Received: 28 September 2022 | Accepted: 27 December 2023

DOI: 10.1111/bjet.13429

ORIGINAL ARTICLE

British Journal of Educational Technology BERA

Towards prescriptive analytics of self-regulated learning strategies: A reinforcement learning approach

Ikenna Osakwe[1] | Guanliang Chen[1] | Yizhou Fan[1,2] | Mladen Rakovic[1] | Shaveen Singh[1] | Lyn Lim[3] | Joep van der Graaf[4] | Johanna Moore[5] | Inge Molenaar[4] | Maria Bannert[3] | Alex Whitelock-Wainwright[1] | Dragan Gašević[1,5]

[1]Centre for Learning Analytics at Monash, Faculty of Information Technology, Monash University, Clayton, Australia

[2]Graduate School of Education, Peking University, Beijing, China

[3]TUM School of Education, Technical University of Munich, Munich, Germany

[4]Behavioural Science Institute, Radboud University, Nijmegen, Netherlands

[5]School of Informatics, University of Edinburgh, Edinburgh, UK

Correspondence
Ikenna Osakwe, Faculty of Information Technology, Centre for Learning Analytics at Monash, Monash University, Clayton, VIC 3168, Australia.
Email: richard.osakwe@monash.edu

Abstract: Self-regulated learning (SRL) is an essential skill to achieve one's learning goals. This is particularly true for online learning environments (OLEs) where the support system is often limited compared to a traditional classroom setting. Likewise, existing research has found that learners often struggle to adapt their behaviour to the self-regulatory demands of OLEs. Even so, existing SRL analysis tools have limited utility for real-time or individualised support of a learner's SRL strategy during a study session. Accordingly, we explore a reinforcement learning based approach to learning optimal SRL strategies for a specific learning task. Specifically, we utilise the sequences of SRL processes acted by 44 participants, and their assessment scores for a prescribed learning task, in a purpose-built OLE to develop a long short-term memory (LSTM) network based reward function. This is used to train a reinforcement learning agent to find the optimal sequence of SRL processes for the learning task. Our findings indicate that the developed agents were able to effectively select SRL processes so as to maximise a prescribed learning goal as measured by predicted assessment score and predicted knowledge gains.

The contributions of this work will facilitate the development of a tool which can detect sub-optimal SRL strategy in real-time and enable individualised SRL focused scaffolding.

KEYWORDS
learning analytics, learning strategies, reinforcement learning, self-regulated learning

Practitioner notes

What is already known about this topic

- Learners often fail to adequately adapt their behavior to the self-regulatory demands of e-Learning environments.
- In order to promote effective Self-regulated learning (SRL) capabilities, researchers and educators need tools that are able to analyze and diagnose a learner's SRL strategy use.
- Current methods for SRL analysis are more often descriptive as opposed to prescriptive and have limited utility for real-time analysis or support of a learner's SRL behavior.

What this paper adds

- This paper proposes the use of Reinforcement Learning for prescriptive analytics of SRL. We train a Reinforcement Learning agent on sequences of SRL processes acted by learners in order to learn the optimal SRL strategy for a given learning task.

Implications for practice and/or policy

- Our work will facilitate the development of a tool which can detect sub-optimal SRL strategy in real-time and enable individualized SRL focused scaffolding.
- The implications of our work can aid in course design by predicting the self-regulatory load imposed by a given task.
- The ability to model SRL strategies using Reinforcement Learning can be extended to simulate or test SRL theories.

INTRODUCTION

Academic researchers have established that self-regulation is vital for effective learning (Butler & Winne, 1995; Panadero, 2017; Winne, 2020, 2021; Winne et al., 1998; Zimmerman, 2013). In an academic setting, self-regulation is the process of setting goals, devising strategies to achieve said goals, and, as the tasks are underway, monitoring and evaluation of progress towards these goals (Hattie & Timperley, 2007; Winne et al., 1998). If an obstacle is encountered or the learner's progress evaluation is deemed unsatisfactory, self-regulating learners may alter strategies to improve progress (Hattie & Timperley, 2007; Winne et al., 1998).

Researchers have documented that students' self-regulated learning (SRL) capabilities can play an important role in academic achievement (Broadbent, 2017; Broadbent & Poon, 2015; Dent & Koenka, 2016; Theobald, 2021). For example, Dent and Koenka (2016) find that academic success will depend on the way a learner deploys self-regulatory strategies (eg, goal-setting, planning, or monitoring) during the learning process. Kizilcec and Schneider (2015); Zheng et al. (2015) and Theobald (2021) showed that SRL strategies such as time management, motivation and self-monitoring play a significant role in goal attainment or academic performance.

As online learning continues to gain prominence, its inability to provide the same level of support and guidance typically seen in a physical setting remains an issue (Wong et al., 2017). Online learners need to be able to strategise what, when and how to engage with the abundance of resources made available to them. Lack of SRL abilities can make it challenging to successfully navigate such environments (Kizilcec & Schneider, 2015; Maldonado-Mahauad et al., 2018). Studies have found that many learners fail to adequately adapt their behaviour to the self-regulatory demands of e-learning environments (Azevedo & Aleven, 2013; Cerezo et al., 2017; Feyzi-Behnagh et al., 2014). Hence, the learner's ability to self-regulate is an increasingly important skill to achieve desired learning goals.

In order to promote effective SRL capabilities in challenging environments such as online learning, research is needed into prescriptive tools that can diagnose a learner's SRL strategy use and provide beneficial scaffolding. However, current methods for SRL analysis (see Section “SRL modelling tools”) are more often descriptive as opposed to prescriptive and have limited utility for real-time analysis of a learner's SRL patterns (see Section “Reinforcement learning to address the shortcomings of SRL analysis methods for prescriptive analytics”). Reinforcement learning techniques show promise in addressing these shortcomings by enabling the real-time prediction of an optimal action required to achieve a desired outcome (Luo, 2020; Nguyen & La, 2019) (see Section “Reinforcement learning to address the shortcomings of SRL analysis methods for prescriptive analytics”). Consequently, this paper analyses the use of reinforcement learning to learn optimal SRL strategy for a given learning task. We believe progress in this area can inform the development of an automated scaffolding tool, that prescribes learners individualised SRL strategies to achieve a desired learning goal. The results and their implications for theory and practice are further discussed in the paper.

BACKGROUND

SRL and trace data

Digital learning environments allow researchers not only to track students' learning performance, but also their learning interactions and activities such as click streams, eye gaze data, resource usage and chat logs (Azevedo & Gašević, 2019; Fan, van der Graaf, Lim, Rakovic, Kilgour, et al., 2022; Li et al., 2020); often referred to as trace data. These traces that learners generate as they engage in digital learning environments result from a variety of internal cognitive and metacognitive states, strategies and processes used by the student (Azevedo et al., 2013). Hence, studies propose the analysis of SRL and learning strategies through the use of learner traces and activity logs (Bannert et al., 2014; Hadwin et al., 2007). For example, Bannert et al. (2014) propose that SRL can be viewed as a trail of actions performed by learners while engaging in a study session. Further buttressing this notion, Hadwin et al. (2007) studied the activity of eight learners across two study sessions using the gStudy platform. They analysed the activity traces and compared the data to learner

self-reports on SRLs and determined that student activity traces are vital towards progressing our understanding of SRL.

SRL analysis can be performed using learner actions (eg, open quiz, play video) captured by trace data (Baker et al., 2020; Winne, 2013). However, learner actions are often platform dependent, which can limit the generalisability of findings (Baker et al., 2020; Fan, van der Graaf, Lim, Rakovic, Singh, et al., 2022). Alternatively, Fan, van der Graaf, Lim, Rakovic, Kilgour, et al. (2022) and Srivastava et al. (2022) have translated sequences of learning actions captured from trace data into theoretically grounded SRL processes. These SRL processes can be categorised according to varying levels of granularity but broadly fall under metacognition, low cognition or high cognition processes (Fan, van der Graaf, Lim, Rakovic, Kilgour, et al., 2022; Srivastava et al., 2022). Analysis can then be conducted on these platform agnostic concepts (Srivastava et al., 2022).

Various research groups have developed statistical and machine learning methods to construct abstract representations of learning strategies from trace log data to analyse the complex temporal patterns of SRL (Aleven et al., 2006; Azevedo et al., 2009; Blikstein et al., 2014; Hadwin et al., 2007; Saint et al., 2022). While efforts have been taken to analyse SRL processes on a frequency basis (Kovanović et al., 2015; Lust et al., 2011), these practices provide more of a summative view and offer insufficient insight into the sequential way learners employ strategies as they engage in a task. Since SRL is a continuous process rather than a standstill frame, SRL analysis techniques should be augmented with tools more suited to capturing the dynamics of temporality in learner engagement (Molenaar & Järvelä, 2014; Saint et al., 2021, 2022; Saint, Whitelock-Wainwright, et al., 2020). These methods largely fall into the categories of process mining, sequential pattern mining and Markov models.

SRL modelling tools

Sequential pattern mining (SPM) is a family of techniques used to extract the most frequently occurring sequences in a dataset (Mabroukeh & Ezeife, 2010). SPM is often used in SRL research to analyse the patterns displayed by learners while engaging with a digital learning environment. For instance, Bouchet et al. (2012) studied frequently occurring action sequences to infer that high-performing students were more systematic with their reading strategy, as their most frequent activity patterns included more relevant reading and full-length re-reads. They also found that academic performance was linked to monitoring patterns. Rabin et al. (2019) compared frequent sub-sequences of learner actions between different groups of learners and found that learners who fulfilled their initial intentions were more strategic in their resource usage; for instance, tending to access video lectures in a sequential manner.

Process mining (PM) is a group of techniques used to extract insights from event log data to track and model processes. Unlike SPM, PM is able to consider other relationships to direct succession such as causality and choice (van der Aalst et al., 2004). PM has steadily gained traction in academic research due to its ability to analyse temporal and sequential data (Maldonado-Mahauad et al., 2018; Romero & Ventura, 2013; Saint, Whitelock-Wainwright, et al., 2020). Cerezo et al. (2020) used a PM technique called inductive mining on learning management system event logs to contrast processual differences in pass versus fail students. They found that both groups were unlikely to follow the recommended course path, but students who passed exhibited significantly more self-regulatory processes.

Markov models (MMs) analyse sequential observations using a set of states and probabilistic transitions between those states under the assumption that future states are dependent only on the current state (Gagniuc, 2017).

Autonomous Markov models (AMMs) are MMs in which the outcome cannot be influenced by feedback or reward signals. Hidden Markov models (HMMs) and Markov chains (MCs) are the most common form of AMMs uses for SRL analysis (Biswas et al., 2017; Galyardt & Goldin, 2015; Kinnebrew & Biswas, 2011; Saint, Gašević, et al., 2020). In analysis of SRL patterns, the states used by AMMs are typically inferred to be a study session (Fincham et al., 2019) or a learning strategy (Biswas et al., 2010; Jovanović et al., 2017; Matcha et al., 2019).Fincham et al. (2019) and Matcha et al. (2019) used AMMs on trace data to extract study tactics from the learner's prior study sessions. The authors then clustered the sequences of learners; study tactics to deduce strategies used during the course. Zhang and Cheng (2019) explored the difference in negotiating behaviours of students with varying SRL capabilities (measured using a questionnaire) in a negotiated online reading assessment system. In the study, states were inferred to be metacognitive strategies. The study found significant differences in the behaviours of high and low SRL students; particularly, high SRL students showed better strategic planning and self-reflection.

Reinforcement learning for prescriptive analytics

Reinforcement learning is a type of machine learning technique that enables an agent to learn in an environment by trial and error using rewards as signals for positive (or negative) behaviour (Doroudi et al., 2019). In each episode, the agent's goal is to learn a sequence of actions that would maximise the total cumulative reward from a given state (Doroudi et al., 2019). The function which determines this action based on a given state is known as the policy. As the agent explores the environment, it receives information in the form of the state and reward and uses this information to take the next optimal action (see Figure 1). Hence, reinforcement learning algorithms are dynamic and flexible in that they are able to update and predict in real time (Luo, 2020; Nguyen & La, 2019).

Reinforcement learning in education

In the education field, reinforcement learning has been utilised successfully for personalised recommendations (Doroudi et al., 2019; Intayoad et al., 2020; Liang et al., 2022; Zhang et al., 2019).

Zhang et al. (2019) use reinforcement learning to filter noisy actions from user profiles for more accurate course recommendations. The proposed approach involves using a two-level

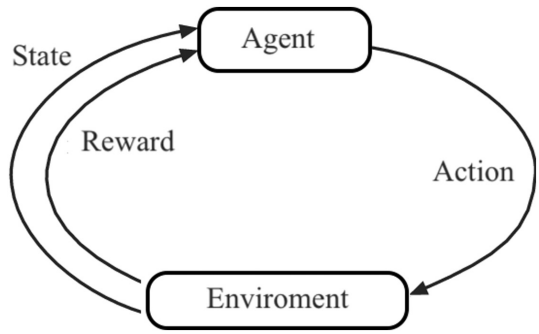

FIGURE 1 Simplified illustration of the reinforcement learning process.

hierarchy of RL agents. The first level agent is responsible for selecting a course category to recommend, while the second level agent is responsible for selecting a specific course within that category. The state space includes both user and course features, such as user demographics, historical activity data and course attributes. The action space for the first level agent includes the different course categories, while the action space for the second level agent includes the specific courses within a category. The reward function is designed to encourage the system to recommend courses that the user is likely to be interested in and to maximise user engagement with the recommended courses hence, the authors use a combination of click-through rate, completion rate, and time spent on course to determine the reward. The authors evaluated their proposed approach on a real-world dataset of MOOC course activities and compared it against several baseline methods, including non-RL approaches and single-level RL approaches and showed that the proposed approach outperformed the baselines in terms of recommendation accuracy and user satisfaction.

Other uses of RL in education relate to scaffolding strategy (Barnes & Stamper, 2008; Fahid et al., 2021; Johnson & Zaiane, 2012). Fahid et al. (2021) develop an RL model to adaptively determine how a student should engage with learning resources after an incorrect answer, based on the integrated cognitive antisocial potential (ICAP) theory. According to the ICAP theory, there exist four types of engagement modes—interactive, constructive, active and passive (Chi & Wylie, 2014). The learning environment used for the study did not provide interactive options, so the developed model works by suggesting to the learner one of the other three modes of engagement or providing no suggestions at all. After each incorrect solution, one of these four actions is proposed to maximise the learner's expected learning gain. The state representation used for this study was a total of 31 features that encompassed the learner's survey features such as gender, age and domain interest; video playback features such as time spent on video; and prior scaffolding engagement features including previous types and numbers of engagement modes delivered. The reward was the increase in predicted normalised learning gains. This prediction function was created from a dataset of participant's post- and pre-test scores and the sequences of scaffolding they received. The data were used to build a prediction model that maps sequences of scaffolding sequences to the difference in post- and pre- test scores. The developed RL model yielded scaffolding policies that outperformed heuristic-based policies such as only constructive scaffolding, no scaffolding or scaffolding at random, as measured by the average expected reward.

Despite the effectiveness of reinforcement learning in in other fields of research (Shao et al., 2019; Zhang & Mo, 2021) as well as in education for personalised recommendations and feedback strategy, to the best of our knowledge, there has been little to no research on the use of reinforcement learning for SRL analysis.

Reinforcement learning to address the shortcomings of SRL analysis methods for prescriptive analytics

The current approaches taken to model SRL patterns indicate that learning strategies can be derived from trace data and are able to provide valuable insights on the learner's processes during study sessions; however, they exhibit significant limitations:

1. *Real-time evaluation of a given pattern is difficult*. The current approaches do not offer insight on intermittent states (such as a learner's SRL processes half-way through a session); hence, various models will have to be built to analyse patterns at different points during a given study session—this makes it difficult to provide real-time personalised feedback. For instance, to evaluate a learner's SRL strategy

using every 5 minutes in an hour-long session using one of SPM, PM or AMM methods highlighted above, we would need to construct 20 (60 minutes/5 minutes) different models to evaluate a learner's patterns. However, we can use a single reinforcement learning model to calculate the predicted outcome at each interval.
2. *Methods are descriptive as opposed to prescriptive*. For instance, we can distinguish the patterns exhibited by low and high SRL strategy groups but cannot determine what actions an individual should take at a given time period to improve from the low to high group. Using one of SPM, PM or AMM methods highlighted above we can attain the average pattern displayed by the high SRL group, but this would not consider the individual learner's prior actions acted thus far in the learning session, which may make recommended actions ill-suited. Reinforcement learning shows promise in its potential to address these shortcomings. By building an abstract representation of the decision making process in SRL strategy selection, we can use reinforcement learning to analyse both the value of a given state and the actions that maximise achievement of a desired goal at any given time period.

We propose the utilisation of reinforcement learning's dynamic capabilities to model a learner's SRL processes as they engage in a study session. These distinctions can provide significant possibilities including (1) learning optimal SRL selection strategy for a given learning task, (2) detecting poor SRL strategy in real-time, (3) providing scaffolding to promote effective SRL strategy and (4) running simulations to explore various SRL theories.

RESEARCH OBJECTIVES

At this stage of the experiment, our goal is to use RL to find the optimal strategy for the learning task. We develop a reinforcement learning agent on sequences of SRL processes acted by learners to analyse to what extent the agent can learn various SRL strategies. We posit progress in our abilities to model SRL strategies using reinforcement learning can provide benefits in future work that include a reinforcement learning directed feedback strategy. Hence, in this paper we study:

> *RQ* To what extent can a reinforcement learning agent learn effective SRL strategy using learner traces?

METHODOLOGY

The process from data collection to model output is illustrated in Figure 2. The process is as follows: trace data are collected in a lab setting as students complete an assigned learning task. The trace data are then translated into sequences of SRL processes. These sequences of SRL processes are used to develop a reward function and train reinforcement learning models. Further information is provided in the following sections.

Data collection and processing

Trace data used in the experiment were collected from an experiment conducted in a lab setting at a university in the Netherlands. Analyses were conducted with 44 participants (Average Age = 21 years, $SD = 3$ years) from a wide range of degree programs but mostly from social sciences. The study uses trace data generated while learners interact with a

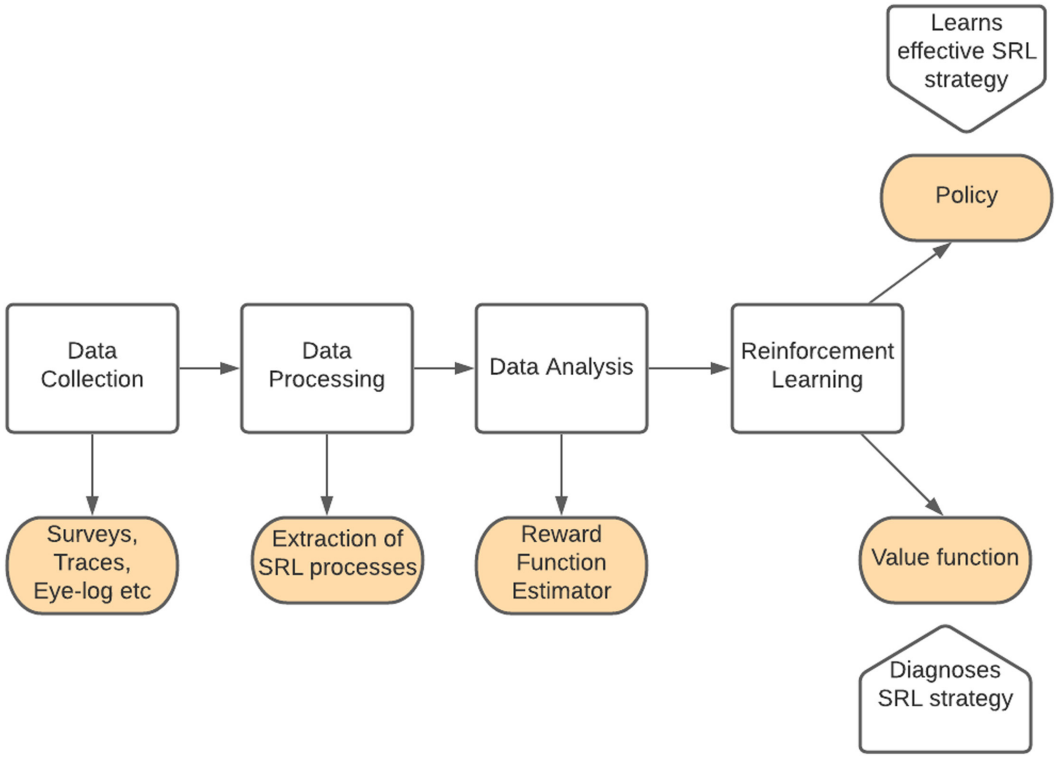

FIGURE 2 An overview of the steps and deliverables of each step to developing the reinforcement learning SRL model.

digital learning environment (Fan et al., 2022). The learning environment comprised of a catalogue and navigation section, a reading material area, multiple learning aids such as annotation, planner, timer and search tools, and a writing panel situated at the bottom right corner of the screen, which can be opened or closed anytime (see Figure 3).

The 90 minute study used a pre-post design with a learning session in between. Specifically, the procedure comprised of the following successive steps: a pre-test aptitude, a pre-survey, a training session, a writing task, a post-test, a transfer-test, and a post-survey (van der Graaf et al., 2022). However, for this study we only made use of the data from the pre/post task aptitude tests and the writing task. Statistics from the assessments are presented in Table 1. The pre/post task aptitude test were used to gauge the level of knowledge in the topic before and after the writing task and had acceptable reliability scores (Kline, 2013),

$$\alpha = 0.60, \lambda_2 = 0.65, \omega_t = 0.68, \text{ at pretest and } \alpha = 0.59, \lambda_2 = 0.64, \omega_t = 0.66$$

at postest. The writing task involved a 45 minute session where they could selectively learn from and read more than 30 web pages on informative texts about three topics: artificial intelligence (AI), differentiation and scaffolds. The learners were instructed to use this information to write a vision essay of 300–400 words about the future of education in the year 2035. The essays were scored by human markers using the following criteria: coverage of topics from readings (9 points), essay cohesion (6 points), future vision (3 points) and word count (3 points). While the learners undertook the reading and writing task, data were collected from the following channels: (1) Navigation data, which stored simple navigational log data and time spent on

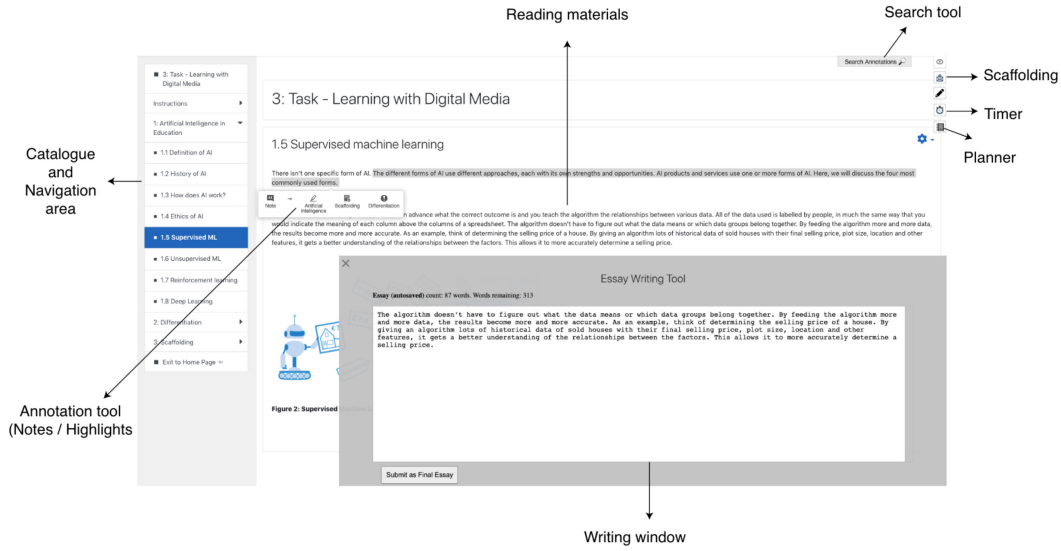

FIGURE 3 A screenshot of the learning environment used by the participants, including annotations of the tools available to the learners.

TABLE 1 Summary statistics of assessments used for developing RL algorithms detailed in Section "Analysis".

	Essay score	Pre-test score	Post-test score
M	0.66	0.55	0.68
SD	0.23	0.14	0.12

pages; (2) peripheral data, which stored data about mouse movements, mouse clicks on pages, mouse scroll and keyboard strokes; and (3) eye-tracking data from a screen-based eye-tracker (Tobii TX300), which was sampled at 300 Hz and consists of fixations, saccades, gaze points and pupil size.

The trace data events (navigation/peripheral/eye-tracking) were translated to learner actions (eg, a learner's click to create or edit a note during learning is indicative of a NOTE_EDITING action). The sequences of these learning actions could be mapped to one of the categories or subcategories mentioned in Table 2. In order to code the data, we followed the framework proposed by Bannert et al. (2014), which describes three major SRL categories: metacognition, cognition and motivation. Due to the difficulty in determining motivation from log data, the Motivation category was excluded from the coding process. Metacognition and cognition can be further decomposed into subcategories (see Table 2). Specifically, Metacognition consists of orientation, planning, monitoring and evaluation. Cognition is divided into first-reading; and re-reading and elaboration/organisation which require more complex processing (Lim et al., 2021, 2023). For example, when a learner highlighted a note while reading the instructions, this sequence of learning actions was labelled as GENERAL_INSTRUCTION<-> EDIT_ANNOTATION and could be mapped to the Orientation process (MC.O), as the learner was orientating towards the task's requirements. In cases where recorded actions could not be mapped to any of the proposed processes, it was labelled as No_Process. This coding process of trace data was checked for validity by using think aloud data in (Fan et al., 2022).

TABLE 2 In processing the data, sequences of learning actions were mapped to one of the following SRL processes listed in the table.

Main categories	SRL processes	Definitions
Metacognition (MC)	Orientation	Orientation on the learning-related activities; on prior knowledge; on the task and feeling about the task. Reading of general instruction and rubric of essay
	Planning	Planning of the reading and writing process by arranging activities and determining strategies. Proceeding to the next topic
	Monitoring	Monitoring and checking the reading and writing process; checking of progress according to instruction or plan
	Evaluation	Evaluation of the learning process; checking of content-wise correctness (eg, the essay content) of learning activities
Low_Cognition (LC)	First-reading	Reading information from the materials and superficial describing of pictorial representations for the first time
	Re-reading	Rereading of information in the text or figures
High_Cognition (HC)	Elaboration/organisation	Elaborate by connecting content-related comments and concepts during reading or writing; organising of content by creating an overview; write down information point by point in notes or essay window; summarising; adding information generated by oneself; and editing information by rephrasing or integrating information with prior knowledge

Note: The definitions of these processes are also presented.

Figure 4 presents a general overview of the processing pipeline. For a more detailed presentation of the process see (Fan et al., 2021).

Analysis

In reinforcement learning, in each episode an agent *A* interacts with an environment *E*, aiming at maximising the accumulated reward *R* along the action trajectories (see Table 3). Each trajectory starts at an initial state S_0 and ends at a final state S_T by doing actions a_i at each step *t* under a policy *P*. For this experiment, an episode was one 45-minute study session, the agent was the learner and the policy decided what SRL process (see Table 2 and below) to take based on the learner's current state.

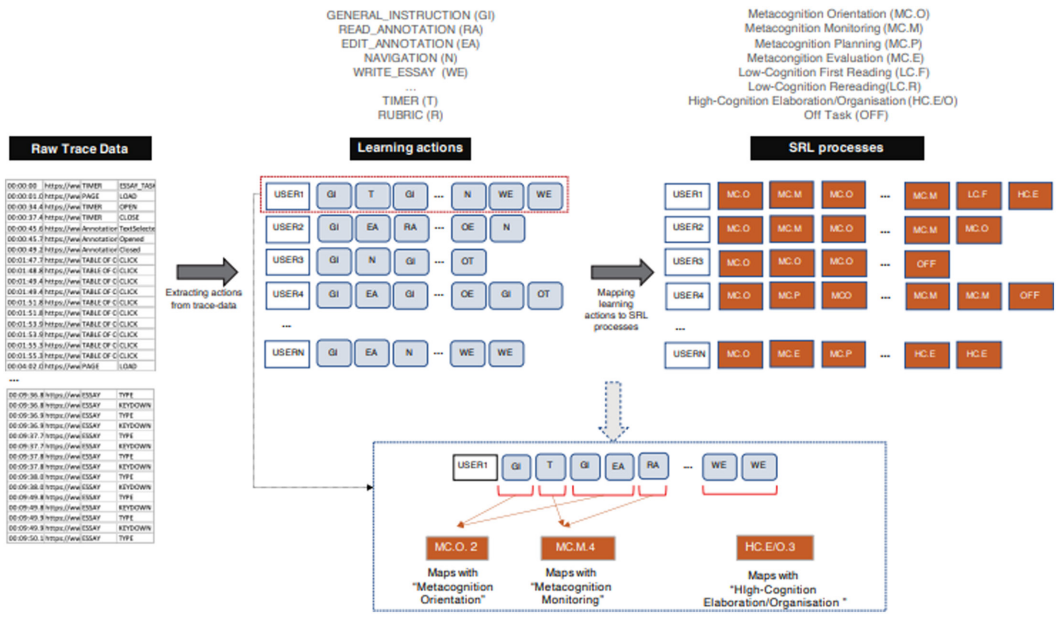

FIGURE 4 An overview of the pipeline of transforming the raw trace data into SRL processes.

TABLE 3 A translation of the reinforcement learning terms presented in the paper into their environmental equivalent.

Reinforcement learning term	Environment equivalent
Agent	Learner
Action	SRL process
State	Sequence of SRL processes taken so far in study session
Policy	Learner's SRL strategic decision making process
Reward	Assessment score or equivalent measure
Episode	Study session

Reward

In determining the reward function best suited for this study, we adopted SRL theories which state that:

1. SRL capabilities play an important role in academic achievement (Broadbent, 2017; Dent & Koenka, 2016; Theobald, 2021).
2. Low and high performing students tend to apply different SRL strategies (DiFrancesca et al., 2016; Proctor et al., 2006; Zhang & Cheng, 2019).
3. High-efficacy SRL learners are more likely to use a diverse range of SRL processes (Fan et al., 2021; Fincham et al., 2019; Nandagopal & Ericsson, 2012).

Hence, we utilised a long short-term memory (LSTM) neural network that maps sequences of SRL processes to learning gains and essay scores. An LSTM model was selected due to its effectiveness with handling sequential data; particularly in regard to predicting processes (Camargo et al., 2019; Tax et al., 2017). Learning gains were calculated by normalising (Marx & Cummings, 2007) the difference between pre-test knowledge scores as a percentage and post-test knowledge scores as a percentage (see Equation 1) (Table 4). The LSTM was implemented using Python Library, Tensorflow (Abadi et al., 2015)

$$\mathrm{NLG} = \begin{cases} \dfrac{\mathrm{post}-\mathrm{pre}}{100-\mathrm{pre}} & \mathrm{post} > \mathrm{pre} \\ 0 & \mathrm{post} = \mathrm{pre} \\ \dfrac{\mathrm{post}-\mathrm{pre}}{\mathrm{pre}} & \mathrm{post} < \mathrm{pre} \end{cases}. \tag{1}$$

The reward function (R) was defined by a weighted combination of the predicted essay assessment score (P) and normalised learning gain (NLG or learning gain) as well as a penalty for lack of diversity, measured using the entropy (H) of actions taken by the reinforcement learning agent, as defined in the formula below (see Equation 2). Entropy was calculated using Shannon entropy (Shannon, 1948) and a standard base of e. The weighted combination α illustrates the fact that learners can approach the study session with different priorities. We can adjust α to toggle the relative balance of the two priorities. For instance, a learner with α of 1, would have the main priority of maximising learning gains, and α at 0 would translate to a learner whose main priority was to perform best on the assessment. The reward would only be received at the end of each episode

$$R = \alpha^{*}\mathrm{NLG} + (1-\alpha) * W + \underbrace{\min\left[H-1, 0\right]}_{\text{Entropy Penalty}}. \tag{2}$$

TABLE 4 Statistics for learning gains obtained from human participants.

	Learning gains	
	Standard	**Normalised**
M	0.14	0.28
SD	0.14	0.25

FIGURE 5 An example of reducing actions into intervals. In our example, from time $t=0.00$ to $t=0.30$ a learner planned for 10 seconds, monitored for 15 seconds then planned again for 5 seconds. In this case, by summarising into 30 second intervals, we lose information on the sequential order of planning to monitoring to planning processes.

The performance of LSTM models can be negatively impacted by longer sequences due to issues such as exploding/vanishing gradients and information decay (Gers et al., 1999); hence, we tested reducing the length of SRL sequence vectors by summarising actions into intervals of varying length in seconds. For instance, using an interval of 30 seconds would translate to every row in the learner's SRL sequences vector being 30 seconds worth of SRL processes acted by the learner (see Figure 5). However, by summarising the sequences in this way, one can lose information on the sequences of processes acted by the learner. Hence, we compared the training loss, as measured by mean squared error, using intervals of 1, 15, 30, 45, 60 and 90 seconds. The training loss was minimised at intervals of 30 seconds (see Figure 6).

State

Similar to the vectors used to construct the LSTM predictor in Section "Reward", the learner's state is a matrix consisting of the SRL processes (see Table 2) taken so far in the current study session, at intervals of 30 seconds.

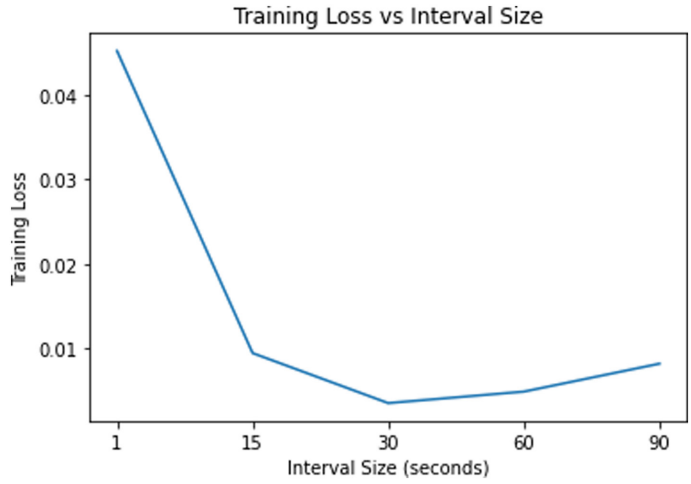

FIGURE 6 The recorded loss from training the LSTM neural network after reducing participants' SRL sequential data to intervals of 1, 15, 30, 45, 60 and 90 seconds.

Actions

To simplify the action selection process and reduce model complexity, we combined First-reading and re-reading into one process—reading. This significantly reduces model complexity and training time by simplifying the action selection process at the cost of a negligible change (0.001) in the LSTM reward function training loss. The total list of available actions is:

- Orientation—MC.O
- Planning—MC.P
- Monitoring—MC.M
- Evaluation—MC.E
- Reading—LC.R
- Elaboration/Organisation—HC.E/O
- No process—NP

Furthermore, to reduce the impact of model training on memory, we introduced a category of actions called magnitude. For each SRL action presented above, the agent determined a magnitude from 1 to the interval size of 30 seconds. This allowed the agent to skip up to 29 additional time steps if necessary. For instance, an output of SRL process orientation and magnitude 10 would translate to performing the orientation process for 10 seconds.

Algorithms

The reinforcement learning algorithms used in this study are Tensorforce (Kuhnle et al., 2017) implementations of actor-critic (AC), proximal policy optimisation (PPO), Q-learning (Q) and random. These are the major categories of algorithms available from Python library, Tensorforce (Kuhnle et al., 2017). The results from the best performing algorithm for this environment are used to draw insights and implications.

Actor-Critic

The AC algorithm (Mnih et al., 2016) consists of a "Critic" that estimates the value of a given action or state and an "Actor" which learns and updates the policy in the direction suggested by the Critic. Both the "actor" and the "critic" were parameterised using neural networks. This study used the advantage actor critic implementation.

Proximal policy optimisation

PPO (Schulman et al., 2017) improves the policy function while maintaining a "trust-region" to minimise the change in policy after each update by measuring the Kullback–Leibler divergence (Kullback & Leibler, 1951) of the updated policy and instilling large penalties for large changes. This ensures the policy is not too heavily swayed by outlier occurrences.

Q-learning

A Q agent will choose an action in each state based on a "Q-value", which is a weighted reward based on the expected highest long-term reward. Specifically, the Q-value of an action is the reward from taking a particular action plus the discounted sum of maximum expected rewards. Traditional Q-learning makes use of a look-up table to store these Q-values; however, in deep Q-Learning, a neural network is trained to parameterise the function which models this expected long term reward (Mnih et al., 2015).

Random

An agent who selects any action with equal probability.

Evaluation

We compared the performance of four reinforcement learning agents: Advantage actor critic (AC), proximal policy optimisation (PPO), Deep Q-learning (Q) and an agent which acts randomly (random). The agents were trained for a total of 5000 episodes. After each batch of 100 episodes of training, we evaluated the agent's performance and record its average reward, using 1000 test episodes. A baseline is included, which is the average reward of a random agent using 1000 test episodes.

Average Cumulative Reward

The Average Cumulative Reward (ACR) metric can be used to measure the effectiveness of a given Reinforcement Learning agent. ACR computes the average total reward an agent accumulates in executing a given task. To evaluate a given agent, we look at its ACR obtained from its test episodes. The developed SRL policy should be able to consistently attain higher ACR over the baseline random agent which does not necessarily select SRL processes in a strategic manner.

We will analyse the output of the best performing agents at each of the three levels of priority:

(i) Essay focused ($\alpha = 0$). The agent whose main priority is maximise essay scores.
(ii) Learning gain focused ($\alpha = 1$). The agent whose main priority is maximise learning gains.
(iii) Balanced focused ($\alpha = 0.5$). The agent whose main priority is to balance the maximisation of learning gains and essay scores.

At each of the three levels we examine:

Predicted Scores. We use the LSTM predictor described in Section "Reward" to analyse the predicted essay scores and learning gains.
Actions Distribution. We observe the distribution of the average timespan allocated to each SRL process during the agent's test run. Plots were created using Python library, Matplotlib (Hunter, 2007).
Epistemic network analysis: Epistemic network analysis (ENA) is a network analytic method used to analyse and visualise network data. ENA measures the connections and strength of connections between nodes in a network (Shaffer et al., 2009). ENA has previously been used to study the dynamics of SRL sequences by Saint, Gašević, et al. (2020). Using each SRL Process as a node, and the relationship measuring the relative frequency of co-occurrence of SRL processes, we can analyse the relationships and interplay between SRL processes generated by an agent's actions during its test run. The thicker the connecting lines between two processes, the more frequently they tend to succeed or precede each other. This can provide deeper insights into the strategy of SRL Process selection. ENA plots were created using rENA, an R Library (Marquart et al., n.d.).

RESULTS

In Table 5, we compare the performance of the various agents at different values of alpha. At an α of zero (an essay focused agent), the best performing agent was Q-learning with an average reward of 1.12. The PPO agent was also able to outperform a Random agent, by obtaining an average reward of 0.74. At every other value of α the best performing agent was PPO with an average reward of 0.57, 0.50, 0.43 and 0.81 at α values of 0.25, 0.50, 0.75 and 1.00, respectively. Hence, the PPO agent was the only algorithm able to consistently outperform a random agent at every value of α.

Using the results from Table 5, we can determine the best performing agents at the three main priority levels of focus. Namely, the PPO agent at the NLG and balanced focus; and the Q agent at the Essay focus. In Table 6, we observe the predicted essay score and learning gain for these best performing agents. As expected, the essay score and learning gain were maximised when the agents were trained to optimise for the respective focus. Furthermore, we also observe an inverse relationship between essay score and learning gain; ie, the greater the focus on maximising learning gains (higher α), the worse predicted essay score attained. However, despite the inverse relationship, the rate of change is notably different. As we move from an α of 0 to 0.5, we see an increase in learning gains of 0.15, with a decrease in essay score of 0.02. Likewise, as we move from an α of 0 to 1, we see an increase in learning gains of 0.56, with a decrease in essay score of 0.17. Essentially, the trade-off

TABLE 5 Performance of trained agents are compared at different values of α, which defines goal priority.

α	Q-learning	Actor-critic	PPO	Random
0.00	**1.12**	−0.66	0.74	0.67
0.25	0.50	−0.28	**0.57**	0.54
0.50	0.47	−0.19	**0.50**	0.41
0.75	0.16	−0.06	**0.43**	0.28
1.00	0.15	0.61	**0.81**	0.15

Note: Performance is measured using the highest average reward obtained from the test batches.
For each value of α, the highest average reward is boldened.

TABLE 6 Predicted essay scores and learning gains for the best performing agents at the three goal priority levels.

α	Description	Agent	Essay score	Learning gain
0.00	Essay focus	Q-learning	0.67	0.15
0.50	Balanced focus	PPO	0.65	0.30
1.00	Learning gain focus	PPO	0.50	0.71
–	Human participant average	–	0.69	0.28

Note: The human participant averages are also included for comparison's sake.

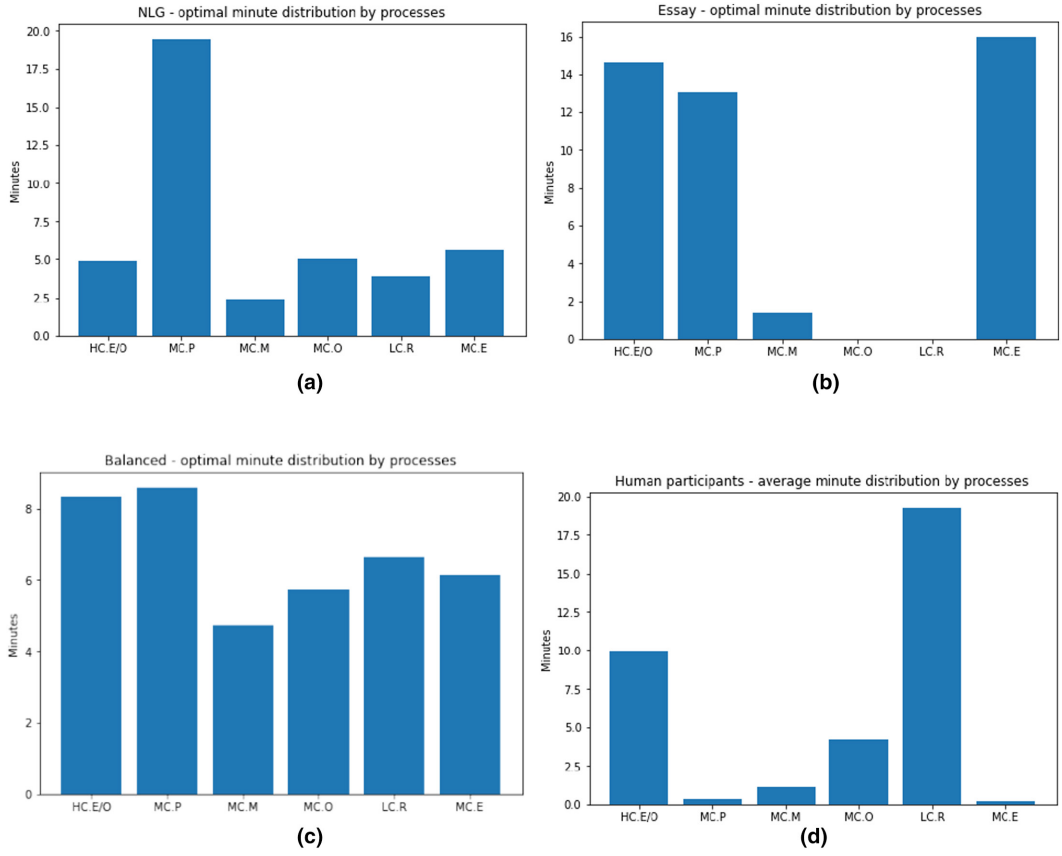

FIGURE 7 A plot of the average minutes' distributions of the best performing reinforcement learning agents at different goal priority levels as well as the average minutes distribution of human participants from the data collection study. The minute totals (*y*-axis) are shown for each of the SRL processes (*x*-axis). (a) Average minutes distribution of the trained learning gains oriented PPO agent. (b) Average minutes distribution of the trained essay oriented Q agent. (c) Average minutes distribution of the balanced oriented PPO agent which accounts for both essay score and learning gains during its best test run. (d) Average minutes distribution of human participants during data collection.

between essay score and learning gains is diminishing as the agent pursues a more learning gain focus.

Figure 7 illustrates the optimal minutes distribution as suggested by the best agents in their respective focus. In regard to learning gains (Figure 7a), the PPO agent allocated

a sizable amount of time to Planning and had a relatively even spread of minutes distributed across the remaining processes. Conversely, the essay score driven Q-learning agent showed less diversity in SRL strategy use, having allocated most of time between three processes: elaboration/organisation, planning and evaluation. The balanced focus agent had the most even spread of minutes allocation, with two highest minutes allocated to planning and elaboration/organisation. The average minute distribution of the human participants showed an uneven allocation, with most of the minutes going to reading and elaboration/organisation.

In Figure 8, we plot an ENA analysis of the network formed by the SRL processes used at the three levels of focus. A plot of the ENA analysis rotated according to accumulated means can be viewed in Figure S1. The learning gain focused agent had a mostly centralised network, with most processes co-occurring with Planning. The strongest of these connections were elaboration/organisation to planning, orientation to planning and evaluation to planning. The essay score oriented agent equally strong connections between planning, elaboration/organisation and evaluation, meaning each of these three processes were just as likely to co-occur. The balanced focus agent's network had planning and elaboration/organisation as the two largest nodes, as well as the strongest connection. The other processes were most likely to co-occur with either of these nodes. For example, reading was most likely to

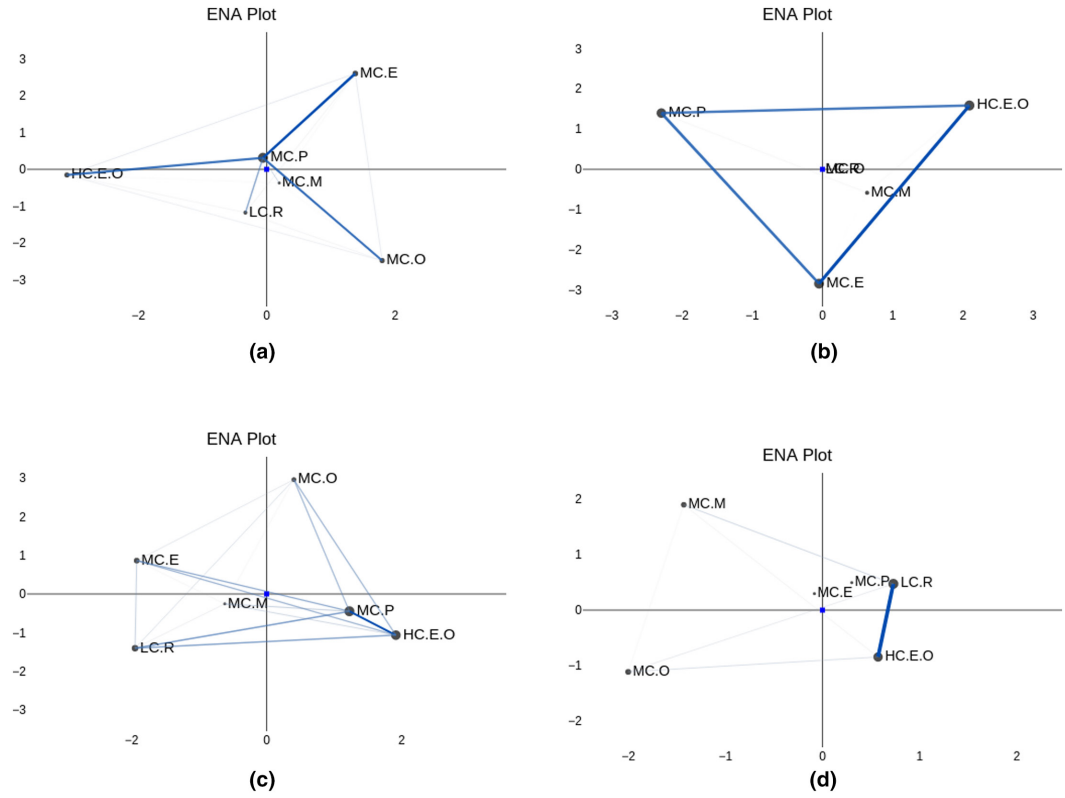

FIGURE 8 A plot of the epistemic network analysis of the best performing reinforcement learning agents at different goal priority levels as well as the human participants from the data collection study. (a) Epistemic network analysis of learning gains oriented PPO agent's actions during its test run. (b) Epistemic network analysis of trained essay score oriented Q agent's actions during its test run. (c) Epistemic network analysis of the balanced oriented PPO agent which accounts for both essay score and learning gains during its best test run. (d) Epistemic network analysis of human participants during data collection.

co-occur with either planning or elaboration/organisation. For all three focus levels, monitoring appears to have a nearly central location in the respective networks. The ENA plot for the average network of human participants had only strong connections between Reading and elaboration/organisation. Monitoring and orientation also had weak connections with both reading and elaboration/organisation processes.

DISCUSSION

In answering the research question to what extent a Reinforcement Learning agent can learn effective SRL strategy, we observed the developed agents were able to effectively select SRL processes so as to maximise a prescribed learning goal as measured by reward. In comparison with the average normalised learning gains of 0.28 and essay score of 0.69 observed from the human participants, the learning gain focused and balanced focus agents were able to attain better learning gains at 0.71 and 0.30, respectively. This may have been due to a greater emphasis on planning, monitoring and meta-cognitive processes as a whole. Specifically, in the minutes distribution chart (Figure 7), significantly more minutes were allocated to Planning, Monitoring and other processes under the MC (meta-cognitive) category by these agents, and in the network analysis (Figure 8) the agents had stronger connections of Planning with all other SRL processes, and had Monitoring processes plotted towards the centre of the graph. In Section "Introduction", we emphasised the key role planning, time-management and self-monitoring play in effective learning, and this is corroborated by the findings of our reinforcement learning experiment.

An additional reason for the higher learning gains obtained by the agents as compared to the results of the human participants may have been due to the greater diversity of SRL strategy use. There was a greater variety of minutes' allocations across all processes for the balanced and leaning gain oriented agents. Furthermore, ENA analysis showed a more connected network, as evidenced by more connection lines between the SRL processes. Literature has shown a more diverse use of SRL strategies can lead to improved learning outcomes (Fan et al., 2021; Fincham et al., 2019; Nandagopal & Ericsson, 2012). It is also noteworthy that the primary discerning factor between the balanced and learning gain focused agent was the much larger prominence of planning, and a larger allocation of minutes to all metacognitive processes (33 vs. 24 minutes). This is line with the findings of Broadbent (2017) and Theobald (2021) which view planning and metacognitive processes as higher-order regulation skills. Furthermore, when analysing the diversity of SRL strategies used at the three focus levels, we can infer the diversity of SRL strategy use is important for learning gains only up to a certain point, after which the use of higher-order regulation skills take precedence.

However, none of the trained agents were able to attain a higher predicted essay score than the average obtained from the human dataset. This may have been due to a lower incentive of the human participants to complete the essay task in an experiment setting in contrast to a graded assessment task leading to inconsistent signals being generated by the reward function. The lower incentive might have been evidenced by the essay score maximising agent which chose to primarily cycle its SRL processes between three processes—planning, elaboration/organisation, and evaluation (Figure 8b). In terms of learning actions this is akin to the learner strategically writing the essay and referring to the information source as they do so, without properly digesting the information through reading processes; essentially a tactical copy-paste strategy. It is also noteworthy the relative dearth of diversity in SRL processes acted by the agent focused solely on maximising the assessment score. Another implication of these findings could be related

to assessment design. Specifically, this suggests the designed task does not require a depth of understanding of the subject matter to attain high scores. The assessment designer can choose to rectify this by utilising materials that require a range of SRL processes in order to succeed.

LIMITATIONS

We were limited by the sise of our dataset, which could have reduced the accuracy of the reward signal to the reinforcement learning agent. However, the primary function of our reward function was to discern positive from negative behaviour, meaning our primary concern was with the polarity of the reward's function output; hence, this effect may have been partially mitigated. Future work will involve the use of larger datasets to improve the accuracy of the developed reward function.

The findings of this research will also need to be studied on other tasks and subject areas to enable more robust conclusions.

IMPLICATIONS FOR RESEARCH AND PRACTICE

This study highlighted the potential for reinforcement learning models to learn the optimal allocation of SRL strategies to maximise for a learning goal. The findings of this research can benefit the design of learning environments in numerous ways. Our findings corroborated that a student who prioritises maximising learning will enact a more diverse set of SRL processes. We also observed a trend of suboptimal use of metacognitive strategies amongst human participants as compared to the trained agents. This suggests a greater need for tools and resources which can facilitate the development of this skillset. Furthermore, we discovered potential limitations in the assessment design, given our essay optimised agent was able to attain high rewards with little to no diversity of SRL Processes as well as minimal reading processes and the competing influences of essay score and learning gains implying insufficient time allotment.

The implications of our work can enable the identification of sections of the course that pose a self-regulatory challenge and result in high cognitive load or sections that are too undemanding and hence fail to invoke or develop elaborate self-regulatory strategies. This can be considered for future course design. Furthermore, the ability to model SRL strategies using Reinforcement Learning can be extended to simulate or test SRL theories; for example, how will varying the experiment length in time affect the distribution of SRL Processes acted by the expert agent?

Our future work will involve adapting this algorithm to diagnose SRL strategy use in real time, enabling the detection of sub-optimal SRL ability in learning environments for use with a scaffolding strategy that improves SRL efficacy. It will also be possible to track a learner's SRL profile over time and assess whether course design is having positive effects on the learner's ability to self-regulate.

ACKNOWLEDGEMENTS

Open access publishing facilitated by Monash University, as part of the Wiley - Monash University agreement via the Council of Australian University Librarians.

FUNDING INFORMATION

None.

CONFLICT OF INTEREST STATEMENT

This study has no conflicts of interest.

DATA AVAILABILITY STATEMENT

Research data are not shared.

ETHICS STATEMENT

The study was approved by the Ethical Committee in the Faculty of Social Sciences of Radboud University for the FLoRA project and the Monash Human Research Ethics Committee under application number 32338.

ORCID

Ikenna Osakwe https://orcid.org/0000-0001-7672-4636
Guanliang Chen https://orcid.org/0000-0002-8236-3133
Mladen Rakovic https://orcid.org/0000-0002-1413-1103
Lyn Lim https://orcid.org/0000-0002-0617-5552
Johanna Moore https://orcid.org/0000-0001-7247-6823
Inge Molenaar https://orcid.org/0000-0003-4639-2524
Maria Bannert https://orcid.org/0000-0001-7045-2764
Alex Whitelock-Wainwright https://orcid.org/0000-0003-3033-4629
Dragan Gašević https://orcid.org/0000-0001-9265-1908

REFERENCES

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., … Zheng, X. (2015). *TensorFlow: Large-scale machine learning on heterogeneous systems*. https://www.tensorflow.org/

Aleven, V., Mclaren, B., Roll, I., & Koedinger, K. (2006, January). Toward meta-cognitive tutoring: A model of help seeking with a cognitive tutor. *International Journal of Artificial Intelligence in Education*, *16*, 101–128.

Azevedo, R., & Aleven, V. (2013). Metacognition and learning technologies: An overview of current interdisciplinary research. In R. Azevedo & V. Aleven (Eds.), *International handbook of metacognition and learning technologies* (pp. 1–16). Springer. https://doi.org/10.1007/978-1-4419-5546-3_1

Azevedo, R., & Gašević, D. (2019). Analyzing multimodal multichannel data about self-regulated learning with advanced learning technologies: Issues and challenges. *Computers in Human Behavior*, *96*, 207–210.

Azevedo, R., Harley, J., Trevors, G., Duffy, M., Feyzi-Behnagh, R., Bouchet, F., & Landis, R. (2013). Using trace data to examine the complex roles of cognitive, metacognitive, and emotional self regulatory processes during learning with multi-agent systems. In R. Azevedo & V. Aleven (Eds.), *International handbook of metacognition and learning technologies* (pp. 427–449). Springer.

Azevedo, R., Witherspoon, A. M., Strain, A. C., Burkett, C., & Fike, A. (2009). MetaTutor: A metacognitive tool for enhancing self-regulated learning. In *AAAI Fall Symposium: Cognitive and Metacognitive Educational Systems*. AAAI.

Baker, R., Xu, D., Park, J., Yu, R., Li, Q., Cung, B., Fischer, C., Rodriguez, F., Warschauer, M., & Smyth, P. (2020, April). The benefits and caveats of using clickstream data to understand student self-regulatory behaviors: Opening the black box of learning processes. *International Journal of Educational Technology in Higher Education*, *17*(1), 13. https://doi.org/10.1186/s41239-020-00187-1

Bannert, M., Reimann, P., & Sonnenberg, C. (2014, August). Process mining techniques for analysing patterns and strategies in students' self-regulated learning. *Metacognition and Learning*, *9*(2), 161–185. https://doi.org/10.1007/s11409-013-9107-6

Barnes, T., & Stamper, J. (2008). Toward automatic hint generation for logic proof tutoring using historical student data. In B. P. Woolf, E. Aïmeur, R. Nkambou, & S. Lajoie (Eds.), *Intelligent tutoring systems* (pp. 373–382). Springer Berlin Heidelberg.

Biswas, G., Baker, R. S., & Paquette, L. (2017). Data mining methods for assessing self-regulated learning. In *Handbook of self-regulation of learning and performance* (2nd ed., p. 16). Routledge.

Biswas, G., Jeong, H., Kinnebrew, J., Sulcer, B., & Roscoe, R. (2010, July). Measuring self-regulated learning skills through social interactions in a teachable agent environment. *Research and Practice in Technology Enhanced Learning*, *5*, 123–152. https://doi.org/10.1142/S1793206810000839

Blikstein, P., Worsley, M., Piech, C., Sahami, M., Cooper, S., & Koller, D. (2014, October). Programming pluralism: Using learning analytics to detect patterns in the learning of computer programming. *Journal of the Learning Sciences*, *23*(4), 561–599. https://doi.org/10.1080/10508406.2014.954750

Bouchet, F., Azevedo, R., Kinnebrew, J. S., & Biswas, G. (2012). *Identifying students' characteristic learning behaviors in an intelligent tutoring system fostering self-regulated learning*. International Educational Data Mining Society. https://eric.ed.gov/?id=ED537188

Broadbent, J. (2017, April). Comparing online and blended learner's self-regulated learning strategies and academic performance. *The Internet and Higher Education*, *33*, 24–32. https://www.sciencedirect.com/science/article/pii/S1096751617300398, https://doi.org/10.1016/j.iheduc.2017.01.004

Broadbent, J., & Poon, W. L. (2015, October). Self-regulated learning strategies & academic achievement in online higher education learning environments: A systematic review. *The Internet and Higher Education*, *27*, 1–13. https://www.sciencedirect.com/science/article/pii/S1096751615000251, https://doi.org/10.1016/j.iheduc.2015.04.007

Butler, D. L., & Winne, P. H. (1995, September). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research*, *65*(3), 245–281. https://doi.org/10.3102/00346543065003245

Camargo, M., Dumas, M., & González-Rojas, O. (2019). Learning accurate LSTM models of business processes. In T. Hildebrandt, B. F. van Dongen, M. Röglinger, & J. Mendling (Eds.), *Business process management* (pp. 286–302). Springer International Publishing. https://doi.org/10.1007/978-3-030-26619-6_19

Cerezo, R., Bogarín, A., Esteban, M., & Romero, C. (2020, April). Process mining for self-regulated learning assessment in elearning. *Journal of Computing in Higher Education*, *32*(1), 74–88. https://doi.org/10.1007/s12528-019-09225-y

Cerezo, R., Esteban, M., Sánchez-Santillán, M., & Núñez, J. C. (2017). Procrastinating behavior in computer-based learning environments to predict performance: A case study in Moodle. *Frontiers in Psychology*, *8*, 1403. https://doi.org/10.3389/fpsyg.2017.01403

Chi, M. T., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist*, *49*(4), 219–243.

Dent, A. L., & Koenka, A. C. (2016, September). The relation between self-regulated learning and academic achievement across childhood and adolescence: A meta-analysis. *Educational Psychology Review*, *28*(3), 425–474. https://doi.org/10.1007/s10648-015-9320-8

DiFrancesca, D., Nietfeld, J. L., & Cao, L. (2016, January). A comparison of high and low achieving students on selfregulated learning variables. *Learning and Individual Differences*, *45*, 228–236. https://www.sciencedirect.com/science/article/pii/S1041608015300133, https://doi.org/10.1016/j.lindif.2015.11.010

Doroudi, S., Aleven, V., & Brunskill, E. (2019, December). Where's the reward? *International Journal of Artificial Intelligence in Education*, *29*(4), 568–620. https://doi.org/10.1007/s40593-019-00187-x

Fahid, F. M., Rowe, J. P., Spain, R. D., Goldberg, B. S., Pokorny, R., & Lester, J. (2021). Adaptively scaffolding cognitive engagement with batch constrained deep Q networks. In I. Roll, D. McNamara, S. Sosnovsky, R. Luckin, & V. Dimitrova (Eds.), *Artificial intelligence in education* (Vol. *12748*). Springer International Publishing.

Fan, Y., Matcha, W., Uzir, N. A., Wang, Q., & Gašević, D. (2021, December). Learning analytics to reveal links between learning design and self-regulated learning. *International Journal of Artificial Intelligence in Education*, *31*(4), 980–1021. https://doi.org/10.1007/s40593-021-00249-z

Fan, Y., van der Graaf, J., Lim, L., Rakovic, M., Kilgour, J., Moore, J., Molenaar, I., Bannert, M., & Gašević, D. (2022). Improving the measurement of self-regulated learning using multi-channel data. *Metacogniton and Learning*, *17*, 1025–1055. https://doi.org/10.1007/s11409-022-09304-z

Fan, Y., van der Graaf, J., Lim, L., Rakovic, M., Singh, S., Kilgour, J., Moore, J., Molenaar, I., Bannert, M., & Gašević, D. (2022). Towards investigating the validity of measurement of self-regulated learning based on trace data. *Metacognition and Learning*, *17*, 949–987. https://doi.org/10.1007/s11409-022-09291-1

Feyzi-Behnagh, R., Azevedo, R., Legowski, E., Reitmeyer, K., Tseytlin, E., & Crowley, R. S. (2014, March). Metacognitive scaffolds improve self-judgments of accuracy in a medical intelligent tutoring system. *Instructional Science*, *42*(2), 159–181. https://doi.org/10.1007/s11251-013-9275-4

Fincham, E., Gašević, D., Jovanović, J., & Pardo, A. (2019, January). From study tactics to learning strategies: An analytical method for extracting interpretable representations. *IEEE Transactions on Learning Technologies*, *12*(1), 59–72. https://doi.org/10.1109/TLT.2018.2823317

Gagniuc, P.. (Ed.). (2017). From Observation to Simulation. *Markov chains*. (pp. 9–24). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119387596

Galyardt, A., & Goldin, I. (Eds.). (2015). Evaluating simplicial mixtures of Markov chains for modeling student metacognitive strategies. *Quantitative psychology research* (pp. 377–393). Springer.

Gers, F., Schmidhuber, J., & Cummins, F. (1999). Learning to forget: Continual prediction with LSTM. In *1999 Ninth International Conference on Artificial Neural Networks ICANN 99. (Conf. Publ. No. 470)* (Vol. 2, pp. 850–855). IET. https://doi.org/10.1049/cp:19991218

Hadwin, A., Nesbit, J., Jamieson-Noel, D., Code, J., & Winne, P. (2007, December). Examining trace data to explore selfregulated learning. *Metacognition and Learning*, *2*, 107–124. https://doi.org/10.1007/s11409-007-9016-7

Hattie, J., & Timperley, H. (2007, March). The power of feedback. *Review of Educational Research*, *77*(1), 81–112. https://doi.org/10.3102/003465430298487

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Intayoad, W., Kamyod, C., & Temdee, P. (2020, December). Reinforcement learning based on contextual bandits for personalized online learning recommendation systems. *Wireless Personal Communications*, *115*(4), 2917–2932. https://doi.org/10.1007/s11277-020-07199-0

Johnson, S., & Zaiane, O. R. (2012). *Deciding on feedback polarity and timing*. EDM.

Jovanović, J., Gašević, D., Dawson, S., Pardo, A., & Mirriahi, N. (2017, April). Learning analytics to unveil learning strategies in a flipped classroom. *The Internet and Higher Education*, *33*, 74–85. https://www.sciencedirect.com/science/article/pii/S1096751617300684, https://doi.org/10.1016/j.iheduc.2017.02.001

Kinnebrew, J., & Biswas, G. (2011, January). Modeling and measuring self-regulated learning in teachable agent environments. *Journal of E-Learning and Knowledge Society*, *7*, 19–35. https://doi.org/10.20368/1971-8829/518

Kizilcec, R. F., & Schneider, E. (2015, March). Motivation as a lens to understand online learners: Toward data-driven design with the OLEI scale. *ACM Transactions on Computer-Human Interaction*, *22*(2), 6:1–6:24. https://doi.org/10.1145/2699735

Kline, P. (2013). *Handbook of psychological testing*. Routledge.

Kovanović, V., Gašević, D., Joksimović, S., Hatala, M., & Adesope, O. (2015, October). Analytics of communities of inquiry: Effects of learning technology use on cognitive presence in asynchronous online discussions. *The Internet and Higher Education*, *27*, 74–89. https://www.sciencedirect.com/science/article/pii/S1096751615000421, https://doi.org/10.1016/j.iheduc.2015.06.002

Kuhnle, A., Schaarschmidt, M., & Fricke, K. (2017). *Tensorforce: A TensorFlow library for applied reinforcement learning*. https://github.com/tensorforce/tensorforce

Kullback, S., & Leibler, R. A. (1951, March). On information and sufficiency. *The Annals of Mathematical Statistics*, *22*(1), 79–86. https://doi.org/10.1214/aoms/1177729694

Li, Q., Baker, R., & Warschauer, M. (2020, April). Using clickstream data to measure, understand, and support self-regulated learning in online courses. *The Internet and Higher Education*, *45*, 100727. https://www.sciencedirect.com/science/article/pii/S1096751620300038, https://doi.org/10.1016/j.iheduc.2020.100727

Liang, Z., Mu, L., Chen, J., & Xie, Q. (2022, July). Graph path fusion and reinforcement reasoning for recommendation in MOOCs. *Education and Information Technologies*, *28*, 525–545. https://doi.org/10.1007/s10639-022-11178-2

Lim, L., Bannert, M., van der Graaf, J., Molenaar, I., Fan, Y., Kilgour, J., Moore, J., & Gašević, D. (2021). Temporal assessment of self-regulated learning by mining students' think-aloud protocols. *Frontiers in Psychology*, *12*, 749749. https://doi.org/10.3389/fpsyg.2021.749749

Lim, L., Bannert, M., van der Graaf, J., Singh, S., Fan, Y., Surendrannair, S., Rakovic, M., Molenaar, I., Moore, J., & Gašević, D. (2023). Effects of real-time analytics-based personalized scaffolds on students' self-regulated learning. *Computers in Human Behavior*, *139*, 107547.

Luo, S. (2020, June). Dynamic scheduling for flexible job shop with new job insertions by deep reinforcement learning. *Applied Soft Computing*, *91*, 106208. https://www.sciencedirect.com/science/article/pii/S1568494620301484, https://doi.org/10.1016/j.asoc.2020.106208

Lust, G., Vandewaetere, M., Ceulemans, E., Elen, J., & Clarebout, G. (2011, November). Tool-use in a blended undergraduate course: In search of user profiles. *Computers & Education*, *57*(3), 2135–2144. https://www.sciencedirect.com/science/article/pii/S0360131511001163, https://doi.org/10.1016/j.compedu.2011.05.010

Mabroukeh, N. R., & Ezeife, C. I. (2010, December). A taxonomy of sequential pattern mining algorithms. *ACM Computing Surveys*, *43*(1), 3:1–3:41. https://doi.org/10.1145/1824795.1824798

Maldonado-Mahauad, J., Pérez-Sanagustín, M., Kizilcec, R. F., Morales, N., & Munoz-Gama, J. (2018, March). Mining theory- based patterns from Big data: Identifying self-regulated learning strategies in massive open online courses. *Computers in Human Behavior*, *80*, 179–196. https://www.sciencedirect.com/science/article/pii/S0747563217306477, https://doi.org/10.1016/j.chb.2017.11.011

Marquart, L. C., Swiecki, Z., Collier, W., Eagan, B., Woodward, R., & Shaffer, D. W. (n.d.). *rENA: Epistemic Network Analysis* (Version Number: 0.2.2.0).

Marx, J. D., & Cummings, K. (2007, January). Normalized change. *American Journal of Physics*, *75*(1), 87–91. https://doi.org/10.1119/1.2372468

Matcha, W., Gašević, D., Uzir, N. A., Jovanovic, J., & Pardo, A. (2019, March). Analytics of learning strategies: Associations with academic performance and feedback. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge* (pp. 461–470). Association for Computing Machinery. https://doi.org/10.1145/3303772.3303787

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T. P., Harley, T., Silver, D., & Kavukcuoglu, K. (2016). *Asynchronous methods for deep reinforcement learning.* (_eprint: 1602.01783).

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., & Hassabis, D. (2015, February). Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529–533. https://doi.org/10.1038/nature14236

Molenaar, I., & Järvelä, S. (2014). Sequential and temporal characteristics of self and socially regulated learning. *Metacognition and Learning*, *9*(2), 75–85.

Nandagopal, K., & Ericsson, K. A. (2012, October). An expert performance approach to the study of individual differences in self-regulated learning activities in upper-level college students. *Learning and Individual Differences*, *22*(5), 597–609. https://www.sciencedirect.com/science/article/pii/S1041608011001798, https://doi.org/10.1016/j.lindif.2011.11.018

Nguyen, H., & La, H. (2019). *Review of deep reinforcement learning for robot manipulation* (p. 595). IEEE. https://doi.org/10.1109/IRC.2019.00120

Panadero, E. (2017). A review of self-regulated learning: Six models and four directions for research. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.00422

Proctor, B., Prevatt, F., Adams, K., Hurst, A., & Petscher, Y. (2006, January). Study skills profiles of normal-achieving and academically-struggling college students. *Journal of College Student Development*, *47*, 37–51. https://doi.org/10.1353/csd.2006.0011

Rabin, E., Silber-Varod, V., Kalman, Y. M., & Kalz, M. (2019). Identifying learning activity sequences that are associated with high intention-fulfillment in MOOCs. In M. Scheffel, J. Broisin, V. Pammer-Schindler, A. Loannou, & J. Schneider (Eds.), *Transforming learning with meaningful technologies* (pp. 224–235). Springer International Publishing. https://doi.org/10.1007/978-3-030-29736-7_17

Romero, C., & Ventura, S. (2013). Data mining in education. *WIREs Data Mining and Knowledge Discovery*, *3*(1), 12–27. https://doi.org/10.1002/widm.1075

Saint, J., Fan, Y., Gašević, D., & Pardo, A. (2022). Temporally-focused analytics of self-regulated learning: A systematic review of literature. *Computers and Education: Artificial Intelligence*, *3*, 100060.

Saint, J., Fan, Y., Singh, S., Gasevic, D., & Pardo, A. (2021, April). Using process mining to analyse self-regulated learning: A systematic analysis of four algorithms. In *LAK21: 11th International Learning Analytics and Knowledge Conference* (pp. 333–343). Association for Computing Machinery. https://doi.org/10.1145/3448139.3448171

Saint, J., Gašević, D., Matcha, W., Uzir, N. A., & Pardo, A. (2020, March). Combining analytic methods to unlock sequential and temporal patterns of self-regulated learning. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge* (pp. 402–411). Association for Computing Machinery. https://doi.org/10.1145/3375462.3375487

Saint, J., Whitelock-Wainwright, A., Gasevic, D., & Pardo, A. (2020). Trace-SRL: A framework for analysis of microlevel processes of self-regulated learning from trace data. *IEEE Transactions on Learning Technologies*, *13*(4), 861–877. https://doi.org/10.1109/TLT.2020.3027496

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017, July). *Proximal policy optimization algorithms*. http://arxiv.org/abs/1707.06347v2

Shaffer, D., Hatfield, D., Svarovsky, G., Nash, P., Nulty, A., Bagley, E., Frank, K. A., Rupp, A. A., & Mislevy, R. (2009, May). Epistemic network analysis: A prototype for 21st-century assessment of learning. *International Journal of Learning and Media*, *1*, 33–53. https://doi.org/10.1162/ijlm.2009.0013

Shannon, C. E. (1948, July). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Shao, K., Tang, Z., Zhu, Y., Li, N., & Zhao, D. (2019, December). *A survey of deep reinforcement learning in video games*. http://arxiv.org/abs/1912.10944v2

Srivastava, N., Fan, Y., Rakovic, M., Singh, S., Jovanovic, J., van der Graaf, J., Lim, L., Surendrannair, S., Kilgour, J., Molenaar, I., Bannert, M., Moore, J. D., & Gasevic, D. (2022, March). Effects of internal and external conditions on strategies of self-regulated learning: A learning analytics study. In *LAK22: 12th International Learning Analytics and Knowledge Conference* (pp. 392–403). ACM. https://doi.org/10.1145/3506860.3506972

Tax, N., Verenich, I., La Rosa, M., & Dumas, M. (2017). Predictive business process monitoring with LSTM neural networks. In E. Dubois & K. Pohl (Eds.), *Advanced information systems engineering* (pp. 477–492). Springer International Publishing. https://doi.org/10.1007/978-3-319-59536-8_30

Theobald, M. (2021, July). Self-regulated learning training programs enhance university students' academic performance, self-regulated learning strategies, and motivation: A meta-analysis. *Contemporary Educational Psychology*, *66*, 101976. https://www.sciencedirect.com/science/article/pii/S0361476X21000357, https://doi.org/10.1016/j.cedpsych.2021.101976

van der Aalst, W., Weijters, T., & Maruster, L. (2004, September). Workflow mining: Discovering process models from event logs. *IEEE Transactions on Knowledge and Data Engineering*, *16*(9), 1128–1142. https://doi.org/10.1109/TKDE.2004.47

van der Graaf, J., Lim, L., Fan, Y., Kilgour, J., Moore, J., Gašević, D., Bannert, M., & Molenaar, I. (2022, December). The dynamics between self-regulated learning and learning outcomes: An exploratory approach and implications. *Metacognition and Learning*, *17*(3), 745–771. https://doi.org/10.1007/s11409-022-09308-9

Winne, P. (2013, January). Learning strategies, study skills, and self-regulated learning in postsecondary education. In *Higher education: Handbook of theory and research* (pp. 377–403). Springer. https://doi.org/10.1007/978-94-007-5836-0_8

Winne, P. H. (2020). Construct and consequential validity for learning analytics based on trace data. *Computers in Human Behavior*, *112*, 106457.

Winne, P. H. (2021, August). *Cognition, metacognition, and self-regulated learning*. https://doi.org/10.1093/acrefore/9780190264093.013.1528

Winne, P. H., Hadwin, A. F., & Graesser, A. C. (1998). Studying as self-regulated learning. In *Metacognition in educational theory and practice* (pp. 277–304). Lawrence Erlbaum Associates. http://site.ebrary.com/id/10349656 (OCLC: 44957327).

Wong, P. C. M., Vuong, L. C., & Liu, K. (2017, April). Personalized learning: From neurogenetics of behaviors to designing optimal language training. *Neuropsychologia*, *98*, 192–200. https://www.sciencedirect.com/science/article/pii/S0028393216303669, https://doi.org/10.1016/j.neuropsychologia.2016.10.002

Zhang, J., Hao, B., Chen, B., Li, C., Chen, H., & Sun, J. (2019, July). Hierarchical reinforcement learning for course recommendation in MOOCs. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(1), 435–442. https://doi.org/10.1609/aaai.v33i01.3301435

Zhang, T., & Mo, H. (2021, May). Reinforcement learning for robot research: A comprehensive review and open issues. *International Journal of Advanced Robotic Systems*, *18*(3), 17298814211007305. https://doi.org/10.1177/17298814211007305

Zhang, X., & Cheng, H. N. H. (2019, August). Mining the patterns of graduate students' self-regulated learning behaviors in a negotiated online academic Reading assessment. In *Proceedings of the 2019 3rd International Conference on E-Society, E-Education and E-Technology* (pp. 109–114). Association for Computing Machinery. https://doi.org/10.1145/3355966.3355977

Zheng, S., Rosson, M. B., Shih, P. C., & Carroll, J. M. (2015, February). Understanding student motivation, behaviors and perceptions in MOOCs. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (pp. 1882–1895). Association for Computing Machinery. https://doi.org/10.1145/2675133.2675217

Zimmerman, B. J. (2013, July). From cognitive modeling to self-regulation: A social cognitive career path. *Educational Psychologist*, *48*(3), 135–147. https://doi.org/10.1080/00461520.2013.794676

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Osakwe, I., Chen, G., Fan, Y., Rakovic, M., Singh, S., Lim, L., van der Graaf, J., Moore, J., Molenaar, I., Bannert, M., Whitelock-Wainwright, A., & Gašević, D. (2024). Towards prescriptive analytics of self-regulated learning strategies: A reinforcement learning approach. *British Journal of Educational Technology*, *00*, 1–25. https://doi.org/10.1111/bjet.13429