THE EUROPEAN SOCIETY OF MEDICINE
Medical Research Archives, Volume 13 Issue 3

RESEARCH ARTICLE

Strategies and Considerations for Safe Reinforcement Learning in Programming Cardiac Implantable Electronic Devices

John Komp[1], Aaptha Boggaram[1], David P. Kao[2], Ashutosh Trivedi[1], Michael A. Rosenberg[2]

[1] College of Engineering and Applied Science, University of Colorado, Boulder, CO, USA
[2] Division of Cardiology, University of Colorado Anschutz Medical Campus, Aurora, CO, USA

OPEN ACCESS

PUBLISHED
31 March 2025

CITATION
Kamp, J., Boggaram, A., et al., 2025. -Strategies and Considerations for Safe Reinforcement Learning in Programming Cardiac Implantable Electronic Devices. Medical Research Archives, [online] 13(3).
https://doi.org/10.18103/mra.v13i3.6363

DOI
https://doi.org/10.18103/mra.v13i3.6363

ISSN
2375-1924

ABSTRACT

The programming of cardiac implantable electronic devices, such as pacemakers and implantable defibrillators, represents a promising domain for the application of automated learning systems. These systems, leveraging a type of artificial intelligence called reinforcement learning, have the potential to personalize medical treatment by adapting device settings based on an individual's physiological responses. At the core of these self-learning algorithms is the principle of balancing exploration and exploitation. Exploitation refers to the selection of device programming settings previously demonstrated to provide clinical benefit, while exploration refers to the real-time search for adjustments to device programming that could provide an improvement in clinical outcomes for each individual. Exploration is a critical component of the reinforcement learning algorithm, and provides the opportunity to identify settings that could directly benefit individual patients. However, unconstrained exploration poses risks, as an automated change in certain settings may lead to adverse clinical outcomes. To mitigate these risks, several strategies have been proposed to ensure that algorithm-driven programming changes achieve the desired level of individualized optimization without compromising patient safety. In this review, we examine the existing literature on safe reinforcement learning algorithms in automated systems and discuss their potential application to the programming of cardiac implantable electronic devices.

A. Introduction

In the past decade, innovation in data science has proceeded at a tremendous pace as larger amounts of data become available for development of increasingly sophisticated machine-learning algorithms. This innovation has brought the promise that treatment approaches, such as programming of cardiac implantable electronic devices (CIEDs), can be tailored dynamically to the benefit of each individual. A type of machine-learning model framework that is increasingly being applied in automated systems is reinforcement learning (RL), whereby the algorithm itself retains the agency to identify better decisions through dynamic interaction with the environment. Utilizing a state-action-reward decision framework[1], the RL algorithm learns the best action to achieve reward, which is individualized based on the state information, in this case the patient's condition (Fig. 1). Unlike a typical clinical decision, in which a human clinician selects the action, in a RL framework the action is selected by a computer algorithm. Reinforcement learning algorithms have been applied successfully in robotics[2,3], automated video game applications[4,5], on-line map applications like Google Maps or Waze, internet search optimization[6], as well as recently in certain medical applications[7-10].

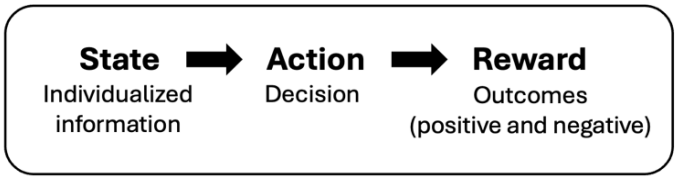

Figure 1. Decision framework. Decisions (action) are made for individuals based on varying state information, toward the goal of achieving a reward, which includes both positive and negative outcomes that can result from the decision.

Reinforcement learning methods are most powerful in settings where the data elements that comprise the individual components of the model are already available in digital formats, such as a CIED that, in addition to providing pacing or defibrillation, also collects various physiological measures. For example, the measure of daily activity derived from the accelerometer of an implanted pacemaker can be used to optimize the pacing protocol for a patient who requires atrial pacing during exertion due to insufficient increase in heart rate (termed *chronotropic incompetence*, see below). If the device-collected daily activity is already available within the storage system of the pacemaker, changes to the pacing protocol can be evaluated prospectively by the device without the need for follow-up exercise testing or collection of patient survey data. As such, the potential for integration of RL methods into programming of CIEDs has great potential to improve outcomes for the various clinical conditions treated with CIED implantation.

However, as is inherent in RL algorithms, the process of ensuring that the reward achieved represents the global optimum rather than a local optimum requires a trade-off between exploitation of existing rewards under the present algorithm, and exploration of new actions that could reach higher levels of reward. While the potential benefit of finding a novel action might be evident, there are risks if this exploration step leads to dangerous or harmful outcomes. *Safe RL* refers to the set of approaches that can be applied to balance exploration and exploitation in such a way that an algorithm can still explore new actions, but with mitigation of risks in exploration. In this review, we outline how safe RL can be applied in models of CIED programming. We start with a brief description of some of the clinical scenarios where programming of CIEDs has an impact on outcomes for certain conditions. We then explain how RL can be utilized in programming CIEDs to improve outcomes for individuals with these conditions, including a brief discussion of the RL framework itself. Finally, we describe the approaches to safe RL that would be needed in these applications, highlighting both technical (quantitative) methodology as well as practical considerations that would need to be addressed for these approaches to be used in practice.

B. Programming decisions for CIEDs based on clinical condition

From the time of the first implanted pacemaker in 1958[11] until present, CIEDs have been increasingly utilized to manage a number of cardiac conditions. While the original pacemakers focused on restoring conduction to patients with asystole or heart block, latter CIEDs have been developed to treat patients with chronotropic incompetence, heart failure, and lethal ventricular arrhythmias. Inherent in all clinical uses is the need to modify programming parameters based on perceived clinical response, but a generalizable strategy for CIED optimization has not been identified. As such, devices are often programmed using standardized settings which may not be optimal for some, if not most patients. Presently, these modifications have been made on two levels: 1) internally, through static changes made by the manufacturer to the rules-based algorithms (i.e., updating the internal pacing algorithm across a version of the device), or 2) externally, by treating clinicians changing standard settings during follow-up clinic visits. Incorporation of next-generation methods of artificial intelligence, such as RL, would bypass both existing methods to design a dynamic, internalized learning algorithm within the CIED, which learns the 'best' settings for each individual patient. In order to develop such algorithms, an understanding of what specific programming changes have already been identified as having clinical impact is required. In this section, we highlight four clinical situations where CIEDs are utilized (Table 1), and outline the broad programming changes that have been investigated for improvement in clinical impact.

Table 1. Examples of CIED programming for clinical conditions. See text for details.

	Chronotropic Incompetence (CI)	**Heart Failure with Preserved Ejection Fraction (HFpEF)**	**Cardiac Resynchronization in Heart Failure with Reduced Ejection Fraction (HFrEF)**	**Detection and Therapy for Ventricular Tachycardia (VT) and Fibrillation (VF)**
State (indication)	Inadequate function of the SAN	Impaired ventricular filling	Cardiac dyssynchrony (left bundle branch block)	Presence of VT or VF
Action (programming)	Rate-responsive pacing	Rate-responsive pacing	Bi-ventricular pacing (cardiac resynchronization)	Anti-tachycardia pacing and defibrillation
Reward (outcome)	Fatigue, perceived exertion	Dyspnea with exertion	Heart failure symptoms, hospitalization, mortality	Termination of VT/VF, mortality

1. Pacing for sinus node dysfunction and chronotropic incompetence. In healthy individuals, the heart's intrinsic pacemaker, the sinoatrial node (SAN), maintains the resting heart rate, and uses input from the autonomic nervous system to adapt to the physiological demands of exercise and stress to maintain an appropriate heart rate by increasing it. Its function is well-documented to decline with age, leading to a reduced heart rate response to exercise and even daily activities in older individuals. In some cases, fibrosis or changes in atrial physiology can result in complete loss of activation of the SAN, leading to electrical atrial asystole, or lack of intrinsic activation of the atrial tissue. In the absence of an escape rhythm from elsewhere in the conduction system, these individuals can suffer from syncope or potentially cardiac arrest. The spectrum of dysfunction of the SAN is categorized under the umbrella of sinus node dysfunction (SND), and when associated with symptoms, represents a class I indication for implantation of a permanent pacemaker[12].

Chronotropic incompetence (CI) refers to a subtype of SND whereby the SAN is unable to increase the heart rate in response to increased activity or exercise[12,13]. Chronotropic incompetence is associated with impaired quality-of-life, and has been found to be an independent predictor of adverse cardiovascular events and overall mortality in certain populations[14]. Although the specific heart rate cutoff used to define CI is not well-defined, it is generally suggested that the failure to achieve 80-85% of the predicted age-dependent maximum heart rate with exertion (220 – age), with associated symptoms such as fatigue or shortness of breath, is sufficient to warrant evaluation for pacing solutions to improve heart rate with exertion.

A common feature of patients with SND, either complete asystole or CI, is the utility of atrial pacing to set the underlying heart rate for normal activities or exertion. In the process of implanting a pacemaker for atrial pacing, regulation of the heart rate is transferred from the body's intrinsic system of autonomic regulation to the device itself, which must learn to adapt to changes in demand as would require increasing the pacing rate to match the needs of physical activity[15]. To match pacing to exertional demands, the device must have a method to ascertain the presence and degree of physical activity, for which some common sensors are integrated in modern CIEDs. Two well-established methods include an accelerometer that uses piezoelectric crystals to identify movement, and a sensor for lung impedance, which monitors for changes in breathing rate. These two sensors are combined in the internal algorithm for most commercial CIEDs in such a way that the initial movement detected by the accelerometer is matched to changes in respiratory rate to confirm that exertion is occurring in the body. This combined methodology thus prevents inappropriate pacing changes resulting from other causes of vibration or movement. However, the relationship between movement and the degree of heart rate increase necessary can vary significantly from person to person, for example in a 40 year-old versus an 85 year-old. Due to a lack of intrinsic adaptability, modifications from standard algorithms are empiric rather than based on the patient's physiologic response to the current pacemaker settings.

Most commercial CIEDs employ adjustable parameters to attempt to match the pacing rate to the level of exertion, which include the sensitivity for detection of activity as well as the degree to which the pacing rate is increased and the peak pacing rate. In standard clinical settings, these adjustments are made offline, occasionally with the additional evaluation during treadmill or ergometer use. The requirement for offline adjustment creates a burden for patients, who must wait for the next clinic visit for adjustments to be made, and the inability to make minor modifications to improve overall exercise capacity.

2. Heart rate modulation for heart failure with preserved ejection fraction. Another condition where heart rate has increasingly drawn attention in clinical decision-making is heart failure with preserved ejection fraction (HFpEF)[16]. HFpEF[17] is sub-type of heart failure in which the predominant pathophysiological deficit lies in the inability of the heart to accommodate filling during diastole, resulting in fluid backup and accumulation in the lungs and other tissues. The symptoms of HFpEF range widely across individuals, with some requiring frequent hospital admission to remove fluid using intravenous diuretic agents, and others primarily noting symptoms only with exertion.

Historically, patients who have HFpEF have been treated empirically with heart rate reduction strategies, such as giving medications to reduce heart rates. However, recent work has demonstrated that certain HFpEF patients may benefit from more permissive or even increased heart rate. Recent studies, such as the myPACE trial, demonstrated patients whose CIEDs were programmed to have an increased heart rate had better quality of life

scores based on the Minnesota Living with Heart Failure Questionnaire, improved changes in heart failure biomarkers (NT-proBNP) and decreased durations of abnormal heart rhythms, such as atrial fibrillation[18]. In contrast, in the RAPID-HF trail, there was no improvement in oxygen consumption or cardiac output with faster heart rates programmed on patients 'pacemakers during exercise[19]. One explanation for these seemingly contradicting findings is that the impact of heart rate on symptoms in HFpEF is highly individualized, such that certain patients benefit from lower heart rates while others may have fewer symptoms with higher rates.

The highly individualized presentation and treatment response across patients with HFpEF presents the opportunity to modify CIED programming to better match the requisite pacing rates to the impact on symptoms, particularly during exertion. In theory, a system that uses lung impedance, which is also a marker of fluid accumulation in the lung and standard in many CIEDs, could be modified to pace at faster or slower rates to match the clinical effects. However, like with heart rate programming for SND, these changes can presently only be made empirically during standard clinical visits, rather than learned dynamically by the device during real-world activities.

3. Cardiac resynchronization therapy for heart failure with reduced ejection fraction. Another major subtype of heart failure includes individuals whose ventricular function becomes depressed or reduced, termed heart failure with reduced ejection fraction (HFrEF), which can be a result of a prior myocardial infarction or ongoing coronary ischemia, as well as from genetic or unknown (termed idiopathic) etiologies. The common underlying problem is that patients with HFrEF suffer from symptoms reflecting both impaired filling, like HFpEF, and reduced cardiac contraction, which can cause fatigue, life-threatening arrhythmias, and eventually death in many cases.

The past few decades have brought a number of medical therapies to improve symptoms and outcomes for patients with HFrEF; however, a particularly interesting technological innovation in management of HFrEF has been the use of CIEDs to provide synchronization to electrical activation among patients whose reduced pump function is also associated with dyssynchony between the left and right ventricles due to abnormalities in the heart's electrical conduction system. These devices, called cardiac resynchronization therapy (CRT) devices, are a type of CIED that include two separate pacing leads that supply energy to the right and left ventricles in a process that restores synchrony to electrical activation and, ideally, contraction of the heart. A number of clinical trials have identified improvement in symptoms and overall mortality with implantation of CRT devices in certain populations with HFrEF[20,21], and for many patients with this diagnosis, implantation of a CIED with CRT represents a standard of care, along with medication.

Like other CIEDs, there are a number of parameters that can be programmed in CRT devices towards the goal of improving cardiac dynamics and improving clinical outcomes. Specifically, the timing of pacing activation of the right and left leads of a CRT can be adjusted between chambers in an effort to restore muscle contraction to as close to normal as possible, as well as in conjunction with intrinsic atrial activation, either detected or provided by a pacing lead in the right atrium. A number of commercial algorithms have been developed and applied seeking to optimize pacing for improved ventricular function; however, like other pacing modifications, the inter-individual variation across patients creates a challenge with finding a one-size-fits-all methodology, as it is highly likely that the optimal settings will be noticeably different between patients.

4. Detection and therapy of ventricular arrhythmias by implantable defibrillators. One of the major innovations in development of CIEDs was the ability to design a completely implantable cardiac defibrillator (ICD) capable of detecting and treating life-threatening ventricular arrhythmias, including ventricular tachycardia and fibrillation. Initially employed in patients who had already suffered from these arrhythmias, clinicians eventually realized that even patients who have not had a prior event could be at risk, and obtained a mortality benefit from implantation of an ICD prophylactically[22,23].

The internal mechanism of an ICD entails two groups of algorithms, with some degree of manual tuning or programming. First, the ICD must be able to detect that a life-threatening arrhythmia is present, and then it must deliver therapy in the form of either rapid ventricular pacing (initially) or a high-voltage shock (if rapid pacing is ineffective). Detection algorithms are generally based on the presence of a rapid ventricular rate, and use characteristics like onset, morphology, and stability to discriminate the etiology of the rapid rate from other causes, such as sinus tachycardia or supraventricular tachycardia (including atrial fibrillation). Appropriate detection of ventricular arrhythmias remains a major challenge, as inappropriate treatment not only causes substantial morbidity (inappropriate shock), but is associated with increased mortality. The focus on algorithms for better therapy for ventricular arrhythmias has primarily centered on the use of rapid, or appropriately timed, ventricular pacing, also called anti-tachycardia pacing during the arrhythmia to 'break the circuit', and potentially forgo the need to deliver a high-voltage shock. Insights from invasive electrophysiology studies have been heavily leveraged to uncover patient-specific features of a ventricular arrhythmia that could suggest the utility of a well-timed paced beat, or set of paced beats, to terminate the arrhythmia. These types of internal algorithms are increasingly being evaluated and applied in programming of ICDs, although to-date, a well-adopted, validated approach remains elusive, leading to many patients continuing to suffer from potentially avoidable shocks that may be worsening outcomes.

C. Reinforcement learning for CIED programming

1. Background on RL algorithms. Reinforcement learning (RL) is a branch of artificial intelligence where the algorithm itself is given agency to select actions

towards the goal of learning the optimal solution through positive and negative rewards based on choices it makes. A typical RL process is depicted in Fig. 2. The RL agent takes actions in an unknown, possibly uncertain (random or probabilistic) environment. The environment's reaction to the action is viewed as a change in state by an observer, which provides a reward to the agent based on how closely the environment comes to some optimally desired condition. The agent uses the reward to update its decision process and selects a new action based on the environment's state, repeating the process with the intent maximizing the reward received for its actions.

The power of RL lies in balancing its exploratory and exploitative nature. To validate its current solution, the agent leverages exploitation by utilizing what it has already learned, reinforcing the selection of actions that have yielded higher rewards. Meanwhile, the exploratory aspect of RL allows the agent to test alternative actions that might lead to even greater rewards. Through repeated iterations of exploration and exploitation, the RL agent progressively converges toward an optimal strategy for interacting with the environment, ultimately maximizing its reward.

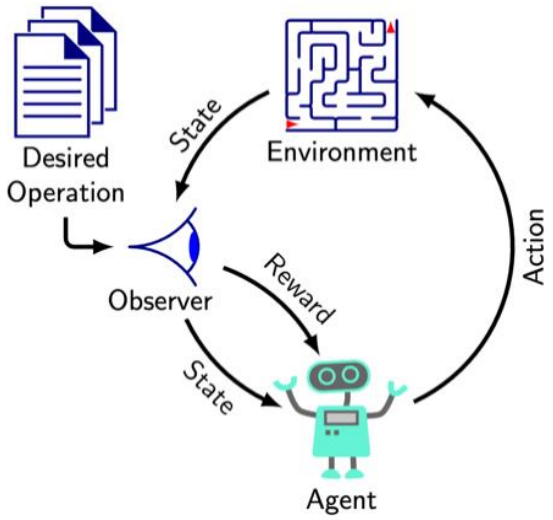

Figure 2. Reinforcement Learning Overview. An agent (computer algorithm) observes state information and rewards provided by the environment and selects actions to interact with it. This information is processed to maximize rewards through the desired operation.

Reinforcement learning can be implemented with or without learning an explicit model of the environment, termed *model-based* and *model-free*, respectively. *Model-based RL* creates an internal representation of the environment based on accumulated interactions that it uses to simulate different combinations of future actions to identify those that yield the maximum reward. In contrast, *model-free RL* does not maintain an explicit representation of the environment, and instead selects the next action based solely on the current state and a table of estimated rewards for each possible action. In simple environments, model-based RL generally converges to an optimal policy more quickly than model-free RL, at the expense of significant memory and computational resources. In contrast, model-free RL is less resource-intensive, but demands more training iterations to achieve convergence. Despite these differences, model-based RL scales more effectively to large and complex environments by leveraging deep neural networks (DNNs)[24] for approximation, enabling efficient representation and computation. These advancements have given rise to deep RL algorithms, which have demonstrated the ability to create novel solutions with superhuman efficiency[5,25-27].

2. Existing non-RL-based approaches for CIED programming. Current CIED programming is based on a method called *process control*, which refers to the regulation of an environment by adjusting its inputs based on the difference between the environment's current state and a desired state. The rate response feature of a pacemaker is an example of a simple process control mechanism. A sensor, such as a piezoelectric accelerometer, detects patient motion and outputs values categorized into predefined ranges. Each range is associated with a specific heart rate. As the sensor output transitions between ranges, the pacemaker adjusts the pacing rate accordingly. These ranges and rates are modified offline by a clinician to optimize exercise capacity for each patient. This method is straightforward to implement and requires minimal processing power and memory, making it suitable for resource-constrained devices like CIEDs. However, this approach is limited in precision and adaptability, as the system cannot self-adjust ranges or heart rates between clinical visits, leaving patients with suboptimal rate responses until their next evaluation.

more dynamic, non-RL method to program CIEDs uses continuous process control, such as the *Proportional-Integral-Derivative* controller, which monitors the difference (or error) between the desired state and the current environment state over time. Continuous process control algorithms enhance pacemaker rate response by enabling the device to self-adjust pacing rates to match the level of activity detected by the accelerometer. However, in practice, this method poses challenges for battery-powered devices like CIEDs, which have limited energy available for ongoing evaluation and adjustment. Moreover, simply matching accelerometer-based activity may not achieve the desired outcome of improving exercise capacity. The resulting design process remains largely manual and error-prone, as it relies on human engineers to define and fine-tune system parameters. Broadly, the non-RL-based methods of CIED programming have provided insights into the potential and challenges with designing smart programming methods, but also greater motivation for even smarter systems using RL.

3. Reinforcement learning for CIED programming. Reinforcement learning offers a promising alternative for automating the design of next-generation pacemakers. By focusing solely on reward design to represent clinical and physiological requirements, RL enables devices to autonomously learn optimal pacing strategies through interaction with their environment. This paradigm shift has the potential to unlock unprecedented levels of adaptability and performance, paving the way for truly intelligent and personalized medical devices.

An example of RL's application to pacing was demonstrated by Dole et al.[28], who applied RL to the pacing function in subjects with Mobitz II, second-degree atrioventricular (AV) block. Starting with a process control algorithm in which the environment was defined by sensed native atrial and ventricular impulses, and the therapy (i.e., action) defined as ventricular pacing, they implemented an RL algorithm to modify the programmed AV interval in response to changes in the intrinsic atrial rate. Unlike the standard static algorithm that uses a fixed AV interval, the RL algorithm developed by Dole et al. dynamically tracked and learned changes in the AV interval. This allowed it to deliver ventricular-paced impulses that closely matched the timing of intrinsic conduction. The result was smoother pacing during dynamic activity, avoiding the pauses or missed beats commonly observed with standard pacemaker settings. This example highlights RL's potential to address critical limitations in current pacemaker algorithms. By enabling dynamic, adaptive control of therapy parameters, RL can significantly improve the precision and responsiveness of pacing, further underscoring its promise as a transformative technology for next-generation CIEDs.

D. Safe Reinforcement Learning

The RL paradigm is inspired by the way organisms reinforce behaviors—good or bad—based on the rewards obtained from past experiences. A critical aspect of this process is the acquisition of new experiences through new decisions, often made without prior knowledge or assurance that those decisions will lead to better outcomes (i.e., higher rewards). In the context of clinical decision-making, this capacity for exploration could be highly beneficial, as it opens the door to discovering novel treatment approaches or paradigms that may not have been previously recognized. However, with novelty comes the inherent risk that these new approaches could lead to worse outcomes for patients. This duality highlights the need to balance exploration and safety, especially when applying RL to safety-critical systems like CIEDs. To address these concerns, identifying safe methods of exploration becomes a key consideration for the successful clinical implementation of RL. This balance ensures that while novel, potentially superior therapies are explored, patient safety is not compromised. In the following section, we discuss strategies and considerations for applying RL in programming safety-critical systems like CIEDs, emphasizing the importance of safety in medical applications of this transformative technology.

1. Defining safety through fomal language specification. Importantly, the concept of 'safety' as interpreted by a clinician must be translated into algorithmic meaning to be integrated with RL. Underlying this translation is the need for a formal, natural language specification that describes what the algorithm should accomplish in plain terms. For many complex systems, such as programming CIEDs, this process of translation can be difficult and error prone[29]. For example, the clinical concept of 'lightheadedness', as might occur if the pacing rate is inadequate to meet the demands of exercise, could be interpreted algorithmically in several ways: a failure to pace the atrium at a sufficient rate, a failure to reach a desired ventricular rate with pacing, or perhaps a sudden drop in heart rate when reaching the upper rate limit (i.e., pacemaker syndrome). In programming a safety feature to operate algorithmically, some definition is required to translate 'lightheadedness' into such a formal language system that is succinct and nonambiguous.

In 2007, Boston Scientific[30] released a natural language specification for a dual-chamber pacemaker to the research community that was subsequently used to define various algorithms of functionality[31-33], although applications were limited in complex clinical situations. Later work by Jiang et al.[34] used a formal modeling language to successfully express and validate the complete pacemaker specification, including complex safety features like pacemaker-mediated tachycardia, a condition in which the pacemaker inappropriately paces the heart at a rapid rate due to mis-sensing non-intrinsic atrial activation (i.e., T wave or retrograde P wave) as sinus node activation. This work was extended by Dole et al.[35], who were the first to specify and validate a complete pacemaker in a formal logic with the capability of capturing all required timing relationships and released the first complete formal pacemaker specification. These types of innovations enable development of algorithms that can connect information seen internally by the CIED with clinical outcomes, measured by the clinician and perceived by the patient. In plain terms, they ensure that the reward detected by the device can be mapped to a clinically meaningful outcome, and allow integration of shielding in RL as outlined below.

2. Reward translation for reinforcement learning. Formal languages provide the tools to define CIED function algorithmically, but they do not themselves define the goal or reward of the system, a requirement in RL. The primary programming challenge in developing RL-driven systems therefore lies in reward translation, or designing a reward function that maps a sequence of programming decisions (by the device or a human making changes) to scalar rewards. The goal is to ensure that an RL agent, by maximizing the aggregated reward sum, converges to a policy that aligns with the specified learning objectives. For example, a RL-based algorithm seeking to avoid causing lightheadedness in a patient must take the formal language relating data obtained from the device to the clinical entity of 'lightheadedness' and apply a reward value toward which the device can learn (in this case, to avoid). As might be expected, complexity arises in real-world attempts at reward translation. The learning agent (i.e., the device), in seeking to avoid a situation associated with lightheadedness, might select pacing settings that cause other adverse outcomes (such as pacing the heart too fast and causing chest pain). As such, reward translation requires constant evaluation and refinement to account for the numerous unexpected outcomes that can result of a given algorithm.

3. Shielding and constrained exploration. As noted previously, the power of RL algorithms comes with the ability of the decision-making agent, through exploration, to identify potentially novel actions that improve the

short- or long-term outcome. A well-designed RL algorithm might identify pacing settings for a given individual that improve exercise capacity in a previously unrecognized manner. However, allowing an algorithm to search across the entire space of pacing settings could be haphazardous if it selected an extremely high pacing rate (e.g., 200 beats per minute) or an extremely low pacing rate (e.g., 10 beats per minute). Clearly, in this example some constraints must be placed on what specific space of possible pacing rates should be explored. This process is called shielding, and it forms the cornerstone of safe RL.

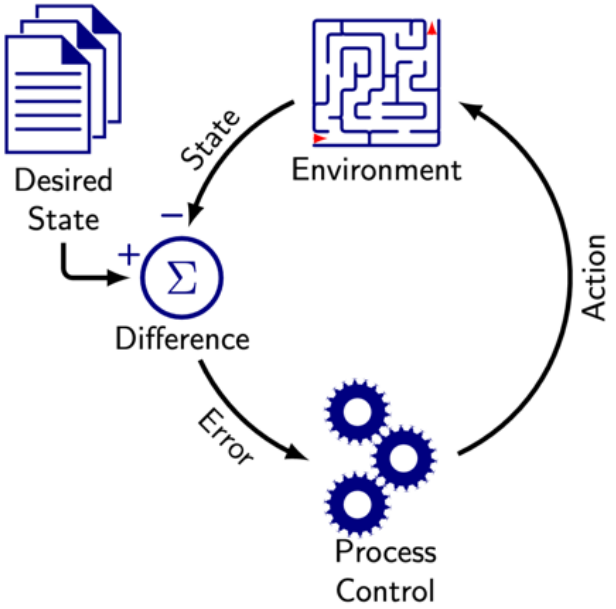

Figure 3. Process control. Regulation of the environment through actions seeking to make the environment match the desired state.

One of the challenges with the inclusion of shielding in RL is that for large, complex environments with many interacting parameters it is virtually impossible to train and test RL for every possible condition that could arise. This gap leaves untrained situations where the RL must interpolate its next action from what it has learned and may unknowingly select an unsafe action, which can include pacing at too high a rate or inhibition of pacing when it is required. Shielded RL (Fig. 4) is becoming a popular new technique that enables safety critical applications like medical devices to benefit from RL's unique ability to improve patient therapy without the dangers of incorrect actions.

The conduction pattern in each patient across the heart is unique, while the heart's response to different electrical impulses from the CIED is complicated and difficult to model or predict with accuracy. With such variability, the therapy must be constrained to maintain safety over a wide population of patients and clinical settings. Bloem, et al[36], introduced the concept of a shield to address this safety constriction for process controllers in complex environments. A shield is a simple automaton modeling the safety properties that must not be violated. When the shield sees the controller (i.e., agent) take an action that will lead to violation of a safety property, the shield modifies the action to maintain safety, otherwise the controller performs unhindered. Alshiekh, et al.[37], extended the concept of shielding to RL to enable its use in safety critical applications. Shielded RL enables safety critical applications like medical devices to benefit from RL's unique ability to improve patient therapy without the dangers of incorrect actions.

There are two basic forms of shielded RL, which are dependent on when and how the safety properties are implemented. In preemptive shielding (Fig. 4, a), the shield examines the current environment state and provides the RL agent with a list of acceptable actions from which it can select. This list assures that the agent's actions are always safe. Alternatively, in post posed shielding (Fig. 4, b), the shield is applied after the action has been selected, where the shield reviews the agent's action in context of other information (i.e., state) that might be informative. If the action would lead to violating a safety property, the shield changes the action to preserve safety. In either design, the shield remains a permanent part of the algorithm, even after RL training is completed.

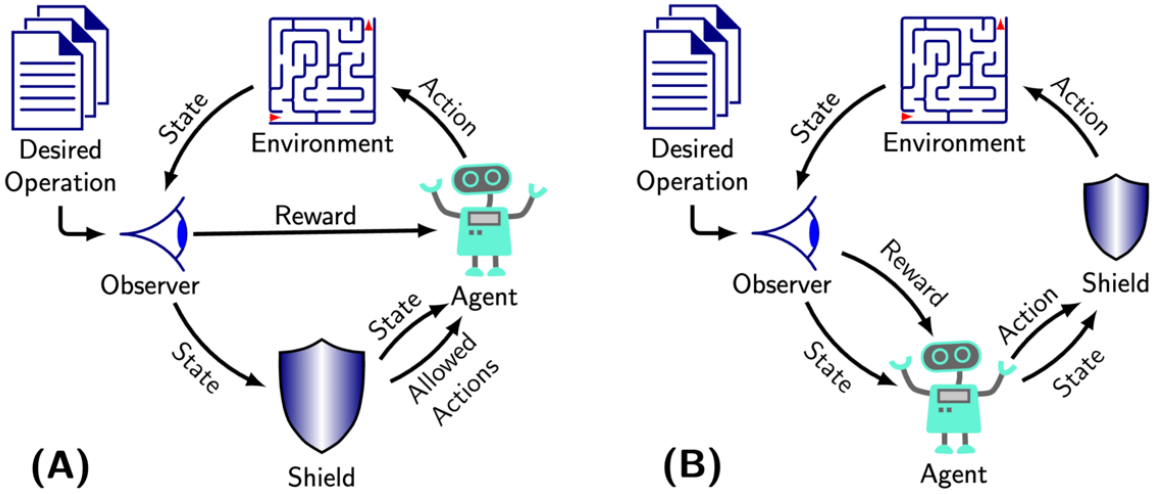

Figure 4. Types of shielded reinforcement learning. A) Pre-emptive shielding entails placing constraints on the space of actions available to the agent. **B) Post posed shielding** applies a shield informed by information from the state to actions selected by the agent. Both methods are trained to avoid potentially unsafe actions during the exploration phase of RL.

4. Testing and verification. The transition from design to implementation is not complete without a formal testing methodology. In the case of safe RL, this system requires a testing framework for both the expected behavior used to reward the RL agent for optimal actions, as well as the shield that prevents the RL agent from violating safety properties. Each of these must be validated to assure correctness. In their pacemaker specification, Dole et al.[35] developed an automaton to enable functional testing using test scenarios that demonstrated expected operation. Interestingly, they found that many test scenarios provided unusual conditions to provoke erroneous operation, emphasizing the justification for such testing. They focused on two common pacemaker features with complex timing, ventricular safety pacing and upper-rate holdoff, and demonstrated optimized Mobitz II pacing therapy to avoid pacemaker syndrome. In this work, the shield consisted of two safety requirements: 1) The AV interval of the paced beat may not be shorter than the most recent intrinsic AV interval and 2) the AV interval should not be longer than the maximum allowed paced AV interval. Their testing showed that indeed, the RL agent was able to dynamically find the optimal AV interval that matched the intrinsic ventricular AV interval with a continuously changing intrinsic heart rate, without inappropriately shortening or prolonging the paced interval in such a way that could inhibit appropriate pacing or result in inadequate ventricular capture (i.e., too short an AV interval). While limited to a specific set of features for pacemaker function, this work provides an example for how safe RL can be deployed within a learning algorithm seeking to improve clinical outcomes.

E. Conclusion.

Contemporary CIEDs have found clinical utility within a growing number of applications, ranging from replacing the heart's intrinsic electrical system to treating potentially life-threatening arrhythmias. Existing programming methodologies for CIEDs remain highly static, with limited opportunities to individualize programming in order to provide the greatest benefit for patients. The RL framework, which through exploration and exploitation is able to learn new settings for a CIED, holds great promise for personalized programming of CIEDs. Within this framework, the role of safe RL is integral to allow sufficient exploration without subjecting patients to increased risk that the algorithm could select settings that could result in adverse outcomes. While the field itself remains in its infancy in terms of real-world applications of RL, the ability of the CIED to capture physiological information about patients provides sufficient motivation to pursue innovations in both RL and safe RL.

Acknowledgements

This work was supported with funding from the AB Nexus award (JK, AT, MAR), and the National Institutes of Health (MAR R01HL146824).

References

1. Sutton RS, AG B. *Reinforcement Learning*. 2nd ed. Cambridge, MA: MIT Press; 2018.
2. Hu Y, Si B. A Reinforcement Learning Neural Network for Robotic Manipulator Control. *Neural computation*. 2018;30:1983-2004. doi: 10.1162/neco_a_01079
3. Peters J, Schaal S. Reinforcement learning of motor skills with policy gradients. *Neural networks : the official journal of the International Neural Network Society*. 2008;21:682-697. doi: 10.1016/j.neunet.2008.02.003
4. Silver D, Huang A, Maddison CJ, Guez A, Sifre L, van den Driessche G, Schrittwieser J, Antonoglou I, Panneershelvam V, Lanctot M, et al. Mastering the game of Go with deep neural networks and tree search. *Nature*. 2016;529:484-489. doi: 10.1038/nature16961
5. Silver D, Schrittwieser J, Simonyan K, Antonoglou I, Huang A, Guez A, Hubert T, Baker L, Lai M, Bolton A, et al. Mastering the game of Go without human knowledge. *Nature*. 2017;550:354-359. doi: 10.1038/nature24270
6. Google RankBrain. 2020.
7. Levy AE, Biswas M, Weber R, Tarakji K, Chung M, Noseworthy PA, Newton-Cheh C, Rosenberg MA. Applications of machine learning in decision analysis for dose management for dofetilide. *PLoS One*. 2019;14:e0227324. doi: 10.1371/journal.pone.0227324
8. Prasad N, Cheng LF, Chivers C, Draugelis M, B. E. A reinforcement learning approach to weaning of mechanical ventilation in intensive care units. *arXiv* 2017;https://arxiv.org/abs/1704.06300.
9. Komorowski M, Celi LA, Badawi O, Gordon AC, Faisal AA. The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. *Nat Med*. 2018;24:1716-1720. doi: 10.1038/s41591-018-0213-5
10. Barrett CD, Suzuki Y, Hussein S, Garg L, Tumolo A, Sandhu A, West JJ, Zipse M, Aleong R, Varosy P, et al. Evaluation of Quantitative Decision-Making for Rhythm Management of Atrial Fibrillation Using Tabular Q-Learning. *Journal of the American Heart Association*. 2023;12:e028483. doi:10.1161/jaha.122.028483
11. Aquilina O. A brief history of cardiac pacing. *Images Paediatr Cardiol*. 2006;8:17-81.
12. Kusumoto FM, Schoenfeld MH, Barrett C, Edgerton JR, Ellenbogen KA, Gold MR, Goldschlager NF, Hamilton RM, Joglar JA, Kim RJ, et al. 2018 ACC/AHA/HRS Guideline on the Evaluation and Management of Patients With Bradycardia and Cardiac Conduction Delay: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines and the Heart Rhythm Society. *Circulation*. 2019;140:e382-e482. doi: doi:10.1161/CIR.0000000000000628
13. Brubaker PH, Kitzman DW. Chronotropic incompetence: causes, consequences, and management. *Circulation*. 2011;123:1010-1020. doi: 10.1161/circulationaha.110.940577
14. Hinkle LE, Jr., Carver ST, Plakun A. Slow heart rates and increased risk of cardiac death in middle-aged men. *Arch Intern Med*. 1972;129:732-748.
15. Świerżyńska E, Oręziak A, Główczyńska R, Rossillo A, Grabowski M, Szumowski Ł, Caprioglio F, Sterliński M. Rate-Responsive Cardiac Pacing: Technological Solutions and Their Applications. *Sensors (Basel, Switzerland)*. 2023;23. doi: 10.3390/s23031427
16. Shang X, Lu R, Liu M, Xiao S, Dong N. Heart rate and outcomes in patients with heart failure with preserved ejection fraction: A dose-response meta-analysis. *Medicine (Baltimore)*. 2017;96:e8431. doi: 10.1097/md.0000000000008431
17. Fonarow GC, Stough WG, Abraham WT, Albert NM, Gheorghiade M, Greenberg BH, O'Connor CM, Sun JL, Yancy CW, Young JB. Characteristics, treatments, and outcomes of patients with preserved systolic function hospitalized for heart failure: a report from the OPTIMIZE-HF Registry. *J Am Coll Cardiol*. 2007;50:768-777. doi:10.1016/j.jacc.2007.04.064
18. Infeld M, Wahlberg K, Cicero J, Plante TB, Meagher S, Novelli A, Habel N, Krishnan AM, Silverman DN, LeWinter MM, et al. Effect of Personalized Accelerated Pacing on Quality of Life, Physical Activity, and Atrial Fibrillation in Patients With Preclinical and Overt Heart Failure With Preserved Ejection Fraction: The myPACE Randomized Clinical Trial. *JAMA cardiology*. 2023;8:213-221. doi: 10.1001/jamacardio.2022.5320
19. Reddy YNV, Koepp KE, Carter R, Win S, Jain CC, Olson TP, Johnson BD, Rea R, Redfield MM, Borlaug BA. Rate-Adaptive Atrial Pacing for Heart Failure With Preserved Ejection Fraction: The RAPID-HF Randomized Clinical Trial. *Jama*. 2023;329:801-809. doi: 10.1001/jama.2023.0675
20. Moss AJ, Hall WJ, Cannom DS, Klein H, Brown MW, Daubert JP, Estes NA, 3rd, Foster E, Greenberg H, Higgins SL, et al. Cardiac-resynchronization therapy for the prevention of heart-failure events. *N Engl J Med*. 2009;361:1329-1338. doi: 10.1056/NEJMoa0906431
21. Aiba T, Hesketh GG, Barth AS, Liu T, Daya S, Chakir K, Dimaano VL, Abraham TP, O'Rourke B, Akar FG, et al. Electrophysiological consequences of dyssynchronous heart failure and its restoration by resynchronization therapy. *Circulation*. 2009;119:1220-1230. doi:10.1161/CIRCULATIONAHA.108.794834
22. Goldberger Z, Lampert R. Implantable cardioverter-defibrillators: expanding indications and technologies. *Jama*. 2006;295:809-818. doi: 10.1001/jama.295.7.809
23. Zipes DP, Camm AJ, Borggrefe M, Buxton AE, Chaitman B, Fromer M, Gregoratos G, Klein G, Moss AJ, Myerburg RJ. ACC/AHA/ESC 2006 guidelines for management of patients with ventricular arrhythmias and the prevention of sudden cardiac death: a report of the American College of Cardiology/American Heart Association Task Force and the European Society of Cardiology Committee for Practice Guidelines (Writing Committee to Develop guidelines for management of patients with ventricular arrhythmias and the prevention of sudden cardiac death) developed in collaboration with the European Heart Rhythm Association and the Heart Rhythm Society. *Europace*. 2006;8:746-837.

24. Goodfellow I, Bengio Y, Courville A. *Deep Learning*. MIT Press; 2016.
25. Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, Graves A, Riedmiller M, Fidjeland AK, Ostrovski G, et al. Human-level control through deep reinforcement learning. *Nature*. 2015;518:529-533. doi: 10.1038/nature14236
26. Vinyals O, Babuschkin I, Czarnecki WM, Mathieu M, Dudzik A, Chung J, Choi DH, Powell R, Ewalds T, Georgiev P, et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*. 2019;575:350-354. doi: 10.1038/s41586-019-1724-z
27. Kaufmann E, Bauersfeld L, Loquercio A, Müller M, Koltun V, Scaramuzza D. Champion-level drone racing using deep reinforcement learning. *Nature*. 2023;620:982-987. doi: 10.1038/s41586-023-06419-4
28. Dole K, Gupta A, Komp J, Krishna S, Trivedi A. Correct-by-Construction Reinforcement Learning of Cardiac Pacemakers from Duration Calculus Requirements. *Proceedings of the AAAI Conference on Artificial Intelligence*. 2023;37:14792-14800. doi: 10.1609/aaai.v37i12.26728
29. Whalen MW, Gacek A, Cofer D, Murugesan A, Heimdahl MPE, Rayadurgam S. Your "What" Is My "How": Iteration and Hierarchy in System Design. *IEEE Software*. 2013;30:54-60. doi: 10.1109/MS.2012.173
30. Laboratory SQR. *PACEMAKER System Specification*. Boston Scientific; 2007.
31. Gomes AO, Oliveira MVM. Formal Specification of a Cardiac Pacing System. Paper/Poster presented at: FM 2009: Formal Methods; 2009//, 2009; Berlin, Heidelberg.
32. Larson BR. Formal semantics for the PACEMAKER system specification. In: *Proceedings of the 2014 ACM SIGAda annual conference on High integrity language technology*. Portland, Oregon, USA: Association for Computing Machinery; 2014:47–60.
33. Méry D, Singh NK. Formal Specification of Medical Systems by Proof-Based Refinement. *ACM Trans Embed Comput Syst*. 2013;12:Article 15. doi: 10.1145/2406336.2406351
34. Jiang Z, Pajic M, Moarref S, Alur R, Mangharam R. Modeling and verification of a dual chamber implantable pacemaker. In: *Proceedings of the 18th international conference on Tools and Algorithms for the Construction and Analysis of Systems*. Tallinn, Estonia: Springer-Verlag; 2012:188–203.
35. Dole K, Gupta A, Komp J, Krishna S, Trivedi A. Event-Triggered and Time-Triggered Duration Calculus for Model-Free Reinforcement Learning. Paper/Poster presented at: 2021 IEEE Real-Time Systems Symposium (RTSS); 7-10 Dec 2021, 2021;
36. Bloem R, Könighofer B, Könighofer R, Wang C. Shield Synthesis. Paper/Poster presented at: Tools and Algorithms for the Construction and Analysis of Systems; 2015//, 2015; Berlin, Heidelberg.
37. Alshiekh M, Bloem R, Ehlers R, Könighofer B, Niekum S, Topcu U. Safe Reinforcement Learning via Shielding. *Proceedings of the AAAI Conference on Artificial Intelligence*. 2018;32. doi:10.1609/aaai.v32i1.11797

