz et al. (2022) and Liu et al. (2023) for a general survey on RL methods in public health.

3.3 Optimized mobile health interventions for family planning and contraceptive use

Mobile-health (mHealth) interventions have gained significant attention in recent years and are now part of the WHO "long-term strategic plan to develop and implement eHealth services, to develop the infrastructure for information and communication technologies for health, to promote equitable, affordable and universal access to their benefits" (WHO 2021). Aided by the use of mobile technologies such as smartphones, mHealth interventions directly target attitudes and behavior changes, by disseminating, gathering, and analyzing health-related data and supporting interventions. Several studies, ranging from physical activity to substance addition, highlight the benefits of mHealth for their potential to improve healthcare delivery and outcomes among the general public (Liu et al. 2023; WHO 2011). Additionally, these tools can be utilized for online consultations, medication adherence, and health literacy.

Various works focusing on sexual and reproductive health showed that mobile technologies have the potential to improve the uptake of services and support family planning. For example, SMS text messages can be used as reminders to improve attendance to doctor appointments and compliance with medications or contraceptives (Halpern et al. 2013; Lopez et al. 2016; McCarthy et al. 2018). In upper-middle and high-income countries, there has been a proliferation of mHealth apps and digital devices, often equipped with sensors, that monitor menstrual cycles, track basal body temperature, and record other relevant data about individuals' fertile windows. By analyzing the collected information, the app can act as a data-driven support system for drug-free cycle-based contraception or for

increasing the chances of successful conception for individuals and couples planning to conceive or already expecting a child (see e.g., the UK *Natural Cycles* app; Pearson et al. 2021). In low- and middle-income countries, studies have tested the use of mobile SMS text messages to encourage discussion about family planning and to prevent unintended pregnancies (Athey et al. 2021; Babalola et al. 2019). It is worthnoting that expansion of contraceptive use in most impoverished countries is also part of the *Family Planning 2020 Commitment to Action*,[3] especially among adolescents (Sánchez-Páez and Ortega 2018).

In addition to providing accessible and up-to-dated educational resources, when equipped with artificial intelligence systems such as RL, these apps can offer recommendations on preconception care, healthy lifestyle choices, or nutrition, and can suggest specific family-planning approaches tailored to individual circumstances. While the adoption of mHealth technologies combined with RL remains low, experimental studies are being conducted.

We now illustrate a study protocol of *The World Bank*, which has been approved by the Cameroon National Ethics Committee to test a digital counseling approach that allows for shared decision making in contraceptive use and family planning.[4] The study occupies a relevant place at the intersection between public health, mortality and fertility. In fact, Cameroon had a maternal mortality ratio of 529 in 2017 and a total fertility rate of about 4.6 in 2018, with a fifth of all births being unwanted or considered mistimed by the mother (Organization 2019). Thus, increasing the uptake of contraceptives may have relevant benefits in terms of the fertility goals for this country, and, in turn, on maternal mortality rates, welfare losses, abortions, school dropouts and early marriages, among others (Bearak et al. 2020).

Specifically, the aim of the study in question is to evaluate whether a tablet-based job-support tool for nurses conducting family planning consultations, along with price discounts for contraceptives, can increase the uptake of long-acting reversible contraceptives among reproductive-age females in Cameroon (including adolescents who may be unmarried). At the time of enrollment in the study, each participant t is assigned to one of the study arms, each given by a combination of:

- The way the contraception options are viewed = {"side-by-side" or *status quo* view for individuals who do not express any contraceptive preference, "sequential" view with their preferred contraceptive method displayed at the top if a preference is expressed};
- A random price = { "Free", "Low", "Mid", "High" }. It is important to note here that even the "High" price constitutes at least a 20% discount relative to the normal prices charged at the hospital.

For each time or individual t, the objective is defined in terms of a loss function, conceived as a weighted sum of "failures" $F_t(k)$ and costs $C_t(k)$ associated to intervention k:

$$Loss_t(k) = \alpha_t F_t(k) + \lambda C_t(k). \tag{12}$$

Here, F_t represents a prediction for the probability of an unwanted pregnancy within the next year and is defined as a function of the contraceptive method:

[3] http://2014-2015progress.familyplanning2020.org/.

[4] For an abridged version of the study protocol see: https://www.worldbank.org/en/programs/sief-trust-fund/brief/cameroon-a-sequential-and-adaptive-experiment-to-increase-the-uptake-of-long-acting-reversible-contraceptives-among-adol.

$$F_t(k) \doteq \begin{cases} 0.05\% & \text{if participant adopted implant} \\ 0.8\% & \text{if participant adopted intra-uterine device} \\ 6\% & \text{if participant adopted pill} \\ 9\% & \text{if participant adopted injectable} \\ 25\% & \text{otherwise} \end{cases}$$

The first term in Eq. (12) takes into account the amount of potential unwanted pregnancies, with α_t representing a deterministic function of a participant's age. It is maximal when a participant is 15 years old (the youngest age allowed in the experiment) and decays linearly until 20 years old. The second term in Eq. (12) relates to the cost of contraception, while the parameter $\lambda > 0$ controls the trade-off between minimizing costs and failures.

The research team proposed an RL framework, more specifically a Thompson sampling variant, to deliver an *online* policy with the advantages of making the experiment more efficient (by making it faster) and more ethical (by assigning more individuals to the treatment condition with the highest probability of success for their context). We refer to the detailed *Pre-Analysis Plan*[5] of the study for further details.

4 Discussion and conclusions

Demographic change is one of the key challenges to be faced by global nations today. The process is complex and its outcomes are both uncertain and consequential. Population dynamics vary widely among different countries (Vollset et al. 2020), and the development process of population policies must be tailored to the specific needs of a country and the specific characteristics of its population. Efforts should be stepped up to incorporate uncertainty in population projections, human behavior, and expected outcomes into flexible policy-making strategies. As stated in the *Modernising Government* White Paper of 1999 (Cabinet 1999), among a number of features, a policy-making process should: (i) take a long-term view; (ii) systematically evaluate early outcomes; and (iii) learn from accumulated experience.

Well suited to the problem at hand, in this work we introduced the powerful reinforcement learning framework for formalizing and solving complex decision-making problems in population research. While a couple of existing works have touched on the promise of RL in applied demography (this is the case of the migration problem illustrated in Sect. 3.1), to our knowledge, this represents the first piece of work that systematically introduces the rich RL framework and illustrates examples of its potential applicability for supporting population policies. The examples presented here only scratch the surface of the promise of RL in population problems. The field is in its infancy, and we expect to see growing interest in the near future.

Our focus in this work was principally on providing mathematical formalization and illustrative examples to guide theoretical and applied researchers into understanding the RL framework. However, applying RL to problems for population policies requires considerations that go beyond the general RL formalization and the development of effective learning algorithms. Therefore, it is not only necessary, but dutiful, to discuss the limitations associated with the application of RL in population research.

[5] Available online at https://www.socialscienceregistry.org/versions/91771/docs/version/document.

In the first place, we note that most demographic and population data are macro (or aggregate) in nature. Within the context of population policies, the *World Population Policies Database*[6] provides up-to-dated and detailed longitudinal country-by-country information on national plans and strategies, as well as implemented programs. These include policies on population ageing, fertility and family planning, urbanization and international migration. Such macro-oriented data may produce strong evidence on population trends and patterns, as well as their associations across time and space. They are crucial in the discovery phase (regarded as the "core" of demography by many scholars), but, as pointed out by Billari (2015), do not provide an adequate means for understanding how the action and interaction of individual units generated what is discovered. Micro-level data are essential to infer how (average) population change arises from individual behavior in response to specific actions and interactions, and their need has been long advocated (see e.g., the "theory of change and response" in Davis 1963). Notably, the quantity of individual-level population data is exploding (Langedijk et al. 2019; Ruggles 2014) and efforts have been made to ensure their free access for the academic community; see e.g., the IPUMS project,[7] releasing data for almost 800 million observations drawn from 300 censuses of about 100 countries (McCaa and Ruggles 2002). Big micro-data represent a new kind of source material and they are expected to provide individual-level data about entire populations spanned over multiple time-points and at high geographic resolution (Ruggles 2014; Rampazzo et al. 2023). However, although the potential comprehensiveness of data continue to improve, due to privacy and security factors, there are significant access restrictions for these data. Further, collection of high-quality longitudinal data, reflecting a potential effect of a policy, remains limited. In the migratory case, for example, one would need individual data at discrete time points for the same unit and information on the origin and host countries at each migration time. Progress towards the collection of high-quality micro-data, arising for example from experimental or quasi-experimental studies, could take policy-making to a different level. In particular, the availability of real-time data would encourage processes of online learning and near-real-time policy adaptation. At the moment, micro-based simulations and agent-based modeling (Billari and Prskawetz 2003), supported by formal demography for ensuring a proper design and validation, remain a solution and may have useful place in RL-based population research. Examples are provided in Heiland (2003), Kniveton et al. (2011), Klabunde and Willekens (2016), Willekens (2017), and in Billari et al. (2007), and Bijak et al. (2013), who advocated a wider use of simulation models for international migration and family planning, respectively. An extensive review of agent-based models of human migration is provided in Klabunde and Willekens (2016). In particular, decision making is emphasized in behavioural models of migration, wherein individuals are likely to leave a location if they expect an improved status elsewhere conditioned on manageable barriers to migration. Such practices may also be useful in terms of understanding how key demographic processes could change if we were able to incorporate in our modeling difficult or impossible to observe quantities (see e.g., Ciganda and Lorenti 2019).

A substantial challenge in applying RL in practice is represented by the need to adequately express each of the basic ingredients of the RL problem (see Sect. 2.1) in relation to the specific application. These include the identification and selection of the state variables, an accurate definition of the reward function or the utility model, and the way we

[6] https://www.un.org/development/desa/pd/data/world-population-policies.

[7] https://usa.ipums.org/usa/.

define the dynamics of the process, among others. In domain applications, RL algorithms are often trained and applied according to some representation of the world, which necessarily involves making some (simplifying) assumptions about reality. In the illustrative example of international student migration, the way we formalize a student's preferences shapes the "reality" that guides a government actions. In the individual migration example introduced in Sect. 3.1, a Markov property is assumed for the states and reward progression, following the *human capital theory* of Sjaastad (1962). Several other theories that give explanations and predictions of why people migrate could be employed; these include the neoclassical theory, the new economics of migration theory, and social capital theory, among others (see e.g., Massey 1999; Van Hear 2010). Any simplifying assumption, such as the Markov property, should be properly investigated in the specific context to which it is applied. For example, in the repeated migration study of Constant and Zimmermann (2012), authors assess the plausibility of this assumption based on the *German Socio-Economic Panel* (GSOEP; Goebel et al. 2019), a nationally representative survey in Germany of people aged 16 or older, including legal immigrants that started in 1984. The authors concluded that it may be an appropriate representation for their setting. All these aspects should be informed by domain knowledge. Considering also the novelty of the approach, potential applications in social sciences, particularly in demography, will require joint collaboration and strong synergy between methodologists and applied researchers with expertise in the specific applied topic, such as migration.

It should also be mentioned that the development of optimal population policies requires an underlying framework for assessing the causal effect of each alternative intervention and, most importantly, each policy (Engelhardt et al. 2009). Real-world applications should be developed under appropriate considerations of causal inference. Finally, as with any primer introduction of new frameworks or methods to an application area, there are many challenges ahead, from the scarcity of sufficiently "good" data to interpretation, ethics, and fairness concerns, among others.

> The challenge, then, is not how to use new technology to entrench old problems, but instead to better understand how we may use this opportunity to imagine and design systems that are more transparent, equitable, and just. (Molnar and Gill 2018)

Acknowledgements The author sincerely appreciates and thanks the two anonymous reviewers for their valuable comments that have substantially improved the manuscript. The author expresses her gratitude for the appreciation received on the work, in particular on its potential value for future research.

Funding Open access funding provided by Università degli Studi di Roma La Sapienza within the CRUI-CARE Agreement. The author acknowledges funding from the H2020 project (Protocol number PH11916B88B59064) titled "*New Statistical Methods for the Analysis of Human Migration*".

Declarations

Conflict of interest The authors have no conflicts of interest to declare.

Ethics approval Not applicable.

Consent for publication All co-authors have seen and agree with the contents of the manuscript. We certify that the submission is original work and is not under review at any other publication.

References

Afsar, M.M., Crump, T., Far, B.: Reinforcement learning based recommender systems: a survey. ACM Comput. Surv. **55**(7), 1–38 (2022)

Agrawal, S., Goyal, N.: Thompson sampling for contextual bandits with linear payoffs. In: International conference on machine learning, pp. 127–135 (2013)

Aguilera, A., Figueroa, C.A., Hernandez-Ramos, R., Sarkar, U., Cemballi, A., Gomez-Pathak, L., Yan, X.: mhealth app using machine learning to increase physical activity in diabetes and depression: clinical trial protocol for the diamante study. BMJ Open **10**(8), e034723 (2020)

Ahn, N., Alho, J., Brücker, H., Cruijsen, H., Laakso, S., Lassila, J., Valkonen, T.: The use of demographic trends and long-term population projections in public policy planning at eu, national, regional, and local level. Report prepared for the European Commission, Brussels: European Commission (2005)

Athey, S., Bergstrom, K., Hadad, V., Jamison, J.C., Ozler, B., Parisotto, L., Sama, J.D.: Shared decision-making: Can improved counseling increase willingness to pay for modern contraceptives? (2021)

Babalola, S., Loehr, C., Oyenubi, O., Akiode, A., Mobley, A.: Efficacy of a digital health tool on contraceptive ideation and use in Nigeria: results of a cluster-randomized control trial. Glob. Health Sci. Pract. **7**(2), 273–288 (2019)

Banha, F., Flores, A., Coelho, L.S.: A new conceptual framework and approach to decision making in public policy. Knowledge **2**(4), 539–556 (2022)

Bansak, K., Ferwerda, J., Hainmueller, J., Dillon, A., Hangartner, D., Lawrence, D., Weinstein, J.: Improving refugee integration through data-driven algorithmic assignment. Science **359**(6373), 325–329 (2018)

Bastani, H., Drakopoulos, K., Gupta, V., Vlachogiannis, I., Hadjichristodoulou, C., Lagiou, P., Tsiodras, S.: Efficient and targeted covid-19 border testing via reinforcement learning. Nature **599**(7883), 108–113 (2021)

Bearak, J., Popinchalk, A., Ganatra, B., Moller, A.B., Tunçalp, Ö., Beavin, C., Alkema, L.: Unintended pregnancy and abortion by income, region, and the legal status of abortion: estimates from a comprehensive model for 1990–2019. Lancet Glob. Health **8**(9), e1152–e1161 (2020)

Bellman, R.: Dynamic Programming, 1st edn. Princeton, NJ, USA (1957)

Berninghaus, S.K., Seifert-Vogt, H.G.: International migration under incomplete information. Stämpfli (1987)

Bertsekas, D.P.: Reinforcement Learning and Optimal Control. Athena Scientific Belmont, MA (2019)

Bijak, J., Hilton, J., Silverman, E., Cao, V.D.: Reforging the wedding ring: exploring a semi-artificial model of population for the united kingdom with gaussian process emulators. Demogr. Res. **29**, 729–766 (2013)

Billari, F.C.: Integrating macro-and micro-level approaches in the explanation of population change. Popul. Stud. **69**(sup1), S11–S20 (2015)

Billari, F.C.: Demography: fast and slow. Popul. Dev. Rev. **48**(1), 9–30 (2022)

Billari, F.C., Prskawetz, A.: Agent-based computational demography: using simulation to improve our understanding of demographic behaviour. Springer Science & Business Media, Berlin (2003)

Billari, F.C., Prskawetz, A., Aparicio Diaz, B., Fent, T.: The "wedding-ring'' an agent-based marriage model based on social interaction. Demogr. Res. **17**, 59–82 (2007)

Boswell, C.: Migration in europe. In: The Politics of Migration, pp. 91–110. Routledge (2018)

Brezzi, M., Lai, T.L.: Optimal learning and experimentation in bandit problems. J. Econ. Dyn. Control **27**(1), 87–108 (2002)

Buettner, T.: Population projections and population policies. In: International Handbook of Population Policies, pp. 467–484. Springer (2022)

Bullock, H., Mountford, J., Stanley, R.: Better policy-making. Centre for Management and Policy Studies London (2001)

Cabinet Office: Modernising Government. A White Paper presented to Parliament. The Stationery Office, London (1999)

Carammia, M., Iacus, S.M., Wilkin, T.: Forecasting asylum-related migration flows with machine learning and data at scale. Sci. Rep. **12**(1), 1457 (2022)
Chakraborty, B., Moodie, E.: Statistical Methods for Dynamic Treatment Regimes, vol. 10, p. 9781. Springer-Verlag, Berlin (2013)
Chakravorty, J., Mahajan, A.: Multi-armed bandits, gittins index, and its calculation. Methods Appl. Stat. Clin. Trials Plan. Anal. Inferential Methods **2**(416–435), 455 (2014)
Charpentier, A., Elie, R., Remlinger, C.: Reinforcement learning in economics and finance. Comput. Econ. 1–38 (2021)
Ciganda, D., Lorenti, A.: Using simulated reproductive history data to re-think the relationship between education and fertility. Social informatics: 11th international conference, socinfo 2019, Doha, Qatar, November 18–21, 2019, proceedings 11, pp. 218–238 (2019)
Constant, A.F., Zimmermann, K.F.: The dynamics of repeat migration: a Markov chain analysis. Int. Migr. Rev. **46**(2), 362–388 (2012)
Davis, K.: The theory of change and response in modern demographic history. Popul. Index **29**(4), 345–366 (1963)
De Lara, M., Doyen, L.: Sustainable Management of Natural Resources: Mathematical Models snd Methods. Springer Science & Business Media, Berlin (2008)
Deliu, N., Chakraborty, B.: Dynamic treatment regimes for optimizing healthcare. The Elements of Joint Learning and Optimization in Operations Management, pp. 391–444. Springer (2022)
Engelhardt, H., Kohler, H.P., Prskawetz, A.: Causal Analysis in Population Studies. Springer, Berlin (2009)
Figueroa, C.A., Aguilera, A., Chakraborty, B., Modiri, A., Aggarwal, J., Deliu, N., Lyles, C.R.: Adaptive learning algorithms to optimize mobile applications for behavioral health: guidelines for design decisions. J. Am. Med. Inform. Assoc. **28**(6), 1225–1234 (2021)
Figueroa, C.A., Deliu, N., Chakraborty, B., Modiri, A., Xu, J., Aggarwal, J., Aguilera, A.: Daily motivational text messages to promote physical activity in university students: results from a microrandomized trial. Ann. Behav. Med. **56**(2), 212–218 (2022)
Fishburn, P.C.: Utility theory for decision making. Research analysis corp McLean VA (1970)
Gilboa, I.: Theory of Decision Under Uncertainty, vol. 45. Cambridge University Press, Cambridge (2009)
Gittins, J.: A dynamic allocation index for the sequential design of experiments. Progress in statistics 241–266 (1974)
Gittins, J., Glazebrook, K., Weber, R.: Multi-armed Bandit Allocation Indices. John Wiley & Sons, New Jersey (2011)
Goebel, J., Grabka, M.M., Liebig, S., Kroh, M., Richter, D., Schröder, C., Schupp, J.: The german socio-economic panel (soep). Jahrbücher für Nationalökonomie und Statistik **239**(2), 345–360 (2019)
Hallsworth, M.: Policy-making in the real world. Polit. Insight **2**(1), 10–12 (2011)
Halpern, V., Lopez, L.M., Grimes, D.A., Stockton, L.L., Gallo, M.F.: Strategies to improve adherence and acceptability of hormonal methods of contraception. Cochrane Datab. Syst. Rev. (10) (2013)
Heiland, F.: The collapse of the berlin wall: Simulating state-level east to west german migration patterns. Agent-Based Computational Demography: Using Simulation to Improve Our Understanding of Demographic Behaviour, pp. 73–96. Springer (2003)
Horton, R.: The Covid-19 Catastrophe: What's Gone Wrong and How to Stop Happening Again. John Wiley & Sons, New Jersey (2021)
Kintner, H.J., Pol, L.G.: Demography and decision-making. Popul. Res. Policy Rev. 579–584 (1996)
Klabunde, A., Willekens, F.: Decision-making in agent-based models of migration: state of the art and challenges. Eur. J. Popul. **32**(1), 73–97 (2016)
Kniveton, D., Smith, C., Wood, S.: Agent-based model simulations of future changes in migration flows for Burkina faso. Glob. Environ. Chang. **21**, S34–S40 (2011)
Kosorok, M.R., Laber, E.B.: Precision medicine. Annu. Rev. Stat. Appl. **6**, 263–286 (2019)
Langedijk, S., Vollbracht, I., Paruolo, P.: The potential of administrative microdata for better policy-making in europe. Data-Driven Policy Impact Eval. 333 (2019)
Lattimore, T., Szepesvári, C.: Bandit Algorithms. Cambridge University Press, Cambridge (2020)
Liu, D.Y. T., Bartimote-Aufflick, K., Pardo, A., Bridgeman, A.J.: Data-driven personalization of student learning support in higher education. Learning Analytics: Fundaments, Applications, and Trends, pp. 143–169. Springer (2017)
Liu, X., Deliu, N., Chakraborty, B.: Microrandomized trials: developing just-in-time adaptive interventions for better public health. Am. J. Public Health **113**(1), 60–69 (2023)
Lopez, L.M., Grey, T.W., Tolley, E.E., Chen, M.: Brief educational strategies for improving contraception use in young people. Cochrane Datab. Syst. Rev. (3) (2016)
Massey, D.S.: Why does immigration occur?: a theoretical synthesis. na (1999)

McCaa, R., Ruggles, S.: The census in global perspective and the coming microdata revolution. Scand. Popul. Stud. **13**, 7–30 (2002)
McCall, J.J., McCall, B.P.: The economics of information: a sequential model of capital mobility. Diskussionsbeiträge-Serie A (1984)
McCarthy, O.L., Wazwaz, O., Osorio Calderon, V., Jado, I., Saibov, S., Stavridis, A., Huaynoca, S.: Development of an intervention delivered by mobile phone aimed at decreasing unintended pregnancy among young people in three lower middle income countries. BMC Public Health **18**, 1–15 (2018)
Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Ostrovski, G.: Human-level control through deep reinforcement learning. Nature **518**(7540), 529–533 (2015)
Molho, I.: Theories of migration: a review. Scot. J. Polit. Econ. **33**(4), 396–419 (1986)
Molnar, P., Gill, L.: Bots at the gate: a human rights analysis of automated decision-making in Canada's immigration and refugee system (2018)
Movsisyan, A., Arnold, L., Copeland, L., Evans, R., Littlecott, H., Moore, G., Rehfuess, E.: Adapting evidence-informed population health interventions for new contexts: a scoping review of current practice. Health Res. Policy Syst. **19**(1), 1–19 (2021)
Nigri, A., Levantesi, S., Aburto, J.M.: Leveraging deep neural networks to estimate age-specific mortality from life expectancy at birth. Demogr. Res. **47**, 199–232 (2022)
OECD: Interrelations between public policies, migration and development. OECD Publishing, Paris (2017). Available at https://doi.org/10.1787/9789264265615-en
World Health Organization (WHO) (2019): Trends in maternal mortality 2000 to 2017: estimates by WHO, UNICEF, UNFPA, world bank group and the united nations population division (2019)
Pearson, J.T., Chelstowska, M., Rowland, S.P., Mcilwaine, E., Benhar, E., Berglund Scherwitzl, E., Scherwitzl, R.: Natural cycles app: contraceptive outcomes and demographic analysis of UK users. Eur. J. Contracept. Reprod. Health Care **26**(2), 105–110 (2021)
Rampazzo, F., Rango, M., Weber, I.: New migration data: Challenges and opportunities. Handbook of Computational Social Science for Policy 345 (2023)
Ruggles, S.: Big microdata for population research. Demography **51**(1), 287–297 (2014)
Sánchez-Páez, D.A., Ortega, J.A.: Adolescent contraceptive use and its effects on fertility. Demogr. Res. **38**, 1359–1388 (2018)
Sjaastad, L.A.: The costs and returns of human migration. J. Polit. Econ. **70**(5, Part 2), 80–93 (1962)
Steinhubl, S.R., Muse, E.D., Topol, E.J.: The emerging field of mobile health. Sci. Transl. Med. **7**(283), 283rv3-283rv3 (2015)
Sugiyama, M.: Statistical Reinforcement Learning: Modern Machine Learning Approaches. CRC Press, Boca Raton (2015)
Sutton, R.S., Barto, A.G.: Reinforcement Learning: An Introduction. MIT press, Cambridge (2018)
Tsiatis, A.A., Davidian, M., Holloway, S.T., Laber, E.B.: Dynamic Treatment Regimes: Statistical Methods for Precision Medicine. Chapman and Hall/CRC, Boca Raton (2019)
UNECE: Mainstreaming ageing - revisited. unece policy brief on ageing no. 27 (2022). Available at https://unece.org/sites/default/files/2022-02/ECE-WG.1-39-PB27.pdf
United Nations Department of Economic and Social Affairs, Population Division (2022). World population prospects 2022: Summary of results. UN DESA/POP/2022/TR/NO. 3
Van Hear, N.: Theories of migration and social change. J. Ethn. Migr. Stud. **36**(10), 1531–1536 (2010)
Vollset, S.E., Goren, E., Yuan, C.W., Cao, J., Smith, A.E., Hsiao, T., Chalek, J.: Fertility, mortality, migration, and population scenarios for 195 countries and territories from 2017 to 2100: a forecasting analysis for the global burden of disease study. Lancet **396**(10258), 1285–1306 (2020)
Weltz, J., Volfovsky, A., Laber, E.B.: Reinforcement learning methods in public health. Clin. Ther. **44**(1), 139–154 (2022)
Willekens, F.: The decision to emigrate: a simulation model based on the theory of planned behaviour. Agent-Based Modell. Popul. Stud. Concepts Methods Appl. 257–299 (2017)
World Health Organization (WHO) (2011). Mhealth: new horizons for health through mobile technologies (2011)
World Health Organization (WHO) (2021). Global strategy on digital health 2020-2025. Geneva: World Health Organization; 2021. Licence: CC BY-NC-SA 3.0 IGO (2021)