Contents

Highlights

Reinforcement Learning for crop management support: review, prospects and challenges.

- Reinforcement Learning is a promising AI framework to support crop management.
- Reinforcement Learning-based crop management support literature is scarce.
- A Reinforcement Learning-based system should learn from interactions on the ground.
- Crop management support is related to many Reinforcement Learning research questions.
- Joint research by the Reinforcement Learning and Agronomy communities is required.

Reinforcement Learning for crop management support: review, prospects and challenges.

Romain Gautron[a,b,c,*], Odalric-Ambrym Maillard[d], Philippe Preux[e], Marc Corbeels[a,b,f] and Régis Sabbadin[g]

[a] CIRAD, UPR AIDA, F-34398 Montpellier, France

[b] AIDA, Univ Montpellier, CIRAD, Montpellier, France

[c] CGIAR Platform for Big Data in Agriculture, Alliance of Bioversity International and the International Center for Tropical Agriculture (CIAT), Regional Office for the Americas, Km 17, Recta Cali-Palmira 763537, Colombia

[d] Université de Lille, Inria, CNRS, Centrale Lille UMR 9189 – CRIStAL, F-59000 Lille, France

[e] Université de Lille, CNRS, Inria, Centrale Lille UMR 9189 – CRIStAL, F-59000 Lille, France

[f] International Institute of Tropical Agriculture, PO Box 30772, Nairobi, 00100, Kenya

[g] Université de Toulouse, INRAE, UR MIAT, F-31320, Castanet-Tolosan, France

ARTICLE INFO

Keywords:
Reinforcement Learning
Multi-Armed Bandit
Machine Learning
Decision Support System
Crop Management

ABSTRACT

Reinforcement Learning (RL), including Multi-Armed Bandits, is a branch of Machine Learning that deals with the problem of sequential decision-making in uncertain and unknown environments through learning by practice. While best known for being the core of the Artificial Intelligence (AI) world's best Go game player, RL has a vast potential range of applications. RL may help to address some of the criticisms leveled against crop management Decision Support Systems (DSS): it is an interactive, geared toward action, contextual tool to evaluate series of crop operations faced with uncertainties. A review of RL use for crop management DSS reveals a limited number of contributions. We profile key prospects for a human-centered, real world, interactive RL-based system to face tomorrow's agricultural decisions and theoretical and ongoing practical challenges that may explain its current low take-up. We argue that a joint research effort from the RL and agronomy communities is necessary to explore RL's full potential.

*Corresponding author:
romain.gautron@cirad.fr (R. Gautron)
ORCID(s): 0000-0002-3218-7215 (R. Gautron)

1. Introduction

Reinforcement Learning (RL), a branch of Machine Learning and more generally Artificial Intelligence (AI), addresses the control of uncertain and unknown dynamical systems. Although information about recent research in RL is widely available, it is too specialized and abstract to be easily understandable (Lapan, 2018, preface). RL is potentially a well suited paradigm to support crop management decisions, but few applications are found in the literature. This paper aims to help the RL and agronomy communities to gain mutual understanding, identify promising research directions and current bottlenecks to foster future joint research to support the design of the next human-centered and data-driven crop management decision support tools. We first define the crop management decision problem as an element of farm decision-making, and describe the dedicated Decision Support Systems. Section 2 introduces the RL paradigm. Section 3 provides a review focused on RL applied to crop management. Finally, Section 4 explores research opportunities and challenges for the use of RL to support crop operation decisions.

Crop management. Crop management is the logical and ordered combination of agricultural practices or operations applied to a field in order to obtain a particular crop production (Sebillotte, 1974, 1978). A field plot is the site of complex interactions happening between biotic (all living organisms) and abiotic components (soil and atmosphere as supports for living organisms) and crop management through physical, biological and chemical processes, as demonstrated by Husson et al. (2021) with soil Eh-pH dynamics. Consequently, decisions about these operations occur in the face of uncertain events (e.g. climatic events), and within a dynamical system that is only partially known. We consistently use the adjective uncertain for events with unsure realizations.

Through crop management, farmers aim to obtain a production result that matches as closely as possible the targets they defined at the beginning of the cultivation period, such as a minimum yield level and certain quality criteria. Typically, at the start of the cropping season, a crop management plan is defined, as illustrated in Figure 1. This plan follows a logical structure, but is an uncertain procedure that requires adaptations to the events occurring during the growing season. Each crop operation is parametrized by multiple factors which determine its outcome and success, further conditioning the remaining crop cycle and future crop operations (Boiffin et al., 2001). For instance, once a cultivar is chosen, the planting operation is defined by a planting date, planting density, sowing depth, possible chemical seed treatment and the choice of machinery (with its own parameters, such as sowing speed) in a mechanized context.

Operational observations during the cultivation period may reveal issues farmers cannot predict with certainty, such as an outbreak of pests and/or diseases, and this will require adaptive operations. Based on the severity of an unforeseen event, the objective defined before cultivation such as a minimum yield might be revised to compensate for these changes (Cerf and Sebillotte, 1988; Papy, 1998). For instance, if a drought occurs after planting, a farmer may not provide a second fertilizer dose to maize as the application cost is not likely to be rewarded by a yield increase. Consequently, the farmer may reduce the yield target.

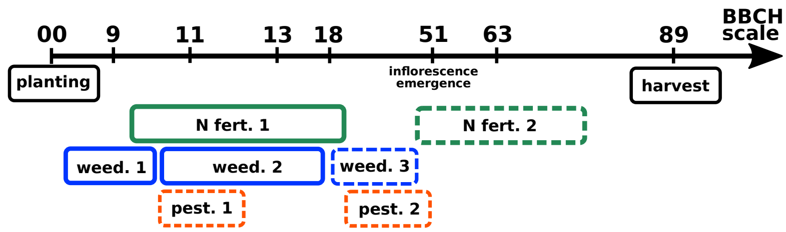

Figure 1: A simplified example of maize management plan. The BBCH scale follows the successive maize growth stages as found in Meier (1997). A dashed box indicates that the operation requirement is uncertain. All operations are made within a time window where the exact date of occurrence is uncertain. 'N fert.' stands for nitrogen fertilization ; 'weed.' for weeding ; 'pest.' for pest and disease control.

Farm decision-making levels. Farm decision-making encompasses multiple nested levels over different time and spatial scales (Chatelin et al., 1993; Papy, 1998). For instance, a cropping system refers to an ensemble of plots equally treated with the same crop rotation (an ensemble of crop types in a given successive order) and crop management (see Sebillotte, 1974; Boiffin et al., 2001). While long-term decisions on a structural production system level are made on an annual to multi-year time line, such as investments in land or machinery, perennial crop implementation or annual cropping systems, crop management decisions are made on a monthly to daily basis.

Levels of decision-making may strongly interact. Indeed, the strategic and tactical[1] levels may be affected by operational events, as a recurrent operational issue may motivate a change in machinery or crop rotation.

Decision Support Systems (DSS). Decision Support Systems (DSS) are computer-based solutions designed to assist decision makers in addressing unstructured or semi-structured problems (Arnott and Pervan, 2005; Power, 2008). Structured problems have unambiguous solutions which can be found with an automatic routine. In contrast, semi-structured or unstructured problems have incomplete or uncertain information with possibly unforeseen events and complex trade-offs between different objectives. DSS provide distilled information as evidence to facilitate and improve human decision-making.

DSS are used in a broad range of domains. For instance, DSS are commonly used in railway track maintenance scheduling to avoid derailments (e.g. Ferreira and Murray, 1997; Guler, 2013), for medical diagnostics (Miller, 2016), or operation planning. As an example, da Silva et al. (2006) designed a DSS to optimize the number of workers, overtime hours and the level of outsourcing in order to meet trade-offs between economic returns to maximize profits while maintaining client and worker satisfaction. DSS can be geared towards a single user from an operator to an executive, to a group that shares decision-making responsibility, or be used to support negotiation between different parties. DSS are not meant to provide off-the-shelf solutions to decision makers to solve a given problem but, rather, to provide a human-machine dialogue, as pointed out by Arnott and Pervan (2005).

Crop Management DSS. Commonly found DSS supporting crop management deal with fertilization, irrigation, pest and disease or weed management; the end users may be researchers, local advisers or farmers. Crop management DSS come in various forms, from advanced user-oriented complex crop models, to easy to use graphical user interface software or even spread sheets (examples can be found in Manos et al., 2004; Cerf and Meynard, 2006; Le Gal et al., 2010; Evans et al., 2017; Jones et al., 2017). In general, they intended to support decisions taken under great uncertainty.

[1]We define the *strategic* level as long term, covering more than a few years; the *tactical* level as intermediate, ranging from a few years to a few months; and the *operational* level ranging from a few months to a daily basis.

For instance, decisions on pest and disease control are usually based on the assessment of the imminence or intensity of crop damage (Gent et al., 2011). They depend on complex interactions of uncertain biotic factors, such as the crowding effect and host-plant response, and a-biotic factors, such as temperature and humidity (Khaliq et al., 2014).

Crop management DSS are based on underlying formal models of various complexity which predict the consequences of actions. These models can take many different formalisms, sometimes combined: a simple set of equations such as soil nitrogen balances (Hébert, 1969; Stanford, 1973), knowledge bases for expert systems (e.g. Lemmon, 1986; Sønderskov et al., 2016), mechanistic models explicitly simulating the processes at stake with crop growth using differential equations (e.g. McCown et al., 1995; Hoogenboom et al., 2019; Brisson et al., 2003) or machine learning models (e.g. Navarro-Hellín et al., 2016; Waghmare et al., 2016; Ip et al., 2018; Sabzi et al., 2018; Barbosa et al., 2020; Saikai et al., 2020). The modeling part is usually done offline, based on prior data. The exploration of candidate crop operations can be made by manual expert guided search (e.g. Thorburn et al., 2011; He et al., 2012), an inference engine for knowledge bases (e.g. Lemmon, 1986), or by using numerical optimization techniques (e.g. Epperson et al., 1993; Bergez et al., 2001; Royce et al., 2001; Saikai et al., 2020).

Despite the existence of numerous applications, the level of crop management DSS use among farmers remains low, as shown by McCown (2002a,b); Hochman and Carberry (2011); Gent et al. (2011); Rose et al. (2016); Evans et al. (2017). The use of DSS in family farming depends on the user's willingness and interest, and is directly related to potential learning through DSS, as emphasized by McCown (2002a); Evans et al. (2017). Thorburn et al. (2011) provide an example of a group comprising sugarcane farmers and local industry representatives who, supported by scientists, learned through a DSS. Based on simulations, the group jointly explored and discussed the environmental benefits of splitting nitrogen applications. While the simulations did not show clear benefits in splitting the applications, the authors concluded that there was an improved understanding of nitrogen dynamics among participants, and thereby a better understanding of the consequences of nitrogen fertilizer management at the individual level. Agricultural DSS have a life cycle where dis-adoption may occur after users have learned and internalized the assessment of risk in decisions, without being a sign of failure (Thorburn et al., 2011; Gent et al., 2011; Evans et al., 2017).

Several critiques and guidelines for the use of DSS in crop management can be found in the literature. In particular, users have deemed that DSS information cannot directly be turned into actions, that farmers' natural decision-making processes are not adequately taken into account, that the sequential nature of decisions is poorly modelled or that risk management is lacking in the decision process (McCown, 2002a,b; Cerf and Meynard, 2006; Hochman and Carberry, 2011; Evans et al., 2017). Ideas of a "discussion support software" from Nelson et al. (2002), or an "information and advice system" from Cerf and Meynard (2006) or Hochman and Carberry (2011) describe DSS that take advantage of the social tissue in which farmers evolve. A DSS should integrate information fluxes at different scales –from plot to regional– and from various actors involved in multi-level decisions such as local suppliers, pest control advisers and environmental protection bodies.

2. Reinforcement Learning

In this section, we shall introduce the ideas behind Reinforcement Learning (RL). In Section 2.1, we informally present the elements of RL. Section 2.2 then formalizes an RL problem. In Section 2.3, we provide a short historical perspective of RL. Section 2.4 presents the famous Q-learning RL algorithm. In Section 2.5 we describe the main RL algorithm categories. Finally, Section 2.6 is dedicated to bandit algorithms, a particular case of RL adapted to small-sample settings.

2.1. Overview of Reinforcement Learning

Machine Learning (ML) is the study of computer programs designed to perform a task and able to self-improve with data or experience (Mitchell et al., 1997). Machine Learning comprises three subfields: Unsupervised Learning, Supervised Learning, and Reinforcement Learning. Unsupervised Learning deals with learning a representation of data, for instance with clustering tasks. Supervised Learning is about learning to label new data based on a set of labelled data (examples) with classification and regression tasks (Mitchell et al., 1997). Reinforcement learning is about learning to control a dynamical system. After a ML model has been trained to perform a given task based on training situations, its performance is measured as its ability to perform the same task in situations that have not been met during the training phase. Overfitting is a recurrent issue in ML, which occurs when, after being trained, a model performs well in training situations but performs poorly in unseen situations.

A Reinforcement Learning problem is a sequential decision-making problem in which a decision maker iteratively interacts with an **environment** which is an unknown and uncertain dynamical system. The decision maker, called the **agent**, learns the task of controlling the evolution of the environment by taking **actions**. A **policy** corresponds to a set of decision rules which determines which action the agent takes, generally depending on an **observation** of the environment. The learning process proceeds through a loop of interactions between the agent and its environment. Each time the agent performs an action according to its policy, the action affects the environment and the agent receives a return. A **return** is a scalar value which indicates how the agent performs with regard to the task to be completed. This process is repeated until a decision sequence eventually ends. The goal of the agent is to compute a policy which maximizes a utility function of the returns it receives during a sequence of decisions, called an **objective function**. To do so, the agent adjusts its policy based on the returns it has collected through its experience. The RL loop is summarized in Figure 2. RL algorithms are inherently online methods, geared towards action, which react to the ongoing uncertain changes in a system and learn to perform a task by trial and error.

2.2. Formalization of a Reinforcement Learning problem

Markov Decision Processes. The canonical RL problem formulation models the environment as a **Markov Decision Process** (MDP, Puterman, 1994). At any moment, the environment is described by its **state** $s \in \mathcal{S}$. $\mathcal{S}$ is the state space, i.e. the set of possible states, known to the learner. Sequentially, at each moment $t \in \{0, 1, \cdots, T\}$ the agent chooses an action $a_t \in \mathcal{A}$ depending on the current state of the environment s_t. $\mathcal{A}$ is the action space, i.e. the set of possible actions, known to the learner. T is the **horizon** which may be known or not, and be finite or not. Performing an action affects the environment which transits to its next state $s_{t+1} \in \mathcal{S}$ according to the MDP

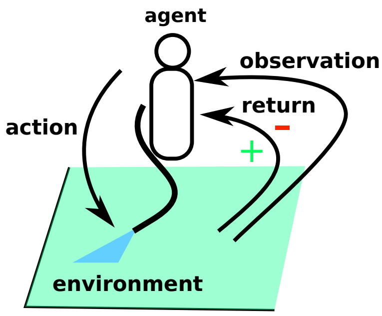

Figure 2: The Reinforcement Learning loop. A decision maker, called the agent, interacts with its environment. The agent's task is to control the environment's evolution. Sequentially, the agent takes an action based on an observation of the environment. The action impacts the environments, and the agent receives a return that indicates how it performs regarding the task to be completed. This loop repeats until the decision sequence eventually ends.

A Markov Decision Process (MDP) $\mathfrak{M}$ is defined by: $\mathfrak{M} = \langle \mathcal{S}, \mathcal{A}, \mathbf{p}, \mathbf{r} \rangle$

- $\mathcal{S}$ the state space,
- $\mathcal{A}$ the action space,
- $\mathbf{p}(s'|s,a)$ is the transition function which give the probability that the environment transits to state s' after action a is performed in state s,
- $\mathbf{r}(s,a,s')$ is the return function, that is the average return after the agent performed action a in state s resulting in a transition to s'.

Figure 3: The four elements of a Markov Decision Processes (MDP). A MDP models the environment in Reinforcement Learning problems.

transition function $\mathbf{p} : \mathcal{S} \times \mathcal{A} \to \mathcal{P}(\mathcal{S})$, where $\mathcal{P}(\mathcal{S})$ denotes the set of probability distributions over states. $\mathbf{p}(s'|s,a)$ is the probability of reaching $s' \in \mathcal{S}$ after action a has been performed in the state s. A random return r accompanies each transition of the environment from a state s to a state s' after taking an action a. We define the return function $\mathbf{r} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ as $\mathbf{r}(s,a,s') = \mathbb{E}[r|s,a,s']$.

In an MDP, the Markov property holds: the probability law of s_{t+1} is fully specified by the knowledge of (s_t, a_t) ; all anterior states and actions can be ignored. The quadruplet $\langle \mathcal{S}, \mathcal{A}, \mathbf{p}, \mathbf{r} \rangle$ is fixed: the environment is **stationary**. For instance, the probability of transiting from one state s to a next state s' after taking an action a is always the same. Figure 3 illustrates the elements forming an MDP. In Figure 4, we model a simplistic irrigation problem as an MDP.

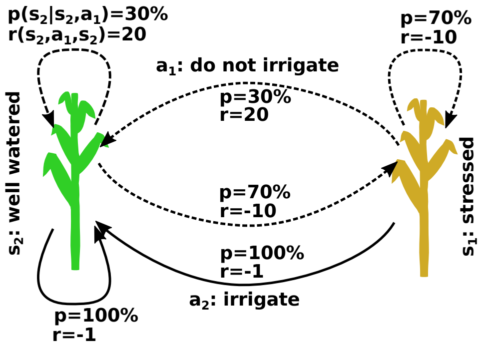

Figure 4: A simplistic irrigation problem modelled as a Markov Decision Process (MDP). Two states are possible: a stressed crop (s_1) or a well watered crop (s_2). Each arrow between two states is a transition which ends in the state pointed by the arrow's head. Watering the crop (a_1) always leads to a well watered state, but it has a cost, hence the negative return. If no irrigation is provided (a_2), 30% of the time rainfall occurs and the crop will be well watered for free, hence the great return. But, 70% of the time, no rainfall occurs and the crop gets stressed, which is highly penalized by the return.

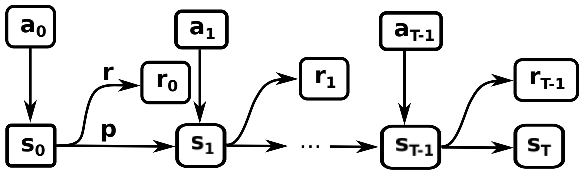

Figure 5: The representation of a sequence of decisions is called an episode. In a canonical Reinforcement Learning problem, starting with the environment in an initial state s_0, at each discrete decision step t, depending on the environment's current state s_t the agent decides on an action a_t thanks to its policy. After the agent takes the action a_t, the environment transits towards its uncertain next state s_{t+1}, given by the transition function **p**. The return function **r** provides a return r_t which indicates to the agent how it performs regarding the task to be completed.

Markov Decision Problems A Markov Decision Problem is the combination of a Markov Decision Process and an objective function to be optimized which is usually defined as the expectation $\mathbb{E}\left[R(t)\right]$ of the **discounted return** $R(t)$ collected by the agent (Puterman, 1994, p. 80):

$$R(t) = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \gamma^3 r_{t+4} + \cdots \quad (1)$$

$$= \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \quad (2)$$

where $\gamma \in [0, 1)$ is the discount factor. The use of γ can be interpreted as with discounted cash flows: future returns are less valuable than immediate returns. A sequence of interactions from an initial state to a given horizon is called a trajectory, or **episode**, which is illustrated in Figure 5.

A policy $\pi : \mathcal{S} \rightarrow \mathcal{P}(\mathcal{A})$ maps a state to probability distributions over actions. The objective of the agent is to find an optimal policy π^* that maximizes the objective function. To measure the performance of a policy π, we define the **Value function** function (V: $\mathcal{S} \rightarrow \mathbb{R}$) and the **Quality function** function (Q: $\mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$). Acting according to policy π, the value of a state s is the expected return starting from state s, denoted $\mathbb{E}_\pi[R(t)|s_0 = s]$; the quality of an action a in state s is defined as the value of first taking action a starting from state s and then following the policy π:

$$V_\pi(s) = \mathbb{E}_\pi\Big[R(t)|s_0 = s\Big], \forall s \in \mathcal{S} \tag{3}$$

$$Q_\pi(s,a) = \mathbb{E}_\pi\Big[R(t)|s_0 = s, a_0 = a\Big], \forall s \in \mathcal{S}, \forall a \in \mathcal{A} \tag{4}$$

Denoting Π the set of possible policies, there exists an optimal policy π^* such that:

$$Q_{\pi^*} \geq Q_\pi, \ \forall \pi \in \Pi \tag{5}$$

We have:

$$Q_{\pi^*}(s,a) = \max_\pi Q_\pi(s,a), \forall s \in \mathcal{S}, \forall a \in \mathcal{A} \tag{6}$$

$$\pi^\star(s) = \operatorname*{argmax}_{a \in \mathcal{A}} Q_{\pi^*}(s,a), \forall s \in \mathcal{S} \tag{7}$$

2.3. A brief historical perspective of Reinforcement Learning

We say that an MDP is known, or fully specified, when we have access to the MDP's transition probabilities given by the transition function **p** and reward function **r**, see Section 2.2. Historically, (Stochastic) Optimal Control (SOC, Kushner, 1967) addresses the control of systems with known MDP. The RL came from the merging of (S)OC and animal psychology to address the problem of controlling a system with an unknown MDP through trial and error: the environment is seen as a black box. (S)OC emphasizes stability analysis, frequency analysis of the controlled systems whereas RL emphasizes the learning process of controlling an unknown dynamical system. (S)OC deals with continuous time and actions while canonical RL problems deal with discrete time, states, and actions. Later, (S)OC and RL converged by addressing decision problems historically belonging to each other's fields, for instance continuous time, states, and actions in RL (e.g. Munos, 1996) and the discrete case in (S)OC (e.g. Bertsekas and Shreve, 1996). Figure 6 summarizes the main difference between SOC and RL.

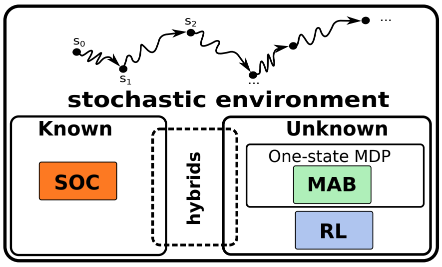

Figure 6: Both Stochastic Optimal Optimization (SOC) and Reinforcement Learning (RL) address the problem of controlling a system with uncertain dynamics. The main historical difference is that SOC supposes the dynamics of the system to be known while RL does not. Recently, hybrid algorithms have been developed, combining RL and SOC. The Multi-Armed Bandit (MAB) is a simplified case of RL with a one-state MDP, see Section 2.6.

2.4. Q-learning: a simple Reinforcement Learning algorithm

Q-Learning (Watkins, 1989) is one of the simplest RL algorithms. It consists of estimating Q_{π^*}, defined in Equation 6. We present its pseudo-code with algorithm 1. Q-learning leverages Bellman's optimality equation which makes explicit a recursive relation between the qualities of states for an optimal policy (Bellman, 1957):

$$Q_{\pi^*}(s,a) = \sum_{s'} \underbrace{\mathbf{p}(s'|s,a)}_{\text{weighing}} \Big[\underbrace{\mathbf{r}(s,a,s') + \gamma \times \overbrace{\max_{a' \in \mathcal{A}} Q(s',a')}^{\text{future optimal returns from } s'}}_{\text{discounted optimal returns } R(t) \text{ transiting from } s \text{ to } s'} \Big] \tag{8}$$

At each time step $t \in \{1, \cdots, T\}$, after the algorithm takes an action a_t depending on s_t and consequently observes return r_t and next state s_{t+1}, it updates:

$$\underbrace{Q(s_t,a_t)}_{\text{new prediction}} \leftarrow \underbrace{Q(s_t,a_t)}_{\text{current prediction}} + \alpha(s_t,a_t) \times \underbrace{\Big(\overbrace{r_t + \gamma \times \max_{a' \in \mathcal{A}} Q(s_{t+1},a')}^{\text{prediction target}} - Q(s_t,a_t) \Big)}_{\text{prediction error}} \tag{9}$$

for instance with learning rate $\alpha(s_t,a_t) = 1/\sqrt{N_{s_t,a_t} + 1}$ where N_{s_t,a_t} is the number of times the action a_t has been taken in state s_t. Assuming a proper learning rate and all (state, action) pairs are asymptotically visited an infinite number of times, the Q-value function which the Q-Learning algorithm learns is guaranteed to converge to Q_{π^*} (Bertsekas and Tsitsiklis, 1996).

2.5. Reinforcement Learning today

Modern RL algorithms stemmed from three archetypal methods shown in Figure 7: the **Critic**, **Actor**, and **Planning** methods. Planning methods focus on deriving a policy by interacting with a simulator of the true environment. Planning methods can be used when a simulator of the environment is available to the agent, or when the agent

Algorithm 1 Q-Learning algorithm

Input: $\epsilon \in (0, 1]$ `// the greediness parameter`
Initialize Q-values for all state-action pairs with arbitrate values
for *episode* $\in \{1, \dots, N\}$ **do**
  for $t \in \{1, \dots, T\}$ **do**
    observe environment's state s_t
    with a probability $1 - \epsilon$ choose the action a_t as $a^* = \arg\max_a Q(s_t, a)$, else randomly choose $a_t \in \mathcal{A}_t \setminus \{a^*\}$
    observe environment's next state s_{t+1} and return r_t
    update $Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha(s_t, a_t) \times \left(r_t + \gamma \times \max_{a' \in \mathcal{A}} Q(s_{t+1}, a') - Q(s_t, a_t)\right)$
  end
end
return *Q-values*

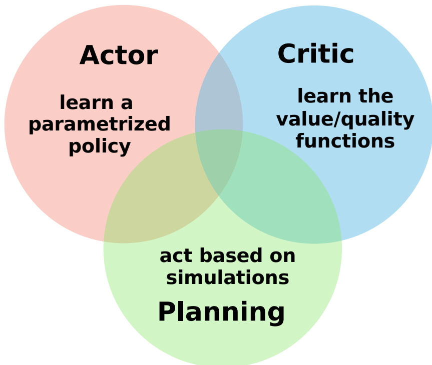

Figure 7: Modern Reinforcement Learning methods are hybrids of three problem solving methods: critic, actor and planning methods, see Section 2.5.

explicitly learns the transition and return functions of the MDP (i.e. a model of the environment) and learns an optimal policy at the same time. Because the potential number of trajectories to be explored is very large, the solutions must be explored efficiently. A celebrated planning algorithm is Monte Carlo Tree Search (Coulom, 2006; Kocsis and Szepesvári, 2006) which explores the most rewarding simulated trajectories to decide on an agent's action in a given state. The second class of algorithms are critic methods which consists in learning a value function V, or Q. One example is the Q-learning (Watkins, 1989) introduced in Section 2.4. Finally, actor methods directly learn an optimal policy in a parametrized fashion (a policy is modeled as a function of a set of parameters) without representing the V or Q functions. For instance REINFORCE (Williams, 1987) searches for an optimal policy using a gradient descent approach in the space of possible policies.

Most of the recent methods are hybrids of the three stems presented in Figure 7, combined with the use of Neural Networks (NN). An NN is made of a set of interconnected units structured in successive layers. Each unit is called a neuron. It computes a function made of simple arithmetic operations from multiple input values and outputs its result. NN are widely used due to the fact that they can approximate any bounded continuous function (Cybenko, 1989). Deep learning is dedicated to the study of the Deep Neural Networks which are neural networks made of multiple layers. Deep neural networks are a powerful way to represent functions when the number of state-action pairs is too large to represent with finite tables. An early achievement of an RL algorithm using NN is Tesauro's TD-Gammon program (Tesauro, 1995) which learned to play the game of backgammon through self-play, succeeding in challenging expert human players. Mnih et al. (2015) reached human performance playing Atari games using a combination of Q-Learning and a neural network (the Deep Q-Network algorithm, DQN). The Alpha-Go program (Silver et al., 2017), the world's best Go player, is a combination of actor, critic and planning methods using NN to deal with the 10^{170} states and 400 actions.

2.6. Multi-Armed Bandit

The Multi-Armed Bandit (MAB) problem (Lattimore and Szepesvári, 2020), originally introduced for drug allocation by Thompson (1933), can be seen as a special case of RL problem with a one-state MDP. For each time step $t \in \{1, \cdots, T\}$, the agent sequentially chooses a single action a among a fixed set of possible actions $\mathcal{A}$. Each time the agent selects an action $a \in \mathcal{A}$, it observes a return r drawn from a fixed distribution of returns of mean value $\mathbf{r}(a) = \mathbb{E}[r|s, a, s]$, and a transition back to the same single state s. In the most common setting, named cumulated regret minimization (Robbins, 1952), the agent's objective is to maximize the expectation of the undiscounted sum of rewards it has collected after time T, that is $\mathbb{E}[\sum_{t=1}^{T} r_t]$. This objective is equivalent to minimizing the expected regret, which is a measure of the expected total loss from sub-optimal action taking up to time T. To correctly identify optimal action(s), the agent must try all actions a sufficient, but *a priori* unknown, number of times –which implies choosing sub-optimal actions-. This is an example of the Exploration/Exploitation dilemma. For various families of algorithms, the bandit theory focuses on providing strong statistical guarantees for the expected regret.

The simpler problem formulation in MAB makes it possible to reduce the sample complexity of the decision problems –that is to say the number of samples required to solve a problem– compared to the general RL setting. MAB algorithms address a rich range of extension settings (Lattimore and Szepesvári, 2020). For instance, risk aware bandits (Cassel et al., 2018) evaluate actions with a risk measure. Considering a random variable X, the mean $\mathbb{E}[X]$ is said to be risk neutral as it equally weighs all possible outcomes whereas risk metrics typically stress bad possible outcomes. To exemplify this, the Conditional-Value-at-Risk (CVaR) at level $\alpha \in (0, 1]$ (Mandelbrot, 1997) can be defined as $\mathrm{CVaR}_\alpha(X) := \mathbb{E}[X|X \leq \mathrm{VaR}_\alpha(X)]$ where $\mathrm{VaR}_\alpha(X)$ is the quantile of probability α of X. When $\alpha \to 0^+$, CVaR_α tends to the worst case analysis and with $\alpha = 1$ it recovers the usual mean. Contextual bandits (Lattimore and Szepesvári, 2020, ch. 5) leverage extra information about the context of a decision, such as demographic data for online advertisements.

3. Review of Reinforcement Learning for agriculture

The following review reveals that while Stochastic Optimal Control (see Section 2.3) has been widely used to support farm level decisions, attempts to use RL for crop-management purposes are scarce and applications only considered simulated environments.

3.1. Early stirrings: farm decision-making under uncertainty

The inclusion of uncertainty and risk to support farm decision-making is not new. Early examples are Tintner (1955) and Freund (1956): stochastic linear programming was used to maximize a utility function for crop allocation under uncertainty and resource constraints at the farm level. The utility function depended on a farmer's net revenue and degree of risk aversion. Hildreth (1957) discussed the use of game theory (Osborne et al., 2004) to make a decision on crop production plans when the environment's dynamics are unknown. Risk treatment assumed that the worst possible scenario occurred. Burt and Allison (1963) later defined decision-making around the choice of crop rotations explicitly as a Markov Decision Problem (see Section 2.2) and addressed it using dynamic programming and Bellman's equation (Bellman, 1957), which are the foundations of modern RL.

Approaches using stochastic linear or dynamic programming and their derivatives are part of Stochastic Optimal Control (SOC). There are numerous examples in which (Stochastic) Optimal Control has been applied to farm level decision-making. These can be found in Kennedy (1986); Norton and Hazell (1986); Glen (1987); Lowe and Preckel (2004); Dury et al. (2012) and Weintraub and Romero (2006). Most of these applications were defined at the farm level addressing cropping plans or farm resource allocation, while this article focuses on crop management at the field level, see Section 1 for the distinction. As a recent example of an application of SOC, Boyabatlı et al. (2019) formalized a farmer's cropping plan decision problem as a finite horizon stochastic dynamic programming problem, to maximize in expectation an uncertain gross margin due to uncertain yield and selling price. They provided a decision heuristic which was nearly optimal and outperformed the ones provided by the literature.

3.2. Seminal works using Reinforcement Learning in agriculture

The seminal works which applied RL to crop-management are summarized in Table 1. Garcia (1999); Trépos et al. (2014) used the R_H-Learning algorithm from (Garcia and Ndiaye, 1998) which introduced adaptations of Q-learning (Watkins, 1989), see Section 2.4, and R-learning (Schwartz, 1993) –a variant of Q-learning with undiscounted returns i.e. $\gamma = 1$ in Equation 1– to the case of non-stationary finite-horizon MDPs. While Garcia (1999) considered continuous actions, Bergez et al. (2001); Trépos et al. (2014) considered discrete actions. Garcia (1999); Bergez et al. (2001); Trépos et al. (2014) all considered continuous state variables.

In all of these works, the use of RL relies on a crop model to simulate real field conditions. Crop models have their own limits: the policies obtained by RL were inherently limited by the simulation biases. The algorithms are not envisioned as using feedback from farmers to continuously improve the policy learned from the simulator. While Garcia (1999) focused on wheat yield maximization under strong limitations on nitrogen pollution of drinking water supplies, Bergez et al. (2001); Trépos et al. (2014) maximized the gross margin which induces *de facto* a great non-stationarity. Fossil-fuels are required to produce nitrogen fertilizer or to pump irrigation water: their price is known to be highly volatile and consequently an optimal management

1955 – Tintner (Stochastic Linear Programming)
1957 – Hildreth (Game Theory)
1963 – Burt and Allison (Dyn. Prog. with MDP)
1999 – Garcia; Ndiaye (R_H learning)

Figure 8: Key contributions towards Reinforcement Learning (RL) use in agriculture. Only Garcia (1999) is categorized as modern RL. Earlier work are based on paradigms that are the historical parents of RL.

strategy is likely to be different every year. Such non-stationarity is not problematic in a simulated setting: many simulations can be run before each season to train an agent to maximize the gross-margin. Nonetheless, for *in situ* field-trial based learning, this non-stationarity will dramatically increase the sample complexity which is already highly challenging. As shown in Table 1, the number of episodes to train agents ranges from 500,000 to 1,000,000, where one episode corresponds to one year in the real world: this clearly precludes any straight application of the learning procedure in real conditions.

In Trépos et al. (2014), for each episode of the learning process of the algorithm, a sample has randomly been chosen from 41 annual weather records to generate weather uncertainty. This limited number of weather records is likely to have induced overfitting. Because Trépos et al. (2014) evaluated their algorithm on the same weather records as the ones used during the training stage, the performance they measured was likely to be overly optimistic for unseen weather conditions. The use of a stochastic weather generator in Garcia (1999); Bergez et al. (2001) guaranteed more robust results with respect to weather uncertainty. Interestingly, after agent learning Garcia (1999) used an *ad hoc* automatic method of rule extraction to express an optimized policy in a naturalistic fashion "if this is observed then do this . . . ", i.e. as a set of simple decision rules that fit farmers' habits (Papy, 1998; Evans et al., 2017). Key contributions towards RL-supported crop management are summarized in Figure 8.

3.3. Deep Reinforcement Learning applications

Recently, Deep RL techniques have been suggested for crop management support. The Internet of Things (IoT) refers to networks of uniquely identified physical devices (sensors and/or actuators) which can autonomously communicate between themselves

Table 1
Principal works which have applied algorithms to crop management. (c) indicates a continuous variable; (integer) indicates the number of discrete elements; (y/n) indicates a binary feature. In all works, decisions are made during a single growing season.

Reference	Number of decisions	State variables	Actions	Return	Algorithm	Number of episodes	Weather generator	Baseline	Results
Garcia (1999)	3	• planting date • tillering date • plant density (c) • N in soil (c) • date of start the stem elongation • aerial biomass (c)	• seed rate (c) • cultivar • basal N date • basal N rate (c) • top N date • top N rate (c)	yield thresholded to 0 if post-harvest nitrogen in soil greater than 30 kg/ha at crop harvest.	R_H-Learning (Garcia and Ndiaye, 1998)	800,000	yes	experts' policy	The algorithm learned strategies for wheat management under strong nitrogen pollution constraint which performed close to the experts' policy without outperforming them.
Bergez et al. (2001)	daily	• soil water deficit (c) • accumulated thermal units (c)	• irrigate (binary)	gross margin at crop harvest.	Q-Learning (Watkins, 1989)	1,000,000	yes	policy obtained by Dynamic Programming (DP) solving	Reinforcement Learning solutions were better than DP with less than 100,000 learning steps which then exhibited similar performances.
Trépos et al. (2014)	4	• N in soil (c) • water in soil (c) • aerial biomass (c) • plant nutrition (c) • planting date • past fertilization • past herbicide applications	• planting date (3) • first fertilization (3) • herbicide application (y/n) • second fertilization (6)	gross margin at crop harvest	R_H-Learning (Garcia and Ndiaye, 1998)	500,000	no	fixed crop management plan obtained by exhaustive search	a 18% margin increase compared to the optimal fixed crop management plan.

or with humans, and process data (Rose et al., 2015). Bu and Wang (2019) have proposed a general IoT architecture for smart decision-making in agriculture based on Deep Q-Learning which combines Deep Neural Networks and Q-learning (see Section 2), to directly learn from field trials. The authors discuss the use of improved efficiency algorithms using Transfer Learning (see Taylor and Stone, 2009; Weiss et al., 2016), which is discussed in Section 4, and relatively Multitask Learning (Zhang and Yang, 2021). In a foresight study, Binas et al. (2019) also see potential in combining RL with Deep Learning for sustainable agriculture and propose similar solutions to overcome learning process limitations, such as the use of crop simulators to pre-train algorithms and the use of short-cycle plants for *in situ* learning.

Several works have recently applied (Deep) RL techniques to support crop-management in simulated environments. Wang et al. (2020) used Deep RL with Transfer Learning to control the CO_2 concentration and humidity in a simulated greenhouse to maximize cucumber cumulative weight. Sun et al. (2017) applied RL and Wang et al. (2020); Yang et al. (2020); Chen et al. (2021) applied Deep RL to control the irrigation at the field level, based on atmospheric, soil and plant state features; Chen et al. (2021) included seven day forecasts in the state. The objective functions of Sun et al. (2017) and Yang et al. (2020) were related to the gross margin at crop harvest; in Chen et al. (2021) the return is a score related to rainfall use efficiency and yield. (Wang et al., 2020; Sun et al., 2017; Yang et al., 2020; Chen et al., 2021) compared the performances of their RL algorithms to already existing decision models based on expert knowledge or machine learning. They measured superior performances of their RL algorithms.

However, we should mention that these recent applications share a common caveat in the method of evaluation of their performances. The authors evaluated their algorithms with a single year of the weather time series and/or with weather time series used during the training phase. Because of the enhanced flexibility of Deep RL techniques compared to more basic RL algorithms, they are more prone to overfitting. The evaluations of the authors are likely to be over-optimistic. A proper evaluation should ideally be done with a great number of weather time series, unused during the training phase and the performances should be presented with a measure of their uncertainty.

3.4. Multi-Armed Bandits

Currently, the use of the MAB framework to support crop management remains anecdotal. Kirschner and Krause (2019) tailored a contextual bandit algorithm, see Section 2.6, for cultivar choice to maximize the yield under uncertain weather forecasts. A decision context was defined as the union of climatic suitability factors (Holzkämper et al., 2013) and the cultivation site. The authors evaluated their algorithm thanks to a regression model of wheat yield trained on multiyear field trials. Their algorithm was substantially outperformed by the exact knowledge of future weather conditions prior to the decision, but showed better performances for other decision problems.

Baudry et al. (2021) provide a MAB example of a risk-aware bandit for crop management. They evaluated their algorithm for maize planting date decision-making using the `DSSAT` crop simulator (Hoogenboom et al., 2019) to maximize the CVaR at level α of grain yield, see Section 2.6, where α models a farmer's risk aversion. For each decision made, the weather used by DSSAT during the growing season was stochastically generated using the WGEN (Richardson and Wright, 1984) weather generator. The algorithm of Baudry et al. (2021) proved to be state-of-art for this decision problem. For practical use, ongoing work addresses the adaptation of the algorithm of Baudry et al. (2021) to batch recommendations, i.e. recommendation to a group of farmers each year to increase the number of samples, the original algorithm being purely sequential (one observation per year).

3.5. RL applications in other domains

Li (2019) presents some examples of RL real-world applications, including recommender systems, computer systems, energy, finance, robotics and transportation. Nevertheless, the practical use of RL remains sporadic in industry at the time this article is being written. Over the past few years, research efforts in the field of RL *sensu lato* have focused on other challenging application domains, such as personalized adaptive treatments in health care. As a particularly interesting *in vivo* bandit application, Durand et al. (2018) designed a contextual MAB for sequential drug administration to maximize the information collected from mouse experiments.

4. Prospects and challenges

In Section 4.1, we first present what conceptually could be an on-farm, human-centered RL-based crop management DSS. Section 4.2 prospects how RL problem solving could help to address the challenges of future agricultural decision-making and to further match farmers' decision-making processes. Section 4.3 details the specific learning challenges associated with learning from interactions in true conditions. Figure 10 wraps up the elements orbiting around a ground-learning RL DSS that we discuss in this Section.

4.1. An RL-based crop management DSS

We start by introducing what could be an on-farm RL-based crop management DSS, learning from on-the-ground experiences. A trained RL agent is viewed as an assistant for a human-centered system in the vein of Evans et al. (2017). For instance, an agent's task is to learn to maximize yield under a pollution constraint, as found in Garcia (1999). We suppose that at any time during the growing season, a farmer can query the RL agent. The agent has access to a snapshot describing the field characteristics at the moment it takes a decision, such as: past fortnight meteorological features, current plant nitrogen and water stress status from leaf inspection (after leaf emergence) and the crop growth stage. Depending on the farmer's settings –such as risk aversion level–, defining its constraints and objectives, and plot field state, the agent provides tailored recommendations.

A farmer may first query a planting date choice at the beginning of the growing season. Once a decision has been made by the farmer, the RL agent is provided with the farmer's decision and a time step later, the field parameters are measured again to evaluate the effect of the action that has been taken. The user may request the next time step to evaluate fertilization in the same fashion with the RL agent's support. This time, nitrogen stress would probably increase pest control needs in the area, thus suggesting a minimal fertilization level requirement. The whole interactive process is eventually repeated until the end of crop cultivation by an ensemble of farmers every season. Such an approach would consequently be a dynamic, interactive system between farmers, fields and agent(s) as illustrated in Figure 9. As an on-farm real-world RL system would learn from an ensemble of individual experiences on the ground, it is *de facto* a cooperative system supported by a community of farmers.

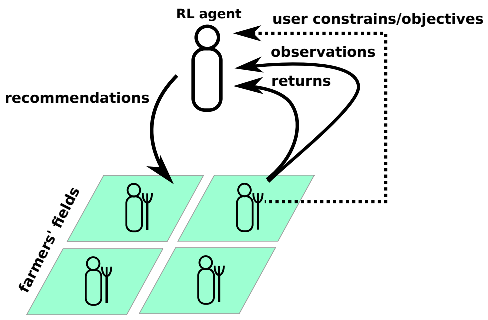

Figure 9: An RL-based Decision Support System for a community of farmers. At any moment, a farmer can query the agent to explore tailored crop management recommendations based on farmer's constraints and objectives. Data should be interactively and iteratively exchanged between farmers and the agent in order to collectively improve the policy for crop management decision problems.

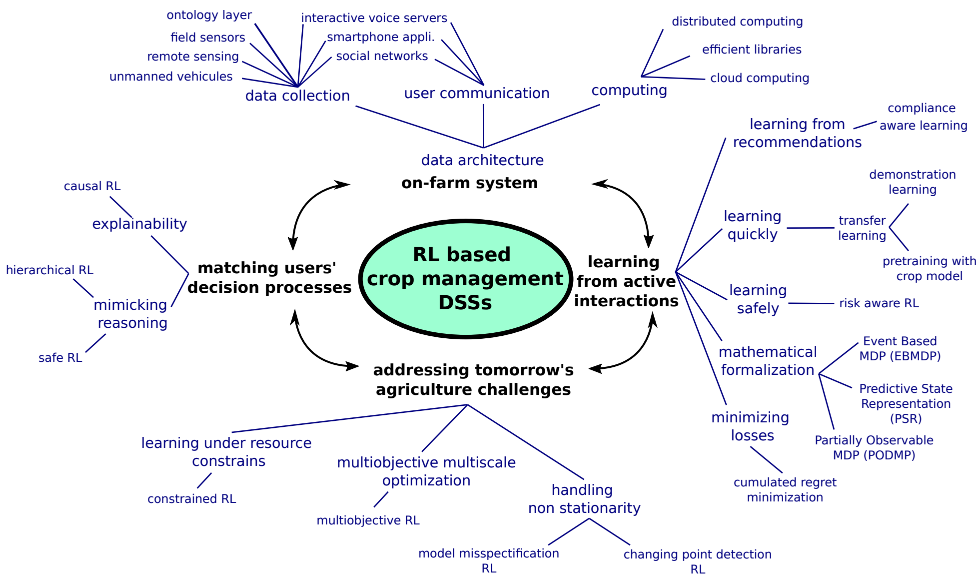

Figure 10: Challenging features and respective prospects for RL-based crop management Decision Support Systems. The inner circle represents the desirable features for an RL based crop management DSS. All of these features inter-relate. The outer circle represents the potential technical or theoretical solutions to reach the corresponding features of the inner circle.

Data collection. An RL on-farm solution would learn from a substantial number of interactions on the ground to evaluate the actions taken. The new data collection techniques and computing frameworks summarized in Table 2 could make this interactive learning possible. With such a system, field data (state measurements) must

Table 2
Technological opportunities for Reinforcement Learning (RL) applications. The interactive communication between a virtual agent and the ground reality with farmers, as shown in Figure 9, require an *ad hoc* data architecture to allow the RL loop. The back end system is dedicated to agent's computational requirements. The data collection elements essentially captures fields states. Finally, the communication elements allow the human-machine dialog.

Technology	Back-end System	Data Collection	Communication
High-level machine Learning libraries	✖		
Distributed systems	✖		
Cloud computing	✖		
Remote sensing		✖	
Unmanned aerial systems		✖	
Field sensors		✖	
Social network platforms		✖	✖
Smartphone applications		✖	✖
Interactive voice response servers		✖	✖

be collected such as human observations (e.g. pest and disease inspection), field sensors (e.g. soil moisture sensors) or remote sensing (e.g. to derive plant stress). Action recommendations must be communicated to the user or additional observations may be requested. Once the user has taken an action, which is not necessarily the recommended one, it should be communicated to the system. The use of field sensors requires the determination of the minimum density for optimal coverage and the minimum frequency of data capture for it to be efficient. More generally, each field observation has a cost which is likely to depend on its precision. A semantic layer is necessary to ensure data harmonization and relevant annotations: digital fieldbooks are an example of such efforts (Shrestha et al., 2010). Crowd sourcing requires specific data management, including *ad-hoc* data quality assessment. Field data traceability is another desirable feature (Quinton et al., 2019).

Data architecture. An overall RL data architecture is necessary to handle recurrent communications between farmers and agent(s) at each decision-making stage. Producing relevant recommendations assumes the storage of past interactions and an *ad-hoc* back-end system to learn from the data. Cloud computing (Hayes, 2008) and distributed computation (Attiya and Welch, 2004) combined with optimized software libraries would be basic tools. Providing personalized recommendations to approximate individual constraints and objectives requires the storage of user-specific information in the data architecture. This consequently raises the common question of data privacy in agriculture (Sykuta, 2016).

4.2. Prospects

RL appears to be a promising paradigm for meeting the challenges of future agricultural decision-making and to further match farmers' decision-making processes.

4.2.1. Tackling tomorrow's challenges

Faced with increasing decision-making complexity and processes that are too complex/uncertain to be jointly modeled, directly learning through the experience thanks to RL provides interesting perspectives. In particular, sharing farmers' tacit individual experiences, as explored in Evans et al. (2017). As Goulet et al. (2008) point out, farmers also innovate and this knowledge should be leveraged.

Researchers usually employ crop model to elaborate crop management in the context of changing climates. As an example, Adam et al. (2020) show that in the Sudano-Sahelian zone, a change in sorghum cultivar has a marginal effect compared to an increase in nitrogen fertilizer use when examining the impact of climate change on grain yield. Nevertheless, Falconnier et al. (2020) point out that the effects of nitrogen fertilization and an elevated CO_2 concentration or nitrogen mineralization combined with high temperatures are so far unsatisfyingly modelled. Agroecology is a promising paradigm for change-resilient agriculture (Altieri et al., 2015). Agroecological systems are highly complex, and modeling has been limited. For instance, simulations of pest and disease dynamics are limited (Donatelli et al., 2017); intercropping modelling is still in its early stage and so highly uncertain (Chimonyo et al., 2015). Even under well simulated processes, climatic projections still remain uncertain, for instance with the impact of climate change on droughts (Cook et al., 2018).

Special RL adaptations have been developed for changing environments, named non-stationary, such as a region under climate change. Change point detection in MAB algorithms (see Hartland et al., 2006; Mellor and Shapiro, 2013; Liu et al., 2018) addresses non-stationary situations and may be extended to MDP (e.g. Padakandla et al., 2020); the Model Misspecification framework also addresses non-stationarity in

MDP (e.g. Mankowitz et al., 2020).

4.2.2. Matching users' decision processes

Hochman and Carberry (2011) write "decision support systems need to better match farmers' naturalistic decision-making processes [. . .]". RL appears to be close to the description of farmers' decision processes. Cerf and Meynard (2006); Evans et al. (2017) point out that farmers usually use small-scale tests and learn by trial and error, repeating experiments under different conditions over the years, given the cyclical nature of crop management. McCown (2002a) uses the expression "learning-in-action". Sebillotte and Soler (1988); Papy (1998) describe how farmers refine crop operations based on successive intermediary crop state checkpoints, as RL does. The use of small-scale tests also directly refers to the Exploration/Exploitation dilemma introduced in Section 2.6: farmers seek to learn potentially better options, but also want to limit potential losses that may occur due to a change in practices. The cumulated regret minimization is largely present in the bandit literature and increasingly found for the general RL setting, for instance with the UCRL algorithm (Auer and Ortner, 2006; Auer et al., 2008). To our knowledge, currently no data-driven crop management support model enjoys such properties.

4.2.3. Learning safely

Farmers have been described to be primarily interested in support for highly uncertain decisions and risk to be a central stressful decision-making determinant (see Cerf and Sebillotte, 1997; McCown, 2002a; Hochman and Carberry, 2011; Evans et al., 2017). The Safe RL (Garcıa and Fernández, 2015) is the generalisation of the Risk-aware bandit setting introduced in Section 2.6. In Safe RL or equivalently Risk-aware RL, the learner has the constraint of avoiding catastrophic failures while learning, e.g. Leurent (2020) with autonomous vehicles, which is of prime interest for subsistence agriculture and food security issues. The use of a risk-aware objective for crop operation evaluation currently remains limited. For instance, Taylor et al. (1999) used the coefficient of variation and Baudry et al. (2021) used the CVaR (see Section 2.6) to compare yield distributions.

4.3. Challenges

Crop management has domain-specific constraints for the *in situ* learning process that we detailed in Section 4.1. Each constraint introduces specific challenges for the RL community that must be addressed.

4.3.1. Learning is costly

RL involves active data collection, where actions and their consequences are explored while learning; this is unconventional in agriculture. Experiments in agriculture are expensive, with the duration of a crop cycle allowing only a limited number of experiments. Plausible confounding factors may turn unclear research results on the effects of crop management practices and subsequently require meta-analysis, as exemplified by Giller et al. (2009) in conservation agriculture. During a season, the effect of actions can exhibit long delays, for instance an uneven sowing depth for maize is likely to result in infertile plants due to uneven growth and therefore competition which leads to reduced grain yields. While having shown great progress recently, the learning efficiency and statistical guarantees of RL are still limited (excepted for bandit algorithms). In other words, the amount of data required is generally impracticable for real-world like problems, and the results are uncertain (Hester et al., 2018; Dulac-Arnold et al., 2019).

To speed up an agent's learning, Transfer Learning (see Taylor and Stone, 2009; Weiss et al., 2016) consists of leveraging prior available knowledge for the task to be learned. For instance, in the field of robotics, one does not want to damage the robot while it learns. Therefore, training may first be performed *in silico*, i.e. in simulated conditions, and then transferred to the real-world, though such an approach is not straightforward (Golemo et al., 2018). With Demonstration Learning (Ravichandar et al., 2020), an expert shows an RL agent how to act before the agent learns on its own. Recently, it has been successfully applied in healthcare to perform complex tasks such as myoelectric prosthesis control (Vasan and Pilarski, 2017), and for ophthalmic microsurgery (Keller et al., 2020).

A need for testbeds. In RL, the first step to address real world problems is generally to create simulated environments to explore the use of candidate algorithms. Despite numerous crop models, very few Open Source RL environments for crop management tasks can be found. More crop models should be turned into RL environments to provide a wide range of crop management learning tasks. The OpenAI gym toolkit is a popular Python encapsulation of complex pre-parametrized underlying models turned into easy to manipulate RL environments with a unified interface. Overweg et al. (2021) introduced an OpenAI gym environment, called `CropGym` which is an interface to the Python Crop Simulation Environment (PCSE) `LINTUL3` (Shibu et al., 2010) wheat crop model and features fertilization tasks. Gautron et al. (2022) turned the DSSAT (Hoogenboom et al., 2019) Fortran crop model in a Python OpenAI gym environment, named `gym-DSSAT`, for both maize nitrogen fertilization and irrigation tasks. In contrast to `CropGym`, `gym-DSSAT` features a stochastic weather generator which is DSSAT's default one (Richardson and Wright, 1984).

4.3.2. Actions are only suggestions

In usual RL problems, the agent has direct control over actions made in the environment. Because recommendations are not authoritative instructions there is no guarantee that an agent's choice of action will be consistent with a farmer's decision, which differs from the usual RL problems. As a consequence, an agent cannot freely explore uncertain action effects and cannot directly evaluate its policy. These kinds of settings, known as Compliance Aware Learning, need to be explicitly considered for practical applications. Examples are found in recommender systems or healthcare applications, e.g. Swaminathan and Joachims (2015); Della Penna et al. (2016) with bandit problems, and Sunehag et al. (2015) in an MDP context.

4.3.3. Substantive rationality and utility in RL.

Substantive rational behavior, as defined by Simon (1976), is equivalent to an algorithmic optimization procedure under a set of constraints, performed by an agent to maximize a specific criterion, such as the economic return. However, human decision makers tend to use procedural rationality (Simon, 1955). Farmers seek sub-optimal pragmatic solutions that they can implement, thereby meeting the minimum requirements that were set, such as a minimum yield (Hochman and Carberry, 2011). Farmer's practices are also influenced by social, cultural and economic conditions (Milleville, 1987) and farmer's health (Edwards-Jones, 2006). Deffontaines and Petit (1985) observed that farmers are often not able to provide a clear definition of their own objectives. In contrast, RL intimately relates to the optimization of an explicit utility function which defines the agent's goal. Practitioners should therefore be careful in the inherent limits for characterizing users' decision determinants and bear in mind that any utility function is a proxy (Hochman and Carberry, 2011).

4.3.4. Mathematical formalization

In practice, real world systems are unlikely to strictly follow the stringent assumptions of an MDP (Section 2.2). All the parameters describing a field plot are not accessible. Some of them may not be directly or precisely measurable, or are even currently not studied in the literature. Overall, they are too numerous to be jointly measured and they continuously and autonomously evolve with time. Garcia (1999) observed that their crop management problem did not strictly follow the Markov property. To model a field plot in an RL problem, several extensions relax the assumptions of the canonical MDP. As an example, in a Partially Observable MDP (POMDP, Åström, 1965) the agent does not fully observe the environment's state, but still knows the state space. The agent only accesses observations of the environment that it can use as proxies of the real states (e.g. noisy sensor data). With Predictive State Representation (PSR, Littman et al., 2001) the agent does not fully observe the environment's state, and nor knows the state space. As an alternative modeling, event-Based MDP (EBMDP, Cao, 2008) focus action taking on a limited number of transition events (subsets of state transitions) rather than considering the whole state space. These extensions are still active areas of research. Finally, other research communities addressing sequential decision-making under uncertainty have also developed approaches of potential interest for agriculture. In particular, Ding et al. (2018) dedicated a review to the applications of Predictive Model Control, a sub-field of Optimal Control (see Section 2.3), for agricultural decision-making.

4.3.5. Policy explainability

It seems natural that a decision maker would like to know why one crop management action is preferable to another. DSS require user trust (Rose et al., 2016; Evans et al., 2017). As pointed out by Garcia (1999), RL-learned policies are often not directly usable in practice by agronomists or farmers. Causability is a desirable feature of solutions based on AI as a measure of the quality of explanations (Holzinger et al., 2019). A novel and promising RL research trend is Causal RL (Dasgupta et al., 2019; Madumal et al., 2020). While learning to act, Causal RL makes it possible to discover and take advantage of cause to effect models at a symbolic level, allowing better generalization capabilities between learning problems and counterfactual reasoning (Roese, 1997). In a perspective of practicability, an RL agent's crop management policy should be provided with some high probability future action-taking and expected results (such as expected yields). This seems necessary to allow farmers to compare alternatives and plan real-world actions such as anticipating fertilizer purchases.

4.3.6. A need for multi-scale, multi-objective, resource-constrained RL

Agroecology requires thinking about taking actions at larger temporal and spatial scales than the typical plot and crop-cycle scales because the sustainability of agricultural practices requires multicriteria evaluations (Duru et al., 2015). As examples, crops from surrounding fields may impact local pollinators and/or pest dynamics (Vasseur et al., 2013). So far, most RL algorithms deal with a single, real-valued objective. Based on expert knowledge, practitioners commonly handcraft the MDP return function to express a desirable tradeoff between multiple objectives, and provide localized advice to the agent (Laud, 2004). Multi-objective RL (MORL, Liu et al., 2014) formally addresses the simultaneous optimization of multiple criteria, and is of increasing interest as it relates to many real-world problems. Crop operations are subject to resource constraints (for example, labor, land or input availability) and feasibility conditions (for example, for the soil to have enough load-bearing capacity to use machinery). Resource arbitration at the farm level should ideally also be taken into account.

5. Conclusions

Reinforcement Learning (RL) deals with the problem of sequential decision making under uncertainty, which appears to fit the purpose of supporting crop management. RL is a contextual, geared toward action tool, which seems to share some similarities with how farmers have been described to deal with crop management while considering inherent uncertainty and evaluating joint action sequences. We have envisioned RL as the core of a human-centered support for learning from real experiments at the community level. RL appears to have great potential for agriculture's future challenges, in particular climate change, in a context of increasingly abundant in-field data, computational resources and theoretical advances. However, a joint research effort by the RL and agronomy communities, supported by ergonomists, is required to turn concepts into practicable tools.

A review of RL applied to crop management has revealed that efforts to apply RL in the agronomy community have so far been limited. A probable explanation is that crop management presents a set of domain-specific practical and theoretical challenges. Decision support cannot be reduced to an algorithmic optimization procedure, user objectives and constraints should be carefully taken into account. Furthermore, data is scarce and costly, and taking the wrong action can be deleterious, especially from a food security perspective. We identified as theoretical challenges how to efficiently learn; how to model crop management decision problems; how to learn explainable crop management policies; how to learn problems with multiple objectives under resource constraints. The multi-armed bandit framework appears one of the most suitable RL approaches for *in situ* learning due to its limited sample complexity and the versatility of the settings found in the literature.

Declaration of competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work has been supported by:

- The French Agricultural Research Centre for International Development (CIRAD).
- The Consultative Group for International Agricultural Research (CGIAR) Platform for Big Data in Agriculture. Special thanks to Brian King.
- The French Ministry of Higher Education and Research, Hauts-de-France region, Inria within the Scool team project and MEL.

The authors would like to thank Marianne Cerf, Ronan Trépos, Eric Penot and Mathieu Seurin for their comments that helped to improve the manuscript. We also thank Andrew Lewer for proofreading.

Glossary

action How the environment's dynamics are controlled by the agent.

action space The set of possible actions.

agent The entity that acts on the environment in order to optimize the objective function.

Deep Neural Network Neural network with several layers. In RL, this number of layers is limited (from a few to say a dozen layers) whereas in machine learning, there may be hundreds and even thousands of layers.

environment The object with which the agent interacts.

episode A single sequence of interactions of the agent with the environment, from a given initial state.

Exploration/Exploitation dilemma The situation in which an agent has the choice between performing an action with consequences which are known (exploitation) and an action with consequences which are unknown (exploration).

horizon Maximum number of time steps of an episode.

in silico A virtual experience.

Internet of Things (IoT) Networks of uniquely identified physical devices which can autonomously communicate between themselves or with humans, and process data.

Markov Decision Process Mathematical formalization of the environment in a Reinforcement Learning problem, see Figure 5.

Neural Networks A neural network is made up of a set of layers of simple computation units, called neurons. Each neuron receives data as input and outputs one or more labels (usually either symbolic, or numeric). Mathematically speaking, a neural network is a function.

objective function The function that the agent optimizes by controlling the environment.

observation In an MDP, a snapshot of the environment's state. In the general case, there is no assumption that an optimal action may be determined using an observation.

overfitting A Machine Learning model that has been trained and performs well in training situations but performs poorly in unseen situations.

policy A function that indicates how the agent acts depending on the environment's state.

quality function The expected value of the objective function when the environment is in a given state and the agent first performs a given action and then follows a given policy.

return A positive or negative stimulus provided by the environment to the agent which indicates if the past actions have been beneficial to the agent with regards to its objective.

sample complexity Number of samples required to solve a problem. The higher the sample complexity, the harder the problem.

state A set of descriptors of the environment that is sufficient to decide on an optimal action.

state space The set of possible states.

stationary A random process in which distributions do not change over time.

value function The expected value of the objective function when the environment is in a given state and the agent follows a given policy.

References

Myriam Adam, Dilys Sefakor MacCarthy, Pierre C Sibiry Traoré, Andree Nenkam, Bright Salah Freduah, Mouhamed Ly, and Samuel GK Adiku. Which is more important to sorghum production systems in the sudano-sahelian zone of west africa: Climate change or improved management practices? *Agricultural Systems*, 185:102920, 2020.

Miguel A Altieri, Clara I Nicholls, Alejandro Henao, and Marcos A Lana. Agroecology and the design of climate change-resilient farming systems. *Agronomy for sustainable development*, 35(3):869–890, 2015.

David Arnott and Graham Pervan. A critical analysis of decision support systems research. *Journal of information technology*, 20(2):67–87, 2005.

Karl Johan Åström. Optimal control of markov processes with incomplete state information. *Journal of Mathematical Analysis and Applications*, 10(1):174–205, 1965.

Hagit Attiya and Jennifer Welch. *Distributed computing: fundamentals, simulations, and advanced topics*, volume 19. John Wiley & Sons, 2004.

Peter Auer and Ronald Ortner. Logarithmic online regret bounds for undiscounted reinforcement learning. *Advances in neural information processing systems*, 19, 2006.

Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. *Advances in neural information processing systems*, 21, 2008.

Alexandre Barbosa, Rodrigo Trevisan, Naira Hovakimyan, and Nicolas F Martin. Modeling yield response to crop management using convolutional neural networks. *Computers and Electronics in Agriculture*, 170: 105197, 2020.

Dorian Baudry, Romain Gautron, Emilie Kaufmann, and Odalric Maillard. Optimal thompson sampling strategies for support-aware cvar bandits. In *International Conference on Machine Learning*, pages 716–726. PMLR, 2021.

Richard Bellman. Dynamic programming. *Princeton, USA: Princeton University Press*, 1(2):3, 1957.

JE Bergez, M Eigenraam, and F Garcia. Comparison between dynamic programming and reinforcement learning: A case study on maize irrigation management. In *Proceedings of the 3rd European Conference on Information Technology in Agriculture (EFITA01), Montpellier (FR) pp*, pages 343–348. Citeseer, 2001.

Dimitri P Bertsekas and Steven E Shreve. *Stochastic optimal control: the discrete-time case*, volume 5. Athena Scientific, 1996.

Dimitri P Bertsekas and John N Tsitsiklis. *Neuro-dynamic programming*. Athena Scientific, 1996.

Jonathan Binas, Leonie Luginbuehl, and Yoshua Bengio. Reinforcement learning for sustainable agriculture. In *ICML Workshop Climate Change: How Can AI Help?*, 2019.

I Boiffin, E Malezieuxz, and D Picard. Cropping systems for the 1 6 future. *Crop science: Progress and prospects*, page 261, 2001.

Onur Boyabatlı, Javad Nasiry, and Yangfang Zhou. Crop planning in sustainable agriculture: Dynamic farmland allocation in the presence of crop rotation benefits. *Management Science*, 65(5):2060–2076, 2019.

Nadine Brisson, Christian Gary, Eric Justes, Romain Roche, Bruno Mary, Dominique Ripoche, Daniel Zimmer, Jorge Sierra, Patrick Bertuzzi, Philippe Burger, et al. An overview of the crop model stics. *European Journal of agronomy*, 18(3-4):309–332, 2003.

Fanyu Bu and Xin Wang. A smart agriculture iot system based on deep reinforcement learning. *Future Generation Computer Systems*, 99:500–507, 2019.

Oscar R Burt and John R Allison. Farm management decisions with dynamic programming. *Journal of Farm Economics*, 45(1):121–136, 1963.

Xi-Ren Cao. Limitation of markov models and event-based learning and optimization. In *2008 Chinese*

Control and Decision Conference, pages 14–17. IEEE, 2008.

Asaf Cassel, Shie Mannor, and Assaf Zeevi. A general approach to multi-armed bandits under risk criteria. In *Conference On Learning Theory*, pages 1295–1306. PMLR, 2018.

Marianne Cerf and Jean-Marc Meynard. Les outils de pilotage des cultures: diversité de leurs usages et enseignements pour leur conception. *Natures Sciences Sociétés*, 14(1):19–29, 2006.

Marianne Cerf and M Sebillotte. Le concept de modele general et la prise de decision dans la conduite d'une culture. *Comptes Rendus de l'Académie d'Agriculture de France 4 (74), 71-80.(1988)*, 1988.

Marianne Cerf and Michel Sebillotte. Approche cognitive des décisions de production dans l'exploitation agricole [confrontation aux théories de la décision]. *Economie rurale*, 239(1):11–18, 1997.

Marie-Hélène Chatelin, Christine Aubry, P Leroy, F Papy, and Jean-Christophe Poussin. Pilotage de la production et aide à la décision stratégique: le cas des exploitations en grande culture. *Cahiers d'Economie et de Sociologie Rurales (CESR)*, 28(905-2016-70228):119–138, 1993.

Mengting Chen, Yuanlai Cui, Xiaonan Wang, Hengwang Xie, Fangping Liu, Tongyuan Luo, Shizong Zheng, and Yufeng Luo. A reinforcement learning approach to irrigation decision-making for rice using weather forecasts. *Agricultural Water Management*, 250:106838, 2021.

Vimbayi Grace Petrova Chimonyo, Albert Thembinkosi Modi, and Tafadzwanashe Mabhaudhi. Perspective on crop modelling in the management of intercropping systems. *Archives of Agronomy and Soil Science*, 61(11):1511–1529, 2015.

Benjamin I Cook, Justin S Mankin, and Kevin J Anchukaitis. Climate change and drought: From past to future. *Current Climate Change Reports*, 4(2):164–179, 2018.

Rémi Coulom. Efficient selectivity and backup operators in monte-carlo tree search. In *International conference on computers and games*, pages 72–83. Springer, 2006.

George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

Carlos Gomes da Silva, José Figueira, João Lisboa, and Samir Barman. An interactive decision support system for an aggregate production planning model based on multiple criteria mixed integer linear programming. *Omega*, 34(2):167–177, 2006.

Ishita Dasgupta, Jane Wang, Silvia Chiappa, Jovana Mitrovic, Pedro Ortega, David Raposo, Edward Hughes, Peter Battaglia, Matthew Botvinick, and Zeb Kurth-Nelson. Causal reasoning from meta-reinforcement learning. *arXiv preprint arXiv:1901.08162*, 2019.

Jean-Pierre Deffontaines and Michel Petit. *Comment étudier les exploitations agricoles d'une région: présentation d'un ensemble méthodologique*. INRA, 1985.

Nicolás Della Penna, Mark D Reid, and David Balduzzi. Compliance-aware bandits. *arXiv preprint arXiv:1602.02852*, 2016.

Ying Ding, Liang Wang, Yongwei Li, and Daoliang Li. Model predictive control and its application in agriculture: A review. *Computers and Electronics in Agriculture*, 151:104–117, 2018.

Marcello Donatelli, Roger D Magarey, Simone Bregaglio, L Willocquet, Jérémy PM Whish, and Serge Savary. Modelling the impacts of pests and diseases on agricultural systems. *Agricultural systems*, 155:213–224, 2017.

Gabriel Dulac-Arnold, Daniel Mankowitz, and Todd Hester. Challenges of real-world reinforcement learning. *arXiv preprint arXiv:1904.12901*, 2019.

Audrey Durand, Charis Achilleos, Demetris Iacovides, Katerina Strati, Georgios D Mitsis, and Joelle Pineau. Contextual bandits for adapting treatment in a mouse model of de novo carcinogenesis. In *Machine learning for healthcare conference*, pages 67–82. PMLR, 2018.

Michel Duru, Olivier Therond, Guillaume Martin, Roger Martin-Clouaire, Marie-Angélina Magne, Eric Justes, Etienne-Pascal Journet, Jean-Noël Aubertot, Serge Savary, Jacques-Eric Bergez, et al. How to implement biodiversity-based agriculture to enhance ecosystem services: a review. *Agronomy for sustainable development*, 35(4):1259–1281, 2015.

Jérôme Dury, Noémie Schaller, Frédérick Garcia, Arnaud Reynaud, and Jacques Eric Bergez. Models to support cropping plan and crop rotation decisions. a review. *Agronomy for sustainable development*, 32 (2):567–580, 2012.

Gareth Edwards-Jones. Modelling farmer decision-making: concepts, progress and challenges. *Animal science*, 82(6):783–790, 2006.

James E Epperson, James E Hook, and Yasmin R Mustafa. Dynamic programming for improving irrigation scheduling strategies of maize. *Agricultural Systems*, 42(1-2):85–101, 1993.

Katherine J Evans, Andrew Terhorst, and Byeong Ho Kang. From data to decisions: helping crop producers build their actionable knowledge. *Critical reviews in plant sciences*, 36(2):71–88, 2017.

Gatien N Falconnier, Marc Corbeels, Kenneth J Boote, François Affholder, Myriam Adam, Dilys S MacCarthy, Alex C Ruane, Claas Nendel, Anthony M Whitbread, Éric Justes, et al. Modelling climate change impacts on maize yields under low nitrogen input conditions in sub-saharan africa. *Global change biology*, 26(10): 5942–5964, 2020.

Luis Ferreira and Martin H Murray. Modelling rail track deterioration and maintenance: current practices and future needs. *Transport Reviews*, 17(3):207–221, 1997.

Rudolf J Freund. The introduction of risk into a programming model. *Econometrica: Journal of the econometric society*, pages 253–263, 1956.

Frédérick Garcia. Use of reinforcement learning and simulation to optimize wheat crop technical management. In *Proceedings of the International Congress on Modelling and Simulation (MODSIM'99) Hamilton, New-Zealand*, pages 801–806, 1999.

Frédérick Garcia and Seydina M Ndiaye. A learning rate analysis of reinforcement learning algorithms in finite-horizon. In *Proceedings of the 15th International Conference on Machine Learning (ML-98*. Citeseer, 1998.

Javier Garcıa and Fernando Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480, 2015.

Romain Gautron, Emilio J. Padrón, Philippe Preux, Julien Bigot, Odalric-Ambrym Maillard, and David Emukpere. gym-DSSAT: a crop model turned into a Reinforcement Learning environment. Research Report RR-9460, Inria Lille, July 2022. URL `https://hal.inria.fr/hal-03711132`.

David H Gent, Erick De Wolf, and Sarah J Pethybridge. Perceptions of risk, risk aversion, and barriers to adoption of decision support systems and integrated pest management: an introduction. *Phytopathology*, 101(6):640–643, 2011.

Ken E Giller, Ernst Witter, Marc Corbeels, and Pablo Tittonell. Conservation agriculture and smallholder farming in africa: the heretics' view. *Field crops research*, 114(1):23–34, 2009.

John J Glen. Mathematical models in farm planning: A survey. *Operations Research*, 35(5):641–666, 1987.

Florian Golemo, Adrien Ali Taiga, Aaron Courville, and Pierre-Yves Oudeyer. Sim-to-real transfer with neural-augmented robot simulation. In *Conference on Robot Learning*, pages 817–828. PMLR, 2018.

Frédéric Goulet, Franck Pervanchon, Cédric Conteau, and Marianne Cerf. Les agriculteurs innovent par eux-mêmes pour leurs systèmes de culture. *R. Reau et T. Doré, Systèmes de culture innovant et durables. Dijon, educagri éditions*, pages 53–69, 2008.

Hakan Guler. Decision support system for railway track maintenance and renewal management. *Journal of Computing in Civil Engineering*, 27(3):292–306, 2013.

Cédric Hartland, Sylvain Gelly, Nicolas Baskiotis, Olivier Teytaud, and Michele Sebag. Multi-armed bandit, dynamic environments and meta-bandits. 2006.

Brian Hayes. Cloud computing, 2008.

Jianqiang He, Michael D Dukes, George J Hochmuth, James W Jones, and Wendy D Graham. Identifying irrigation and nitrogen best management practices for sweet corn production on sandy soils using ceres-maize model. *Agricultural Water Management*, 109:61–70, 2012.

J Hébert. La fumure azotée du blé tendre d'hiver. *Bull Tech Inf*, 244:755–766, 1969.

Todd Hester, Matej Vecerik, Olivier Pietquin, Marc Lanctot, Tom Schaul, Bilal Piot, Dan Horgan, John Quan, Andrew Sendonaris, Ian Osband, et al. Deep q-learning from demonstrations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Clifford Hildreth. Problems of uncertainty in farm planning. *Journal of Farm Economics*, 39(5):1430–1441, 1957.

Zvi Hochman and PS Carberry. Emerging consensus on desirable characteristics of tools to support farmers' management of climate risk in australia. *Agricultural Systems*, 104(6):441–450, 2011.

Andreas Holzinger, Georg Langs, Helmut Denk, Kurt Zatloukal, and Heimo Müller. Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(4):e1312, 2019.

Annelie Holzkämper, Pierluigi Calanca, and Jeffrey Fuhrer. Identifying climatic limitations to grain maize yield potentials using a suitability evaluation approach. *Agricultural and Forest Meteorology*, 168:149–159, 2013.

G Hoogenboom, CH Porter, KJ Boote, V Shelia, PW Wilkens, U Singh, JW White, S Asseng, JI Lizaso, LP Moreno, et al. The dssat crop modeling ecosystem. *Advances in crop modelling for a sustainable agriculture*, pages 173–216, 2019.

Olivier Husson, Jean-Pierre Sarthou, Lydia Bousset, Alain Ratnadass, Hans-Peter Schmidt, John Kempf, Benoit Husson, Sophie Tingry, Jean-Noël Aubertot, Jean-Philippe Deguine, François-Régis Goebel, and

Jay Ram Lamichhane. Soil and plant health in relation to dynamic sustainment of eh and ph homeostasis: A review. *Plant and Soil*, Jul 2021. ISSN 1573-5036. doi: 10.1007/s11104-021-05047-z. URL https://doi.org/10.1007/s11104-021-05047-z.

Ryan HL Ip, Li-Minn Ang, Kah Phooi Seng, JC Broster, and JE Pratley. Big data and machine learning for crop protection. *Computers and Electronics in Agriculture*, 151:376–383, 2018.

James W Jones, John M Antle, Bruno Basso, Kenneth J Boote, Richard T Conant, Ian Foster, H Charles J Godfray, Mario Herrero, Richard E Howitt, Sander Janssen, et al. Brief history of agricultural systems modeling. *Agricultural systems*, 155:240–254, 2017.

Brenton Keller, Mark Draelos, Kevin Zhou, Ruobing Qian, Anthony N Kuo, George Konidaris, Kris Hauser, and Joseph A Izatt. Optical coherence tomography-guided robotic ophthalmic microsurgery via reinforcement learning from demonstration. *IEEE Transactions on Robotics*, 36(4):1207–1218, 2020.

John OS Kennedy. *Dynamic programming: applications to agriculture and natural resources*. Springer Science & Business Media, 1986.

AM Khaliq, M Javed, M Sohail, and Muhammad Sagheer. Environmental effects on insects and their population dynamics. *Journal of Entomology and Zoology studies*, 2(2):1–7, 2014.

Johannes Kirschner and Andreas Krause. Stochastic bandits with context distributions. *arXiv preprint arXiv:1906.02685*, 2019.

Levente Kocsis and Csaba Szepesvári. Bandit based monte-carlo planning. In *European conference on machine learning*, pages 282–293. Springer, 2006.

Harold J Kushner. Stochastic stability and control. Technical report, Brown Univ Providence RI, 1967.

Maxim Lapan. *Deep Reinforcement Learning Hands-On: Apply modern RL methods, with deep Q-networks, value iteration, policy gradients, TRPO, AlphaGo Zero and more*. Packt Publishing Ltd, 2018.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Adam Daniel Laud. *Theory and application of reward shaping in reinforcement learning*. University of Illinois at Urbana-Champaign, 2004.

P-Y Le Gal, Anne Merot, C-H Moulin, Mireille Navarrete, and Jacques Wery. A modelling framework to support farmers in designing agricultural production systems. *Environmental Modelling & Software*, 25 (2):258–268, 2010.

Hal Lemmon. Comax: An expert system for cotton crop management. *Science*, 233(4759):29–33, 1986.

Edouard Leurent. *Safe and Efficient Reinforcement Learning for Behavioural Planning in Autonomous Driving*. PhD thesis, Université de Lille, 2020.

Yuxi Li. Reinforcement learning applications. *arXiv preprint arXiv:1908.06973*, 2019.

Michael L Littman, Richard S Sutton, and Satinder P Singh. Predictive representations of state. In *NIPS*, volume 14, page 30, 2001.

Chunming Liu, Xin Xu, and Dewen Hu. Multiobjective reinforcement learning: A comprehensive overview. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 45(3):385–398, 2014.

Fang Liu, Joohyun Lee, and Ness Shroff. A change-detection based framework for piecewise-stationary multi-armed bandit problem. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Timothy J Lowe and Paul V Preckel. Decision technologies for agribusiness problems: A brief review of selected literature and a call for research. *Manufacturing & Service Operations Management*, 6(3):201–208, 2004.

Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. Explainable reinforcement learning through a causal lens. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 2493–2500, 2020.

Benoit B Mandelbrot. The variation of certain speculative prices. In *Fractals and scaling in finance*, pages 371–418. Springer, 1997.

Daniel J Mankowitz, Dan A Calian, Rae Jeong, Cosmin Paduraru, Nicolas Heess, Sumanth Dathathri, Martin Riedmiller, and Timothy Mann. Robust constrained reinforcement learning for continuous control with model misspecification. *arXiv preprint arXiv:2010.10644*, 2020.

Basil D Manos, Adriano Ciani, Thomas Bournaris, I Vassiliadou, and J Papathanasiou. A taxonomy survey of decision support systems in agriculture. *Agricultural Economics Review*, 5(389-2016-23416):80–94, 2004.

RL McCown, GL Hammer, JNG Hargreaves, D Holzworth, and NI Huth. Apsim: an agricultural production system simulation model for operational research. *Mathematics and computers in simulation*, 39(3-4): 225–231, 1995.

Robert L McCown. Changing systems for supporting farmers' decisions: problems, paradigms, and prospects. *Agricultural systems*, 74(1):179–220, 2002a.

Robert L McCown. Locating agricultural decision support systems in the troubled past and socio-technical complexity of 'models for management'. *Agricultural systems*, 74(1):11–25, 2002b.

Uwe Meier. *Growth stages of mono-and dicotyledonous plants*. Blackwell Wissenschafts-Verlag, 1997.

Joseph Mellor and Jonathan Shapiro. Thompson sampling in switching environments with bayesian online change point detection. *arXiv preprint arXiv:1302.3721*, 2013.

Randolph A Miller. Diagnostic decision support systems. In *Clinical decision support systems*, pages 181–208. Springer, 2016.

Pierre Milleville. Recherches sur les pratiques des agriculteurs. *Les cahiers de la Recherche Développement*, 16:3–7, 1987.

Tom M Mitchell et al. Machine learning. 1997. *Burr Ridge, IL: McGraw Hill*, 45(37):870–877, 1997.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

Rémi Munos. A convergent reinforcement learning algorithm in the continuous case : the finite-element reinforcement learning. In *International Conference on Machine Learning*. Morgan Kaufmann, 1996.

Honorio Navarro-Hellín, Jesus Martinez-del Rincon, Rafael Domingo-Miguel, Fulgencio Soto-Valles, and Roque Torres-Sánchez. A decision support system for managing irrigation in agriculture. *Computers and Electronics in Agriculture*, 124:121–131, 2016.

Seydina Moussa Ndiaye. *Apprentissage par renforcement en horizon fini: application à la génération de règles* *pour la Conduite de Culture*. PhD thesis, Toulouse 3, 1999.

RA Nelson, DP Holzworth, GL Hammer, and PT Hayman. Infusing the use of seasonal climate forecasting into crop management practice in north east australia using discussion support software. *Agricultural Systems*, 74(3):393–414, 2002.

Roger D Norton and Peter BR Hazell. *Mathematical programming for economic analysis in agriculture*. Macmillan New York, NY, USA, 1986.

Martin J Osborne et al. *An introduction to game theory*, volume 3. Oxford university press New York, 2004.

Hiske Overweg, Herman NC Berghuijs, and Ioannis N Athanasiadis. Cropgym: a reinforcement learning environment for crop management. *arXiv preprint arXiv:2104.04326*, 2021.

Sindhu Padakandla, KJ Prabuchandran, and Shalabh Bhatnagar. Reinforcement learning algorithm for non-stationary environments. *Applied Intelligence*, 50(11):3590–3606, 2020.

François Papy. Savoir pratique sur les systèmes techniques et aide à la décision. *La conduite du champ cultivé. Points de vue d'agronomes. IRD*, pages 245–259, 1998.

Daniel J Power. Decision support systems: a historical overview. In *Handbook on decision support systems 1*, pages 121–140. Springer, 2008.

Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 1994.

Eric Quinton, Cécile Pignol, Hector Linyer, Julin Ancelin, Sébastien Cipière, Wilfried Heintz, Mathias Rouan, Sylvie Damy, Vincent Bretagnolle, et al. Towards better traceability of field sampling data. *Computers & Geosciences*, 129:82–91, 2019.

Harish Ravichandar, Athanasios S Polydoros, Sonia Chernova, and Aude Billard. Recent advances in robot learning from demonstration. *Annual Review of Control, Robotics, and Autonomous Systems*, 3:297–330, 2020.

Clarence W Richardson and David A Wright. Wgen: A model for generating daily weather variables. *ARS (USA)*, 1984.

Herbert Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535, 1952.

Neal J Roese. Counterfactual thinking. *Psychological bulletin*, 121(1):133, 1997.

David C Rose, William J Sutherland, Caroline Parker, Matt Lobley, Michael Winter, Carol Morris, Susan Twining, Charles Ffoulkes, Tatsuya Amano, and Lynn V Dicks. Decision support tools for agriculture: Towards effective design and delivery. *Agricultural systems*, 149:165–174, 2016.

Karen Rose, Scott Eldridge, and Lyman Chapin. The internet of things: An overview. *The internet society (ISOC)*, 80:1–50, 2015.

FS Royce, JW Jones, and JW Hansen. Model–based optimization of crop management for climate forecast applications. *Transactions of the ASAE*, 44(5):1319, 2001.

Sajad Sabzi, Yousef Abbaspour-Gilandeh, and Gines Garcia-Mateos. A fast and accurate expert system for weed identification in potato crops using metaheuristic algorithms. *Computers in Industry*, 98:80–89, 2018.

Yuji Saikai, Vivak Patel, and Paul D Mitchell. Machine learning for optimizing complex site-specific management. *Computers and Electronics in Agriculture*, 174:105381, 2020.

Anton Schwartz. A reinforcement learning method for maximizing undiscounted rewards. In *Proceedings of the tenth international conference on machine learning*, volume 298, pages 298–305, 1993.

Michel Sebillotte. Agronomie et agriculture. essai d'analyse des tâches de l'agronome. *Cahiers Orstom, série biologie*, 24:3–25, 1974.

Michel Sebillotte. Itinéraires techniques et évolution de la pensée agronomique. *CR Acad. Agric. Fr*, 64(11): 906–914, 1978.

Michel Sebillotte and Louis Georges Soler. Le concept de modele general et la comprehension du comportement de l'agriculteur. 1988.

Muhammed E Shibu, Peter A Leffelaar, Herman Van Keulen, and Pramod K Aggarwal. Lintul3, a simulation model for nitrogen-limited situations: Application to rice. *European Journal of Agronomy*, 32(4):255–271, 2010.

Rosemary Shrestha, Elizabeth Arnaud, Ramil Mauleon, Martin Senger, Guy F Davenport, David Hancock, Norman Morrison, Richard Bruskiewich, and Graham McLaren. Multifunctional crop trait ontology for breeders' data: field book, annotation, data discovery and semantic enrichment of the literature. *AoB plants*, 2010, 2010.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *Nature*, 550(7676):354, 2017.

Herbert A Simon. A behavioral model of rational choice. *The quarterly journal of economics*, 69(1):99–118, 1955.

Herbert A Simon. From substantive to procedural rationality. In *25 years of economic theory*, pages 65–86. Springer, 1976.

Mette Sønderskov, Per Rydahl, Ole M Bøjer, Jens Erik Jensen, and Per Kudsk. Crop protection online—weeds: a case study for agricultural decision support systems. In *Real-World Decision Support Systems*, pages 303–320. Springer, 2016.

George Stanford. Rationale for optimum nitrogen fertilization in corn production. *Journal of Environmental Quality*, 2(2):159–166, 1973.

Lijia Sun, Yanxiang Yang, Jiang Hu, Dana Porter, Thomas Marek, and Charles Hillyer. Reinforcement learning control for water-efficient agricultural irrigation. In *2017 IEEE International Symposium on Parallel and Distributed Processing with Applications and 2017 IEEE International Conference on Ubiquitous Computing and Communications (ISPA/IUCC)*, pages 1334–1341. IEEE, 2017.

Peter Sunehag, Richard Evans, Gabriel Dulac-Arnold, Yori Zwols, Daniel Visentin, and Ben Coppin. Deep reinforcement learning with attention for slate markov decision processes with high-dimensional states and actions. *arXiv preprint arXiv:1512.01124*, 2015.

Adith Swaminathan and Thorsten Joachims. Counterfactual risk minimization: Learning from logged bandit feedback. In *International Conference on Machine Learning*, pages 814–823. PMLR, 2015.

Michael E Sykuta. Big data in agriculture: property rights, privacy and competition in ag data services. *International Food and Agribusiness Management Review*, 19(1030-2016-83141):57–74, 2016.

Matthew E Taylor and Peter Stone. Transfer learning for reinforcement learning domains: A survey. *Journal of Machine Learning Research*, 10(7), 2009.

SL Taylor, ME Payton, and WR Raun. Relationship between mean yield, coefficient of variation, mean square error, and plot size in wheat field experiments. *Communications in Soil Science and Plant Analysis*, 30 (9-10):1439–1447, 1999.

Gerald Tesauro. Temporal difference learning and td-gammon. *Communications of the ACM*, 38(3):58–68, 1995.

William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

PJ Thorburn, E Jakku, AJ Webster, and YL Everingham. Agricultural decision support systems facilitating co-learning: a case study on environmental impacts of sugarcane production. *International Journal of Agricultural Sustainability*, 9(2):322–333, 2011.

Gerhard Tintner. Stochastic linear programming with applications to agricultural economics. In *Proceedings of the Second Symposium in Linear Programming*, volume 1, pages 197–228. National Bureau of Standards Washington, DC, 1955.

Ronan Trépos, Stephane Lemarié, Helene Raynal, Muriel Morison, Stéphane Couture, and Frederick Garcia. Apprentissage par renforcement pour l'optimisation de la conduite de culture du colza. In *JFPDA'14. Journées Francophones Planification, Décision, Apprentissage pour la conduite de système.*, pages 1–13, 2014.

Gautham Vasan and Patrick M Pilarski. Learning from demonstration: Teaching a myoelectric prosthesis with an intact limb via reinforcement learning. In *2017 International Conference on Rehabilitation Robotics (ICORR)*, pages 1457–1464. IEEE, 2017.

Chloé Vasseur, Alexandre Joannon, Stéphanie Aviron, Françoise Burel, Jean-Marc Meynard, and Jacques Baudry. The cropping systems mosaic: how does the hidden heterogeneity of agricultural landscapes drive arthropod populations? *Agriculture, ecosystems & environment*, 166:3–14, 2013.

Harshal Waghmare, Radha Kokare, and Yogesh Dandawate. Detection and classification of diseases of grape plant using opposite colour local binary pattern feature and machine learning for automated decision support system. In *2016 3rd international conference on signal processing and integrated networks (SPIN)*, pages 513–518. IEEE, 2016.

Lu Wang, Xiaofeng He, and Dijun Luo. Deep reinforcement learning for greenhouse climate control. In *2020 IEEE International Conference on Knowledge Graph (ICKG)*, pages 474–480. IEEE, 2020.

Christopher John Cornish Hellaby Watkins. Learning from delayed rewards. 1989.

Andrés Weintraub and Carlos Romero. Operations research models and the management of agricultural and forestry resources: a review and comparison. *Interfaces*, 36(5):446–457, 2006.

Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big data*, 3 (1):1–40, 2016.

Ronald J Williams. *Reinforcement-learning connectionist systems*. College of Computer Science, Northeastern University, 1987.

Yanxiang Yang, Jiang Hu, Dana Porter, Thomas Marek, Kevin Heflin, and Hongxin Kong. Deep reinforcement learning-based irrigation scheduling. *Transactions of the ASABE*, 63(3):549–556, 2020.

Yu Zhang and Qiang Yang. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, 2021.