Robustness assessment of energy policies: Reinforcement learning approach to support myopic energy transition

Xavier Rixhon[1,2,*], Hervé Jeanmart[1], and Francesco Contino[1]

[1]Institute of Mechanics, Materials and Civil Engineering (iMMC), Université catholique de Louvain (UCLouvain), Place du Levant, 2, 1348 Louvain-la-Neuve, Belgium
[2]Lead contact
[*]Correspondence: xavier.rixhon@uclouvain.be

1 Context and scale

SUMMARY

The summary (abstract) should consist of a single paragraph of 150 words or fewer.

KEYWORDS

Whole-energy systems, transition pathways, energy policy, reinforcement learning, optimisation, uncertainty, myopic decision-making

INTRODUCTION

To help decision makers in paving the way of the energy transition and setting investment plans to meet the climate targets, Energy System Optimisation Models (ESOM) explore possibilities the decision makers rely models to invest this transition pathway

On top of dealing with these uncertainties, the reality of decision-makers leads to a limited foresight into the future[1]. In a perfect foresight approach, decision-makers would be able to, from now on, see the “finish-line of the transition” in 2050 and accordingly make the planning decisions once and for all. On the contrary, they uncover the realisation of these uncertainties step-by-step, in a “myopic” way, and progressively act on them to, hopefully, meet the set target to reduce the anthropogenic Greenhouse Gas (GHG). In the objective to respect an overall CO_2 budget rather than to follow a prescribed CO_2 emissions trajectory, there is a need for a framework to explore these multiple transition pathways and provide insight into intermediate milestones not to miss. On top of the “what to do?”, this framework would aim at helping the policymakers to answer the question “how to do it?”.

Due to the increasing complexity of the systems and the integration of uncertainties, the last decades have seen the emergence of publications where Reinforcement Learning (RL) is applied to energy systems[2,3]. In their respective reviews, Cao et al.[2] and Perera and Kamalaruban[3] highlighted groups of problems addressed with RL in the research field of energy systems: building energy management system (BEMS), optimisation of dispatch and operational control closely linked with the energy market and the optimal power flow problem in the grid, micro-grid management, electro-mobility or even demand-side management or optimal control of energy system devices like maximum power point tracking (MPPT) of wind turbines and Photovoltaic (PV) panels.

In this paper, we propose a novel method based on the application of RL to a new kind of energy system problem: the optimisation of the transition pathway of a whole-energy system. In this sense, the objective is to optimise a policy to support this transition in myopic conditions and subject to uncertainties and a CO_2 budget rather than a prescribed GHG emissions trajectory (see Figure 1). This method highlights the importance of taking short-term actions and points out the no-go zones where succeeding the energy transition is very unlikely.

Even though the following messages result from studies on Belgium, the trends can be transposed to other countries with a high demand and low renewable potentials such as the Netherlands or Germany[4,5]. After showing the convergence of the learning process, intermediate milestones and effective energy policies are pointed out.

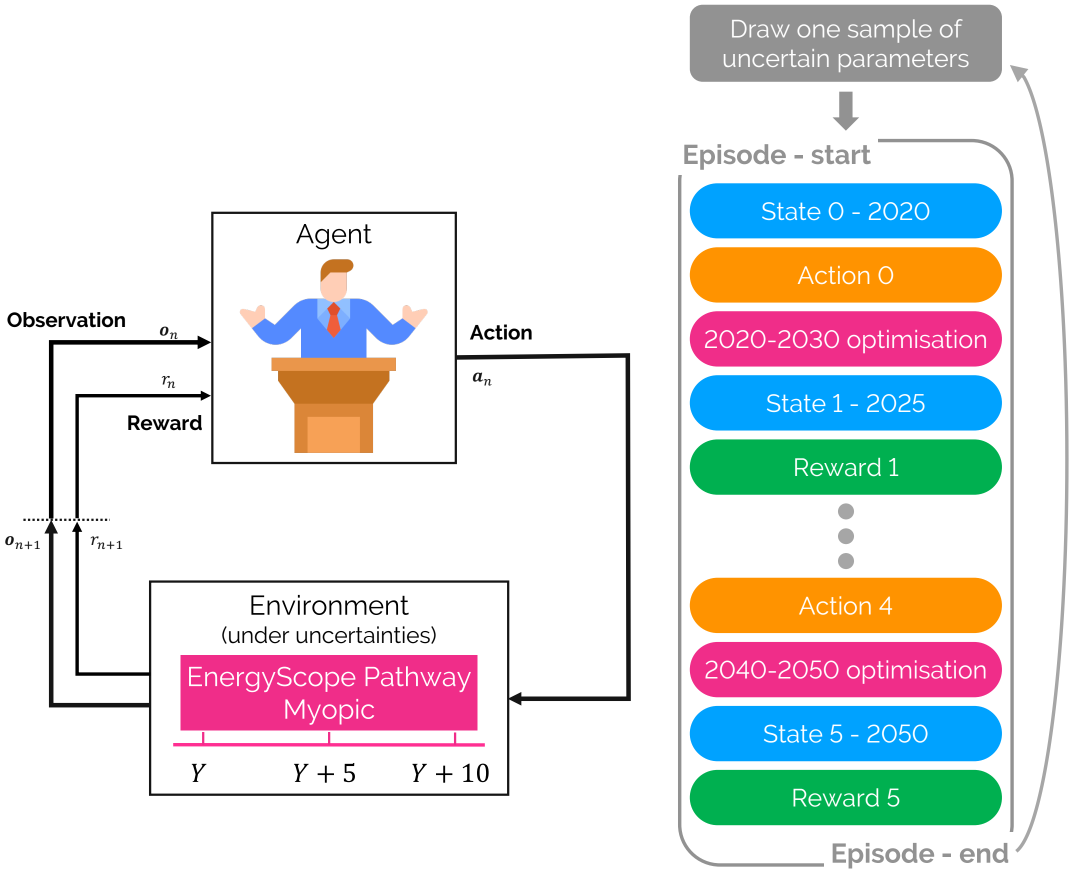

Figure 1: The Reinforcement Learning (RL) framework applied to the myopic optimisation of the energy transition pathway between 2020 and 2050. Here, the agent interacts with its environment, i.e., the energy-system model on a limited decision window of 10 years. At the beginning of each episode, a different sample of uncertain parameters is drawn and affects the environment, EnergyScope Pathway, according to the methodology detailed in Section **??**.

RESULTS

Reward and success

The learning phase has been split into batches of 500 steps, i.e., 500 sequences of state-action-reward-new state. At the end of each batch, the up-to-date policy, i.e., the Neural Network (NN), is saved. This way, we can assess the progress in the learning process and its convergence (see Figure 2). The mean reward increases rapidly at the beginning of the learning process before reaching a plateau where the optimisation of the policy becomes more marginal. As these successes indirectly drive the agent's optimisation, it shows that the reward function (see Figure 8) leads towards more and more successes. However, given the wide range of uncertainty of some parameters and the agent's levers of action, this success rate stays limited at the end of the learning process. In other words, there are conditions where it is impossible for the agent to succeed the transition.

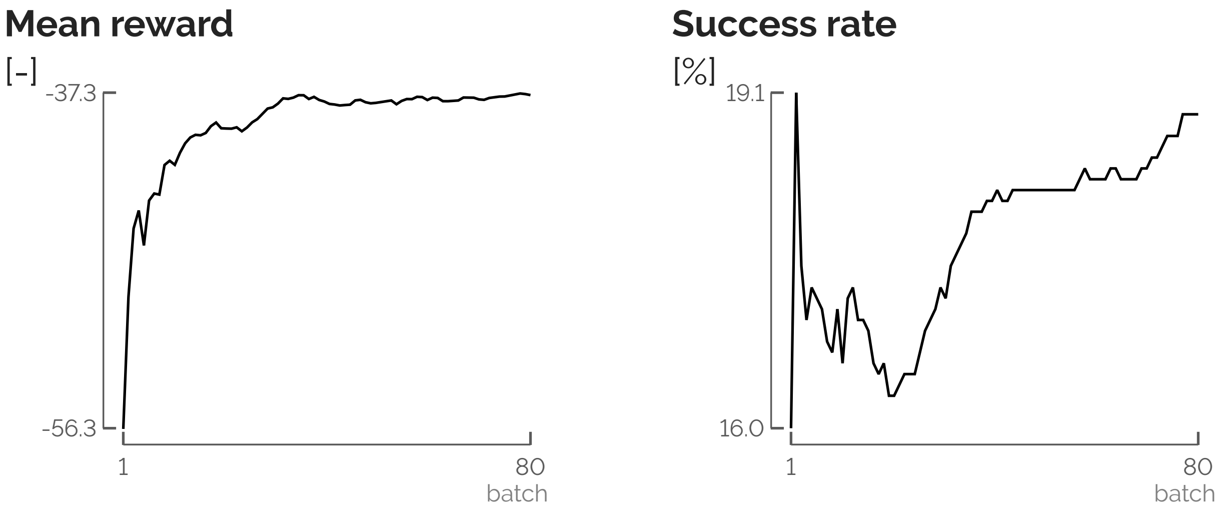

Figure 2: Mean reward and success rate of the different learning batches. The stabilisation of the reward curve shows a convergence of the learning process from the agent's point of view. The evolution of the success rate also shows that the reward function aims at more and more successful transitions.

When assessing the distributions of the values of reward in the failure and success cases, one notices a range where these distributions overlap (see left-hand side of Figure 3). This area corresponds to either transitions that exceed the CO_2 budget in 2050 but are cheaper than the total transition cost of reference (see Section **??**) or successful transitions that are more expensive. Besides this overlap, we observe that successes account for the majority of the cases with higher rewards. This is another indication that the reward function is appropriate in this exploration of successful transition pathways. More indirectly, in case of applying this methodology to another case study, this substantiates that the weights between emissions and cost defined in Section 1 could be used as an initial step and fine-tuned afterwards.

Considering the end of the time window where the CO_2 budget is exceeded, the right-hand side of Figure 3 shows that 2040 is the "tipping year" for the agent. Beyond this point, through this learning process, the chances to succeed the transition were 38%. In other words, near-term (2025-2030) actions are necessary to hope to succeed the transition.

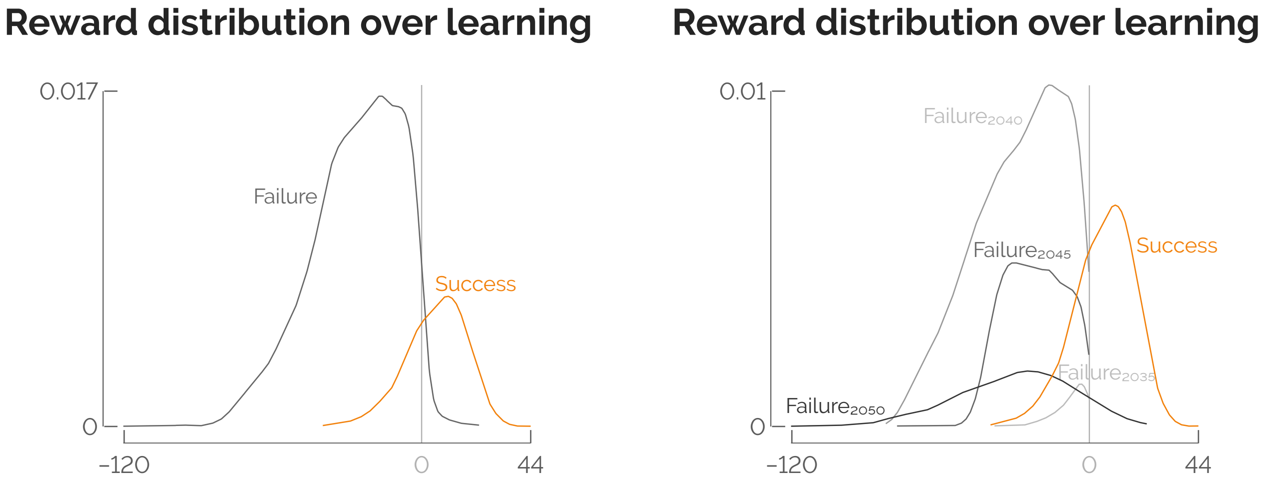

Figure 3: Reward distribution between successes and failures. Graph on the right-hand side details at the end of which time window the failure occurred. The "tipping year" is 2040 as failing the transition by 2040 represents 57% of all the failures. Beyond this point, through this learning process, succeeding the transition represents 38% of the episodes.

States

The first two dimensions of the state space are the cumulative emissions and costs. They drive the value of the reward and, consequently, the optimisation of the agent's policy. Per definition, the threshold of $1.2\,\mathrm{Gt}_{\mathrm{CO_2,eq}}$ splits the episodes reaching 2050 into successes and failures (see Figure 4). Since infeasible cases or those that overshoot the CO_2 budget are discarded before 2050 (see Section 1), the number of attempts that reach further steps in the transition progressively decreases. Consequently, the share of successful transitions compared to failures progressively increases with time.

In the successful transitions, the median cumulative emissions, P_{50}, are about $0.9\,\mathrm{Gt}_{\mathrm{CO_2,eq}}$. Reaching cumulative emissions significantly lower than the CO_2 budget is possible thanks to efforts made at earlier stages of the transition and the potential to install Small Modular Reactor (SMR) later on. Considering the failures in 2050, half of these episodes ended up with cumulative emissions lower or equal to $1.4\,\mathrm{Gt}_{\mathrm{CO_2,eq}}$. As illustrated in Figure 3, 2040 is identified as the tipping year. Where 98% of the failures were below the CO_2 budget in 2035, only 37% passed

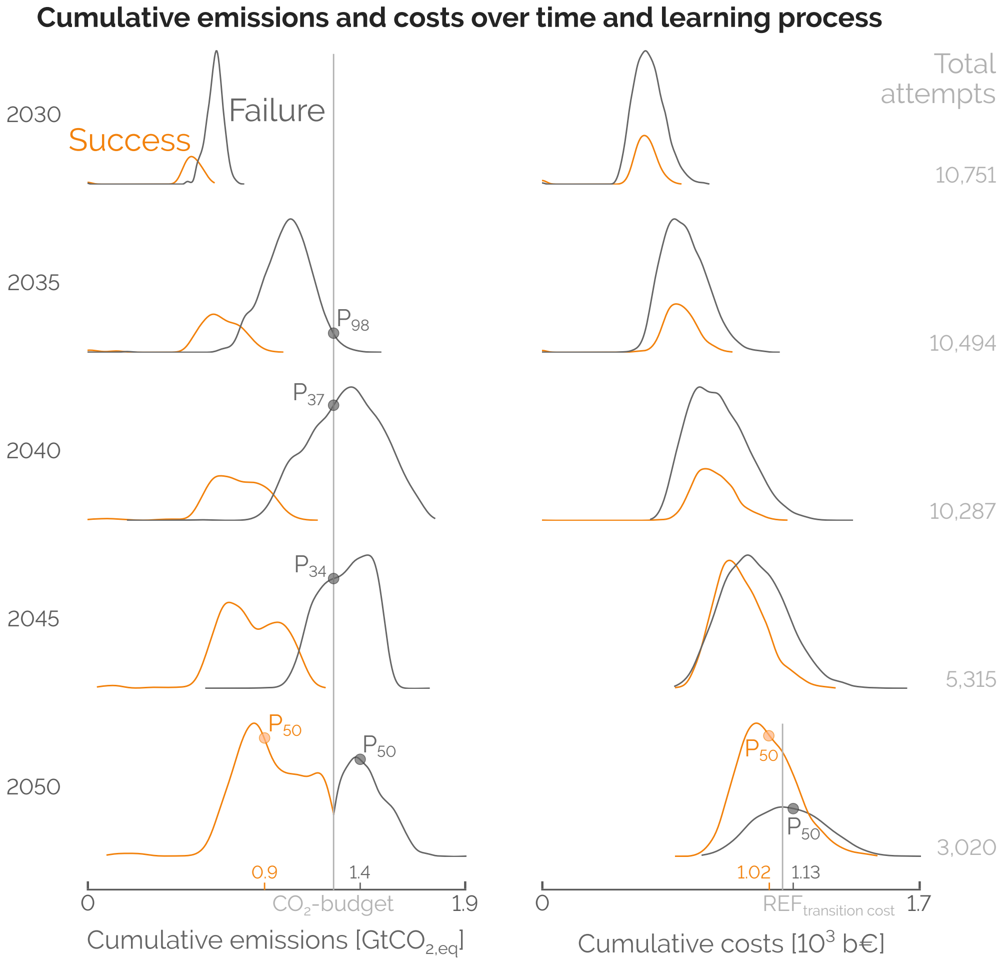

Figure 4: Exploration of the state space over the learning process: distribution of occurrence of cumulative emissions (left) and costs (right). The number of remaining attempts decreases with time since the infeasible problems and the solutions overshooting the CO_2 budget are discarded prematurely, i.e., before 2050. Besides infeasible problems, distributions labelled as “Failure” represent the attempts that overshot the CO_2 budget by 2050 at the latest. The majority of successful transitions have cumulative emissions much lower than the CO_2 budget and are cheaper than the REF case.

this threshold in 2040. This reminds the importance of near-term, 2025-2030, actions to hope for a successful transition.

Given the reward function (see Figure 8), the agent optimises its policy by aiming at lowering the total transition cost as soon as it meets the CO_2 budget. The skewness of the cumulative emissions and costs in 2050 are indications of this reward function (see Table 1). When succeeding the transitions, the cumulative emissions have a negative skewness: the agent successfully stayed within the budget and most of the cases were close to that budget (median at 0.9 $\mathrm{Gt}_{\mathrm{CO_2,eq}}$). On the contrary, the cumulative cost of successful transitions has a positive skewness: the agent successfully reduces the cost of the system as a secondary objective with 30% of the cases above the reference transition cost. The hierarchy of the agent's objectives is verified with the failures. When it failed the transition, the agent aimed at reducing the emissions (skewness of 0.61) before minimising the total transition cost (skewness of 0.24).

Table 1: Skewness of cumulative emissions and costs in 2050. Cumulative emissions are skewed to the left and to the right for the successes and failures, respectively. The skewness of the cumulative costs for successful transitions is higher compared to failures. On top of being the results of the optimisation through EnergyScope, these are influenced by the agent's policy that aims only at lowering the total transition cost as soon as it meets the CO_2 budget.

Status of episode in 2050	**Skewness of cumulative emissions**	**Skewness of cumulative costs**
Success	-0.52	0.50
Failure	0.61	0.24

Finally, we observe that the majority of the successful transitions are cheaper than the reference transition cost, 1.1 b€. Among the parameters impacting the most the total transition cost, we observe that success occurs when, on average, the cost of purchasing fossil fuels is increased more than the one of electrofuels (see Table 2). In other words, to have higher chances to succeed a myopic transition, the key factor is to reduce the uncertainty on the cost of purchasing electrofuels or to increase the cost of fossil fuels. Given the skewness that is positive and negative for the electrofuels and the fossil fuels, respectively, these cases represent more than the majority of the successful cases. On top of this, total transition costs of successful episodes are lower due to lower industrial End-Use Demand (EUD) and discount rates. These favourable conditions combined with the right agent's actions led to transitions respecting the CO_2 budget.

Table 2: Uncertain parameters impacting the most the total transition cost and, for the successful transitions, the mean of their values between 0 and 100%, μ, and their skewness, γ. On top of being supported by the agent's actions, successful transitions occur when the cost of purchasing fossil fuels is more increased than the one of electrofuels.

Parameter	μ	γ
Purchase electrofuels	50.4%	0.004
Industry EUD	49.8%	0.026
Discount rate	48.4%	0.089
Purchase fossil fuels	55.0%	-0.068

Besides the cumulative emissions and costs, the agent also observes the share of renewable energy carriers in the primary mix and the efficiency of the system. The share of renewable energy carriers in the primary mix allows identifying intermediate milestones along successful transitions (see Figure 5). From the initial state of 10% in 2020, a boost of integration of renewables in the near term is needed to hope for a successful transition. For the successful occurrences to exceed failures, this share increases to 54% in 2025. Along the transitions, this increase goes with the import of electrofuels and the full deployment of local Variable Renewable Energy Sources (VRES). In 2050, the threshold where successes occur more often than failures was at 82% renewable share. In the REF case of Chapter **??**, this share reached 86% by 2050. However, by 2050, Figure 5 shows another "bump" at lower shares of renewables in the mix. This area corresponds to the possibility of installing SMR. As uranium is considered as a non-renewable resource[6], installing SMR allows lowering the threshold as in the SMR case of Chapter **??**. Besides these milestones to respect the CO_2 budget of the transition, one can also look at the other side of the thresholds. Below the near-term threshold of $\sim$60%, this is the "no-go zone" where succeeding the transition becomes unlikely, except if betting on the future installation of SMR.

The efficiency, as defined in Section **??**, gives less valuable information towards successful transitions. Through the transition, besides the share of success increasing over the failures, the distributions of success and failure indistinguishably spread over the whole range. Similarly to the emissions, we observe a bump at lower efficiencies by 2050 due to the installation of SMR.

Actions

After investigating the intermediate milestones to meet the CO_2 budget by 2050, this section details the actions the agent has taken during the learning process (see Figure **??**). Rows represent the beginning of the time window at which the set of actions is taken. Similarly to the state space, we observe a wide exploration of the action space. The more the agent was able to progress through transition, without exceeding the CO_2 budget, the bigger is the share of successes compared to failures. Besides this observation, no specific range of values for the different actions at the different timings seems to lead to more successes. Looking at action individually, there does not seem to be any that supports more effectively the transition. The success comes from the combination of these actions.

After filtering out failures of the learning episodes and keeping only the successful transitions, only a limited set of the actions are binding and have an actual impact on the result of the optimisation in EnergyScope Pathway (see Figure 6). Supplementary Note 1 provides further details on the binding characteristic of a constraint. This allows identifying key actions to support the myopic transition. Limiting the Global Warming Potential (GWP) in the near term is a key factor for success. However, this action has a binding effect on the environment only at the end of the transition. The range over which limiting the use of fossil gas binds the optimisation is wider. Compared to other non-renewable fuels, this is due to the longer use of this energy carrier favoured by its low GWP (the second after uranium) and its versatility (applications in the electricity, heat and mobility sectors). In line with Vogt-Schilb et al.[7], the early constraints on generation and the sharper decrease of the emissions avoid lock-in situations.

When it comes to limiting the use of LFO and coal, the conclusions are more straightforward. At the beginning of the transition, most of the 159 TWh of LFO are consumed by naphtha-crackers (46%) and decentralised boilers (45%). The remaining 10% are consumed by industrial boilers. Even though LFO represents 30% of the primary energy mix in 2020, the cost-based model removes it from the mix without requiring the action of the agent. Naphtha-crackers, decentralised and industrial boilers get substituted by Methanol-to-Olefins (MTO), decentralised Heat Pump (HP) and industrial resistors and Combined Heat and Power (CHP), respectively. This "non-bindness" of limiting LFO is an indication that this action could be removed from the agent's levers of action without impacting the optimisation of its policy.

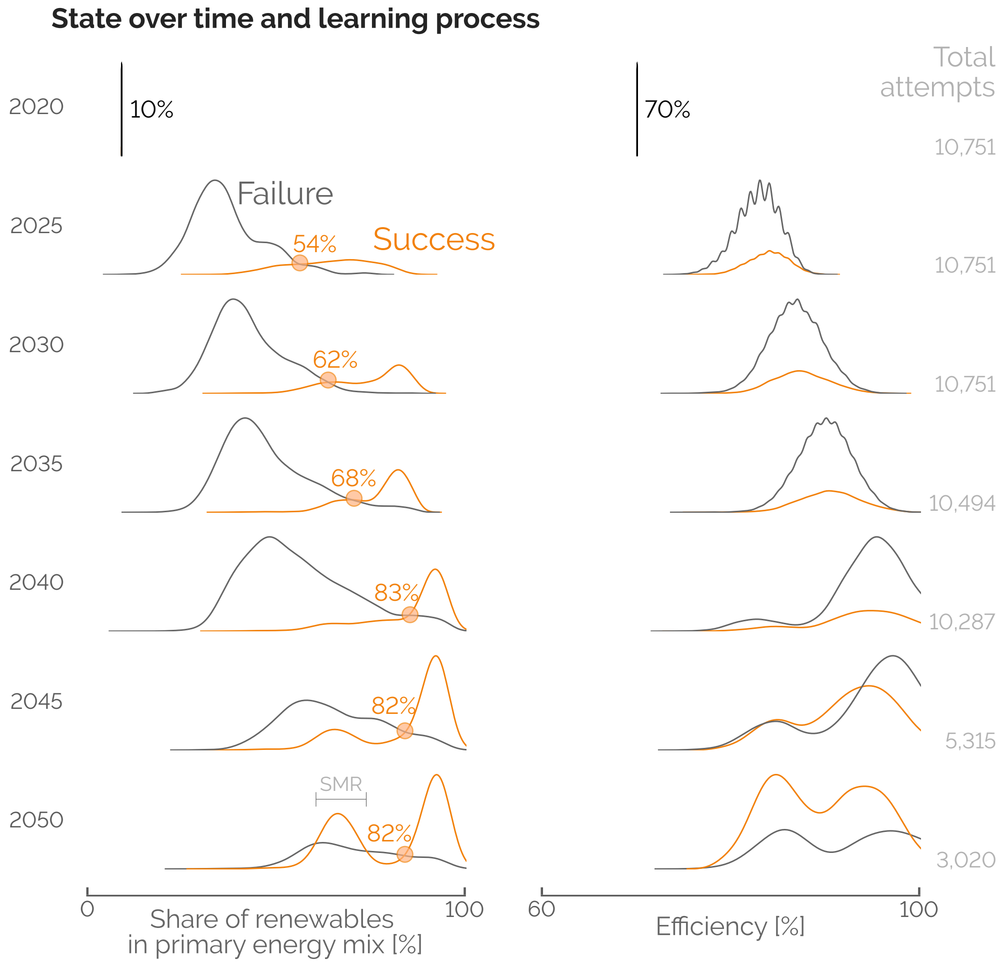

Figure 5: Exploration of the state space over the learning process: distribution of occurrence of share of renewable energy carriers in the primary energy mix (left) and efficiency (right). The number of remaining attempts decreases with time since infeasible problems and solutions overshooting the CO_2 budget are discarded prematurely, i.e., before 2050. Besides infeasible problems, distributions labelled as “Failure” represent the attempts that overshot the CO_2 budget by 2050 at the latest. Integration of local VRES at early stages then massive import of electrofuels later are needed to secure successful transitions. Below a near-term threshold (∼60%), the chances of success are limited, i.e., no-go zones. Efficiency is less valuable information for the agent to succeed transitions as failures and successes indistinguishably spread over the whole range.

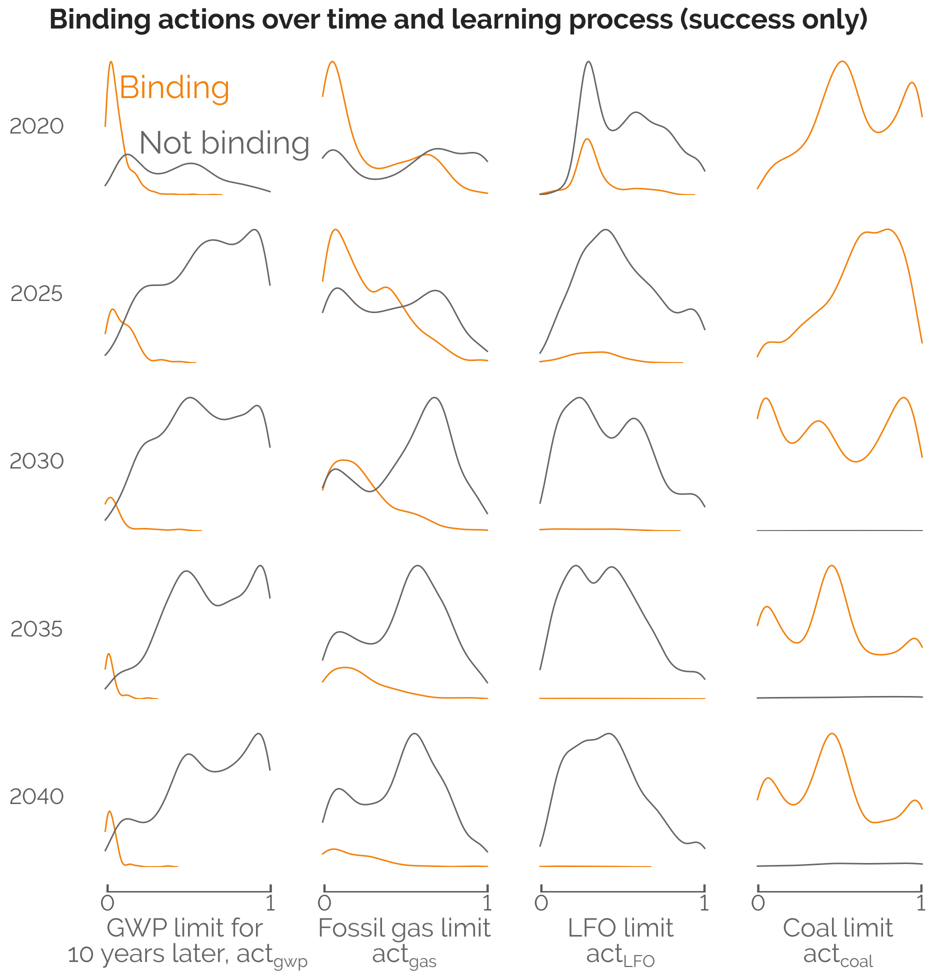

Figure 6: Keeping only the successful transitions, distribution of occurrence of binding and not binding actions. Depending on the action and its timing, it is actually constraining the optimisation through EnergyScope Pathway or not. Sweet spots can be identified when considering the limits of GWP and fossil gas consumption. Limiting coal consumption is always constraining, unlike LFO which is “naturally” substituted by EnergyScope Pathway in the near term.

On the contrary, limiting coal is always binding. Before all, this is due because coal is a cheap resource (17 €/MWh). In other words, the cost-driven environment will favour it. Then, as the maximum amount of coal (28 TWh) is much smaller than fossil gas and LFO, high values of $\mathrm{act}_{\text{coal}}$ still represent small consumptions of coal. Whatever the stage in the transition, a policy limiting the use of coal will always be effective. However, to maximise the chance in succeeding the transition, the sooner the better.

DISCUSSION

Reinforcement Learning (RL) was found to be an appropriate approach to explore myopic transition pathways under uncertainties and subject to CO_2 budget over this transition. It also allows assessing the robustness of policies to support such pathways.

Applying this method on a multi-cells model where regions can exchange energy carriers and materials, like Europe **??**, could provide insights into the varying effectiveness of energy policies applied to different energy systems. This extension could be done either by keeping a single agent ruling for the whole multi-cells system, similarly to the European Commission delivering directives to the Member States under the constraint of a collective CO_2 budget. Or, that could be a multi-agent approach where each agent would have to interact with its own energy system and with the other agents, aiming at respecting their own CO_2 budget. Even though a multi-agent environment is considerably more exigent[3], this method would allow to highlight different individual national policies in the context of a collective objective to succeed the transition at a continental scale.

As a novice user of a RL framework considering continuous action and state spaces, we would recommend opting for Soft Actor-Critic (SAC) as it is sample efficient, ensures a wide exploration and has a low sensitivity to hyper-parameters[8]. Using the SAC package developed by STABLE-BASELINES3 allows a handy introduction to apply RL.

Most of the work in applying RL is the definition of the interactions between the agent and its environment (i.e., actions, reward and states) that are very dependent on the case study and the research questions to answer. The elements presented in this work result from several trials and errors to end up with meaningful results according to our research questions.

Besides mimicking potential actual policies, the actions chosen in this work have a direct translation into constraints and we can therefore assess their effectiveness through the fact they are binding or not. In this work, we have investigated other actions like incentivising solar PV panels and wind turbines by "artificially" reducing their Capital Expenditure (CAPEX). The result is not conclusive as these technologies must take part in the Belgian energy transition and because it was harder to assess the actual impact of this action.

The reward function was designed to first aim at respecting the CO_2 budget and then minimising the total transition cost. The -300 penalty given in case of an infeasible optimisation problem was arbitrarily set. *A posteriori*, it seems to be a well-defined penalty given its significant relative difference with the values taken by the reward otherwise, between -120 and 44 (see Figure 3). Besides this penalty and given the observed results, we recommend starting with the same reward function if the objective is similar, first target cumulative emissions then cumulative costs. This requires defining the CO_2 budget according to a certain sharing principle (see Section **??**) and computing the reference total transition cost. However, other research focusing on reaching carbon neutrality by 2050 could define a binary reward function as +1 for reaching the objective and -1 otherwise.

Finally, states aim at representing the information relevant to the agent to efficiently learn and progress through the transitions. For this reason, on top of reward-related features (cumulative emissions and costs), we added other indicators that are actually monitored to help decision-makers assess their policies to reach their targets (share of renewables in the mix and the overall efficiency of the system). For other studies, one might consider other information like the metrics considered by Pickering et al.[9] (e.g., heat electrification, average national import or level of curtailment).

In conclusion, future studies might start from the actions-reward-states defined in this work and adapt these rules depending on the research questions to answer, the case study and the energy system optimisation model.

The discussion should explain the significance of the results and place them into a broader context. Subheadings are permitted.

Limitations

A ”limitations” or ”limitations of the study” subsection in the discussion is encouraged and may be required for some journals and some article formats.

Recommendations for future researchers

EXPERIMENTAL PROCEDURES

Problem formulation and rules of the game

Before starting an episode, a sample of uncertain parameters is drawn and affects the environment, EnergyScope Pathway, according to the methodology detailed in Section **??**. At the initial state, i.e., the energy system in 2020, the agent gets an initial observation, $\boldsymbol{o}_0$. An observation represents a set of the characteristics of the environment accessible to the agent for it to take the next action. The state, though, is the exhaustive list of these characteristics. Even though an observation is a subset of the state, this work uses these two words interchangeably. From this state, the agent takes a step: action, reward, and new state. The action, $\boldsymbol{a}_0$, impacts the environment, i.e., the energy system limited transition over the first decision window (2020-2030). Through this interaction with its environment, the agent is given a reward, $r_1 = r\left(\boldsymbol{a}_0|\boldsymbol{o}_0\right)$, and ends up in a new state, i.e., the energy system in 2025, characterised by a new observation, $\boldsymbol{o}_1$, and so on (see Figure 1).

A learning episode is a succession of such learning steps. In the context of the transition pathway between 2020 and 2050, an episode can come to an end for different reasons. First, if the actions taken by the agent make the optimisation infeasible, the episode is prematurely stopped before reaching 2050. Similarly, cumulative emissions of the system over the predefined CO_2 budget (see Section **??**) lead to an anticipated end of the episode. Finally, the "natural" end is the prescribed end of the transition, i.e., 2050. Consequently, the maximum value of steps for an episode is equal to $N = 5$.

The environment with which the RL-agent interacts is the optimisation of the transition pathway of a whole-energy system on a specific time window, e.g., 2020-2030 then 2025-2035 and so on, until 2040-2050 (see Figure 1). In a nutshell, starting from the initial state of the environment (i.e., the whole-energy system in 2020), the agent takes a set of actions that influence the environment, i.e., that affects parameters of the Linear Programming in EnergyScope Pathway. Then, the window 2020-2030 is optimised via EnergyScope. Some of the outputs of this optimisation feed the agent with either the new state of the system or the reward, i.e., telling the agent how good the actions were at the state the agent took it. Based on the new state and the reward, the agent takes another set of actions and the window 2025-2035 is optimised. This goes on until 2050.

Actions

Defining the levers of action, the core of the policy, to support the transition of a country-size whole-energy system is challenging, especially when accounting for political and socio-technical aspects[10]. In our work, focusing only on the techno-economic aspect, we assume that the actions taken by the agent are directly implemented and impact the environment. In other words, considering only the techno-economic lens, there is no moderation nor contest towards the agent's actions, as the objective is to assess how far and when within the transition to push the different levers of action. Given the overall objective of the agent to succeed the transition, i.e., respecting the CO_2 budget by 2050, we have defined the actions in this sense. The first action, $\mathrm{act}_{\mathrm{gwp}} \in [0,1]$, aims at limiting the emissions at the representative year ending the concerned time window, $\textbf{GWP}_{\textbf{tot}}(y_{\text{end of the window}})$, between the level of emissions in 2020, i.e., $\textbf{GWP}_{\textbf{tot}}(2020) = 123\,\text{Mt}_{\mathrm{CO_2,eq}}$, and carbon neutrality:

$$\textbf{GWP}_{\textbf{tot}}(y_{\text{end of the window}}) \leq \mathrm{act}_{\mathrm{gwp}} \cdot \textbf{GWP}_{\textbf{tot}}(2020). \qquad (1)$$

This action is equivalent to setting a national CO_2 quota.

Three additional actions support the strict limitation of yearly emissions: limiting the consumption of oil, fossil gas and coal. Out of the total GHG emissions in Belgium in 2020, oil (i.e., so-called "LFO" in the model) and fossil gas account for roughly 40% and 31%, respectively. In 2020, solid fossil fuels (i.e., so-called "coal" in the model) is much less consumed than oil and gas: i.e., 28 TWh of solid fossil fuels versus 159 and 142 TWh for oil and fossil gas, respectively. Even though its cost (17€/MWh) makes coal cost-competitive, it is a highly-emitting resource, 0.40 kt$_{\mathrm{CO_2,eq}}$/GWh. For these reasons, three independent actions limit the consumption of these three fossil resources up to the level of consumption in 2020, $\mathbf{Cons}_{\textbf{fossil gas}}(2020)$, $\mathbf{Cons}_{\textbf{LFO}}(2020)$ and $\mathbf{Cons}_{\textbf{coal}}(2020)$, over the entire concerned time window, except the first one as this year is the initial condition of the time window and cannot be optimised any more:

$$\mathbf{Cons}_{\textbf{fossil gas}}(y) \leq \mathrm{act}_{\text{fossil gas}} \cdot \mathbf{Cons}_{\textbf{fossil gas}}(2020) \qquad \forall y \in \text{time window} \quad (2)$$

$$\mathbf{Cons}_{\textbf{LFO}}(y) \leq \mathrm{act}_{\text{LFO}} \cdot \mathbf{Cons}_{\textbf{LFO}}(2020) \qquad \forall y \in \text{time window} \quad (3)$$

$$\mathbf{Cons}_{\textbf{coal}}(y) \leq \mathrm{act}_{\text{coal}} \cdot \mathbf{Cons}_{\textbf{coal}}(2020) \qquad \forall y \in \text{time window} \quad (4)$$

where $\mathrm{act}_{\text{fossil gas}}$, $\mathrm{act}_{\text{LFO}}$ and $\mathrm{act}_{\text{coal}}$ can take values between 0 and 1. These complete the action space of the agent, $A \in \mathbb{R}^4_{[0,1]}$ (see Figure 7).

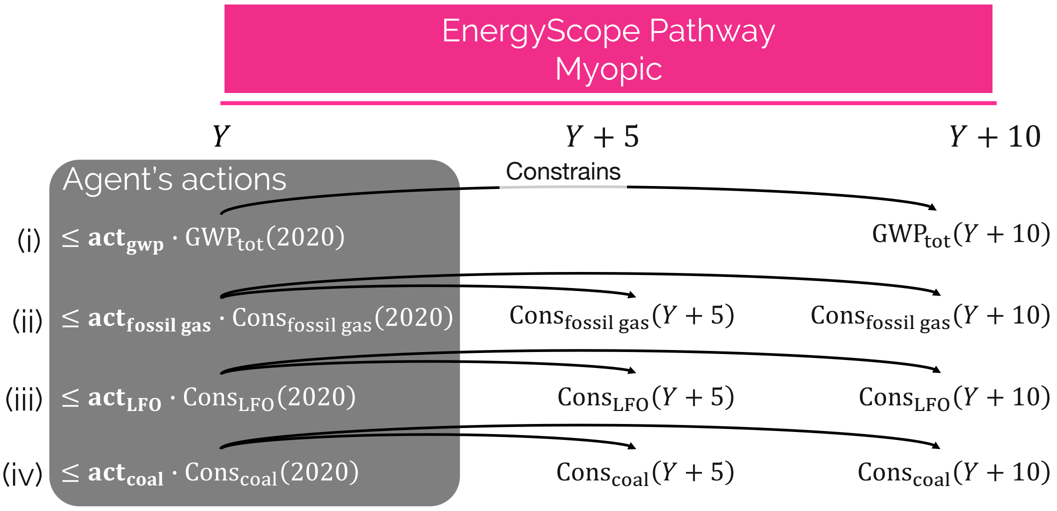
EnergyScope Pathway
Myopic
Y
$Y+5$
$Y+10$
Agent's actions
Constrains
(i) $\leq \mathbf{act_{gwp}} \cdot \mathrm{GWP_{tot}}(2020)$
$\mathrm{GWP_{tot}}(Y+10)$
(ii) $\leq \mathbf{act_{fossil\ gas}} \cdot \mathrm{Cons_{fossil\ gas}}(2020)$
$\mathrm{Cons_{fossil\ gas}}(Y+5)$
$\mathrm{Cons_{fossil\ gas}}(Y+10)$
(iii) $\leq \mathbf{act_{LFO}} \cdot \mathrm{Cons_{LFO}}(2020)$
$\mathrm{Cons_{LFO}}(Y+5)$
$\mathrm{Cons_{LFO}}(Y+10)$
(iv) $\leq \mathbf{act_{coal}} \cdot \mathrm{Cons_{coal}}(2020)$
$\mathrm{Cons_{coal}}(Y+5)$
$\mathrm{Cons_{coal}}(Y+10)$

Figure 7: Actions available to the decision-maker. Taken at the beginning of the time window to optimise (year Y), the four actions impact (i) the emissions of the system at the end of the time window (year $Y+10$) and, (ii-iv) the consumption of fossil gas, LFO and coal at years $Y+5$ and $Y+10$. Unlike the first action that sets a target for the end of the time window, the last three aim at limiting the consumption of these fossil resources over the whole time window.

Reward

When the reward is not properly defined, the agent may optimise its policy for an unintended objective, leading to undesired or suboptimal behaviour, i.e., the so-called misalignment of the learning objective[11]. Even worse, it can lead to reward hacking (or reward tampering) where the agent exploits loopholes in the reward function to achieve higher rewards without actually performing the desired task[12]. On the contrary, a proper definition of the reward function increases the sample efficiency, i.e., requiring fewer episodes to converge to the optimal policy. It also makes the policy more stable and able to withstand variations and uncertainties in the environment[13].

Through its maximisation of the expected return (see Section **??**), a RL-agent is as sensitive to positive reward, i.e., the carrot, as negative reward, i.e., the stick. When the former encourages desired behaviours, the latter can be seen as a penalty or a punishment and discourages undesirable behaviours[14]. In our case, we have decided to combine these two approaches (see Figure 8).

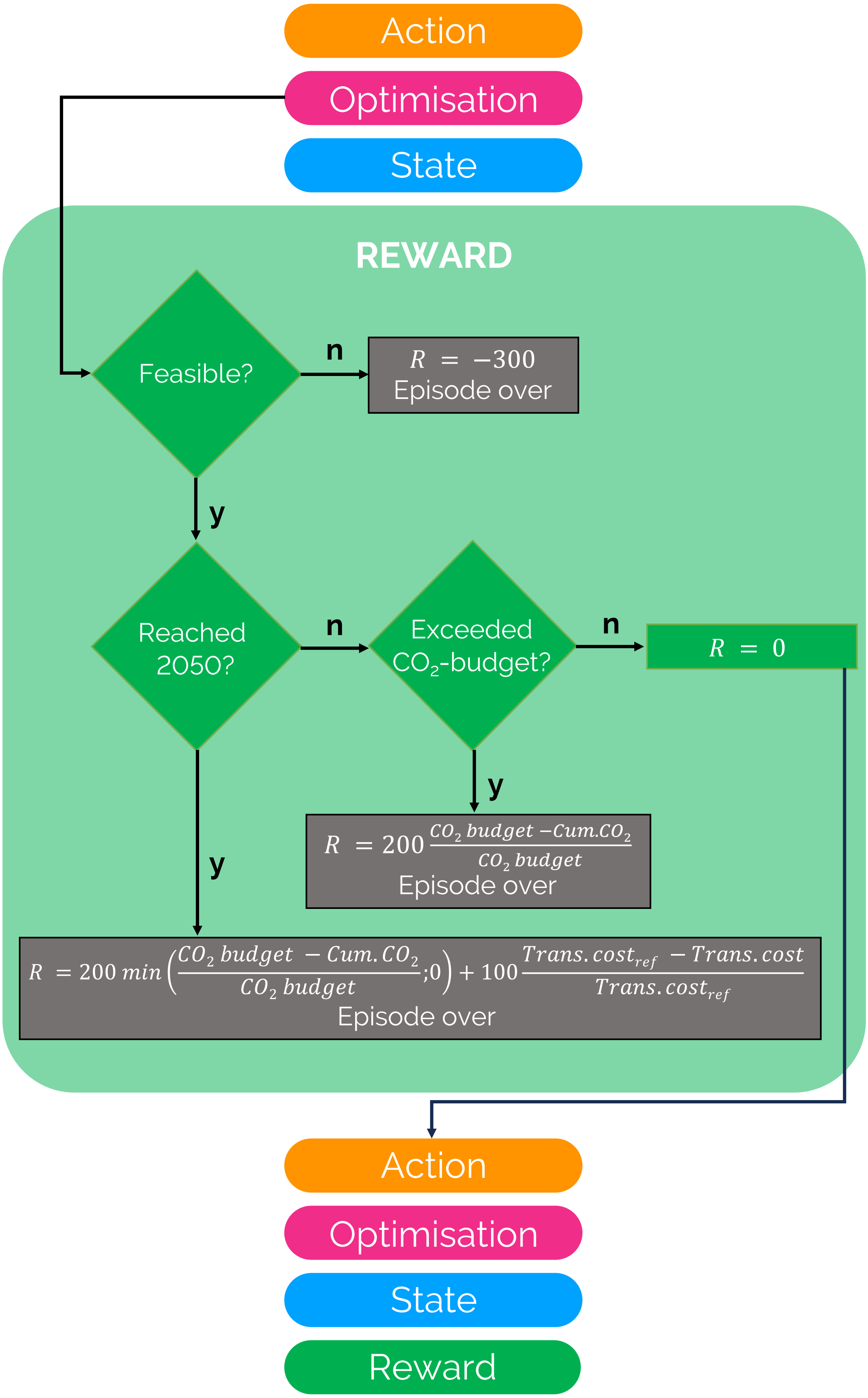
Action
Optimisation
State
REWARD
Feasible?
n
$R = -300$
Episode over
y
Reached 2050?
n
Exceeded CO_2-budget?
n
$R = 0$
y
$R = 200 \frac{CO_2\,budget - Cum.CO_2}{CO_2\,budget}$
Episode over
y
$R = 200\,min\left(\frac{CO_2\,budget - Cum.CO_2}{CO_2\,budget};0\right) + 100\frac{Trans.cost_{ref} - Trans.cost}{Trans.cost_{ref}}$
Episode over
Action
Optimisation
State
Reward

Figure 8: Reward function, R. Before 2050, the episode is prematurely ended and a negative reward is given if the optimisation is infeasible or if the CO_2 budget is exceeded. If the optimisation provides a solution and the CO_2 budget is not exceeded, the episode continues. Finally, if the episode goes until 2050, the reward is a weighted sum between the capped cumulative emissions and the total transition cost, and the episode terminates. After terminating an episode, the process starts over at the initial state, i.e., 2020.

The reward function is defined in three steps. First of all, taking a set of actions at a certain state might lead to an infeasible optimisation problem. In other words, as actions have a direct impact on some constraints of the problem, they might limit too much the feasible domain to the point where no solution can be found. For instance, the extreme case of aiming at carbon neutrality, i.e., $\mathrm{act}_{\mathrm{gwp}} = 0$, and forbidding the use of the three aforementioned fossil fuels, i.e., $\mathrm{act}_{\text{fossil gas}} = \mathrm{act}_{\text{LFO}} = \mathrm{act}_{\text{coal}} = 0$, from the beginning of the transition makes the optimisation impossible to solve. In this case, the episode is prematurely ended and the reward is "highly" negative, -300. If the optimisation is feasible and the end of the transition, i.e., 2050, is not reached, the cumulative emissions so far are evaluated. On the one hand, if these cumulative emissions exceed the CO_2 budget, $1.2\,\text{Gt}_{\mathrm{CO_2,eq}}$ (see Section **??**), the episode is also ended and a penalisation is given to the agent. This penalisation is proportional to the difference between the CO_2 budget and the actual cumulative emissions. On the other hand, the episode continues with a zero reward if the CO_2 budget is not exceeded. Eventually, when reaching 2050, given the main objective of the agent to respect the CO_2 budget and not to be more "CO_2-ambitious", we cut short the contribution of the cumulative emissions as soon as they are lower or equal to the CO_2 budget. On top of that, the reward function includes a secondary objective: the cumulative transition cost. To make the agent sensitive to the cost impact of its policy, we added the total transition cost in the reward function where the *Trans. cost$_{ref}$* on Figure 8 is equal to $1.1 \cdot 10^3$ b€. This value comes from the mean of the total transition costs obtained through the Global Sensitivity Analysis (GSA) performed on the perfect foresight transition pathway optimisation (see Section **??**). In this final form of the reward, one will notice that overshooting cumulative emissions is more penalising than an overshooting transition cost, i.e., a weight of 200 for the emissions versus 100 for the cost. The values of these weights are the results of a trial and error to fine-tune the balance between more expensive successes and cheaper failures. This way, we observed that the agent first targeted the respect of the CO_2 budget and then, to a lesser scale, avoided reaching over-costly transitions.

States

Besides the reward, the states are the other piece of information provided by the environment to the agent. In RL, the purpose of states is to represent the current situation or configuration of the environment in which the agent operates. The primary function of states in RL is to provide the necessary context for the agent to choose appropriate actions based on its current observations and goals[14]. The challenge in the definition of the states is to provide enough information but not too much to avoid overwhelming the agent with non-informative features.

Consequently, after testing several state spaces and observing the convergence of the reward, we have converged to a four-dimensional state space characterizing the energy system at the end of the optimised time window. The first dimension is directly related to the main objective of the agent: respecting the CO_2 budget until 2050. Therefore, the cumulative emissions emitted so far up to the current step of the transition is the first dimension of the states. Similarly, the cumulative cost of the transition so far constitutes the second dimension of the states to inform the agent about the cost-impact of its actions on the environment. Finally, to enrich the level of details, we have added two other dimensions representative of the key-to-the-transition indicators identified in the Renewable Energy Directive (RED) III of the European Commission[15]: the share of renewables in the primary energy mix and, the energy efficiency. The former is computed as the share of local renewables (i.e., wind, solar, hydro and biomass) and imported renewable energy carriers (i.e., biofuels and electrofuels) in the total consumption of primary energy. Electricity imported from abroad is not considered in the set of renewable energy carriers even if it can be assumed to be fully renewable by 2050. Finally, even though energy efficiency is usually defined as the ratio between the Final Energy Consumption (FEC) and the primary energy mix, we decided to define this efficiency with a focus on the EUD, like in the rest of this thesis. Where electricity, heat and non-energy EUD are expressed in terms of energy content, we needed to convert passenger and freight transports into their respective FEC to integrate them in the ratio. The information of efficiency fed back by the environment to the agent is the ratio between a “hybrid” EUD and the consumption of primary energy resources.

Custom methods subheading 2

Custom methods subheading 3

Custom methods subheading 4

RESOURCE AVAILABILITY

Lead contact

Requests for further information and resources should be directed to and will be fulfilled by the lead contact, Xavier Rixhon (xavier.rixhon@uclouvain.be).

Materials availability

Materials will be deposited to Zenodo (urls will be added).

Data and code availability

All code and data associated with this study will be available on GitHub and Zenodo (urls will be added).

ACKNOWLEDGMENTS

Authors acknowledge the support of the Energy Transition Fund of Belgium.

AUTHOR CONTRIBUTIONS

Conceptualization, X.R., H.J and F.C.; Methodology, X.R.; Investigation, X.R.; Writing-–Original Draft, X.R.; Writing-–Review & Editing, X.R., H.J and F.C.; Funding acquisition, H.J. and F.C.; Resources, H.J. and F.C.; Supervision, H.J. and F.C.

DECLARATION OF INTERESTS

The authors declare no competing interests.

SUPPLEMENTAL INFORMATION INDEX

Figures S1-S5 and their legends in a PDF

Table S1. A descriptive title for an Excel file that was too large to appear in the PDF

Table S2. Another descriptive title for a different Excel file

Data S1. Raw data on x, y, and z

Reinforcement Learning fundamentals and algorithm

RL is a subfield of machine learning focused on training an agent to make sequential decisions by interacting with an environment to achieve specific goals (see Figure 9). Unlike supervised learning, where data is labelled, and unsupervised learning, where patterns are inferred from unlabelled data, reinforcement learning deals with learning from interaction, typically through trial and error. This way, RL is considered as active learning[2]. Starting from an initial state, the agent takes an action that impacts its environment. The latter feeds back the agent with a reward and the new state. This goes on until the end of the episode. When the episode is done, the agent starts again from an initial state, takes an action and so on.

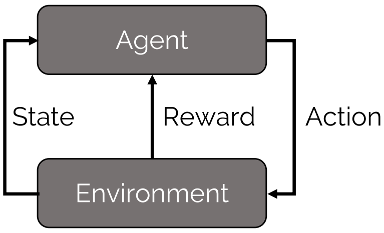

Figure 9: General concept of Reinforcement Learning (RL) as the interactions between the agent and its environment. The agent takes some action that has an impact on the environment which feeds back the agent with a reward and the new state. The objective of the agent is to optimise its policy, i.e., the mapping between the state it is at and the action to take, by maximising its cumulative reward.

The agent learns to optimise its policy by maximising the cumulative reward over time. This policy refers to the strategy or mapping from states to actions that the agent employs to make decisions. Essentially, it defines the behaviour of the agent in the environment. The ultimate goal of the agent is often to find an optimal policy, which maximises the expected cumulative reward over time. All these concepts and interactions between the agent and its environment are formalised as a Markov Decision Process (MDP)[14], represented by the tuple $< s, a, T, r, \pi, \gamma >$. The Markov property of such a decision process states that a decision is made only based on this tuple and not on the history/path that has led to it. In this tuple, $s \in S$ is the state defined in a certain state space, S, that represents the observable parts of the environment that the agent uses to make decisions; $a \in A$ is the action among the action space, A; T is the probability of transitioning from one state s to another state s' given a specific action, a: $T(s, a, s')$: $\Pr(s'|s, a)$; r is the reward received by the agent when taking the action a from state s, $r(s, a)$; π is the policy telling the action to take depending on the current state and; γ is the discount factor that controls the importance of future rewards versus immediate rewards. During the learning/optimisation process, the agent acts according to the exploitation-exploration trade-off. In the exploitation, the action a is directly given by the mapping provided by the current policy π, depending on the state s. In the exploration, the action is randomly picked within the action space. For further information, the interested reader is invited to refer to the work of Sutton and Barto[14] or the course given by David Silver[16] available online.

Before jumping to the choice of the learning algorithm, it is worth noting that we opted for the combination of RL with Deep Neural Network (DNN), called Deep Reinforcement Learning (DRL). Among others, one of the main drawbacks of traditional RL algorithms, i.e., without the use of NN, is that it suffers from the "curse of dimensionality" when facing problems with continuous action and state spaces (see Chapter **??**). By approximating the state-action function with its parameters (i.e., weights and biases), DNN can address this difficulty.

Given the assumed absence of knowledge of the agent about the dynamics of the environment, i.e., its transition or reward functions, we needed a so-called "model-free" learning algorithm. In Reinforcement Learning, the "model" stands for the dynamics "action-state-reward" between the environment and the agent. In practice, in a model-free approach, the agent estimates the optimal policy directly from experience and without estimating the dynamics of the environment. However, model-free methods suffer from two major drawbacks: their sample inefficiency and their sensitivity concerning their hyper-parameters (e.g., learning rates, exploration constants)[8]. The former leads to a too-expensive computational burden while the second requires meticulous settings to get good results. To overcome these two challenges, we needed to choose between an "on-policy" or "off-policy" algorithm. In a nutshell, in on-policy learning, the agent learns the value function or policy based on the data it generates by following its current policy whereas, in off-policy, the agent can learn from data collected by any policy, not just the one it is currently following, which provides greater flexibility and potential for reusing data stored in the so-called replay buffer. This makes off-policy algorithms more data efficient and ensures better exploration by reusing past experiences or even following random exploration[8].

To optimise the mapping between the observations and the actions, the policy $\pi\left(\boldsymbol{a}_n|\boldsymbol{o}_n\right)$, an objective function, $\boldsymbol{J}(\pi)$, is built on the cumulative rewards collected during each episode. Finally, a back-propagation process updates the weights and biases of the NN during the learning of the agent. Among the wide variety of RL algorithms applied in energy systems[3], this work opted for SAC[8] to train and update the NN. Like other actor-critic-based algorithms, SAC works with two NN in parallel: the actor learning the control policy and the critic judging the actor (see Figure 10).

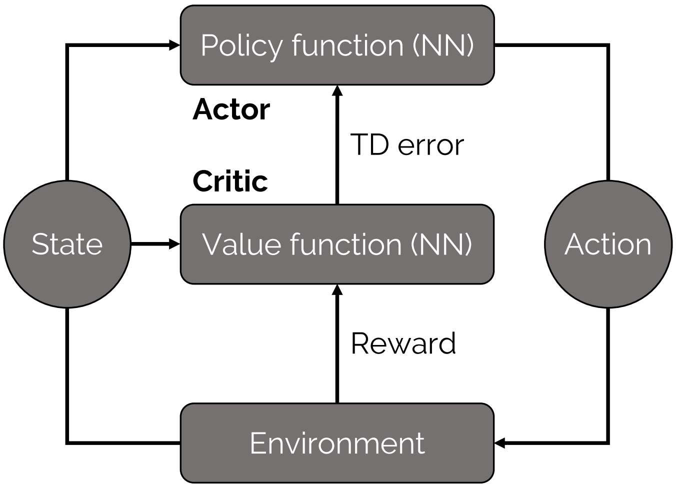

Figure 10: General concept of actor-critic-based algorithms. The two NN are trained against each other for the actor to improve the control policy and for the critic to provide a better judgement of the actor's action via the temporal-difference (TD) error. Graph adapted from[2].

SAC is a model-free and off-policy actor-critic deep RL algorithm based on the entropy-augmented objective function (see Equation (5)). The word "augmented" here is in opposition to the conventional RL objective function that is only based on the cumulative reward, i.e., the first term of Equation (5). In the RL context, entropy, also called "Shannon entropy", stands for the randomness or stochasticity of the policy.

$$\boldsymbol{J}(\pi) = \underset{\pi}{\mathbb{E}}\left[\sum_{n=0}^{N_{ep}} \gamma^n r_n\left(\boldsymbol{o}_n, \boldsymbol{a}_n\right) - \zeta \log\left(\pi\left(\boldsymbol{a}_n | \boldsymbol{o}_n\right)\right)\right], \tag{5}$$

where γ is the discount factor and ζ the temperature parameter. γ determines how much importance we want to give to future rewards within an episode. ζ balances the trade-off between the exploitation of proven actions via the return maximisation, i.e., $\sum_{n=0}^{N_{ep}} \gamma^n r_n\left(\boldsymbol{o}_n, \boldsymbol{a}_n\right)$, and the exploration through the entropy term, i.e., $\log\left(\pi\left(\boldsymbol{a}_n | \boldsymbol{o}_n\right)\right)$. This way, SAC ensures sample efficiency while improving exploration[17] and robustness[18]. In their work, Haarnoja et al.[17] showed a lower sensitivity of SAC to hyper-parameters. These make SAC a state-of-the-art algorithm and one of the most efficient model-free deep RL methods nowadays[17]. In this thesis, we used the open-source SAC package developed by STABLE-BASELINES3[19] where the policy NN is a fully connected multilayer perceptron (MLP) built with TENSORFLOW[20]. For further information on RL and the SAC algorithm, the interested reader is invited to refer to the works of Sutton and Barto[14] and Haarnoja et al.[8], respectively.

Comparison with perfect foresight under uncertainties

This section compares these results under myopic conditions supported by the RL-agent with the perfect foresight under uncertainties that is considered as a reference.

RL-based myopic optimisation provides CO_2 emissions pathways different from the perfect foresight approach to respect the same CO_2 budget (see left side of Figure 11). However, driven first by this CO_2 budget, the agent often reaches much lower cumulative emissions when succeeding the transition (see Figure 4). This comes from the agent's actions that limit the emissions and/or the consumption of fossil resources at the early stages. Thanks to the bigger emission reduction at these early stages, the RL-based optimisation can benefit from a "CO_2 buffer" at the end of the transition. This buffer is compensated by the end of the transition where 50% of the myopic transitions reach 2050 with 10 or more remaining $\text{Mt}_{\text{CO}_2\text{,eq}}$ compared to 4 for the perfect foresight approach. These remaining emissions by 2050 come from the consumption in industrial boilers of waste and coal accounting for 3.5% and 2.4% of the primary mix on average by 2050. Finally, the long-term vision of the perfect foresight approach results in a smoother reduction of emissions to end up with less emissions by 2050.

The comparison between the failures and the successes demonstrates the added value brought by myopic pathway optimisation. In the near term (2025-2030), levels of emission are similar between perfect foresight and myopic cases that have failed. This shows that limited foresight encourages to strongly act at the early stages. On top of this, following the initial steps of CO_2 emissions pathways resulting from the PF approach would likely ($\sim$80%) lead to failure of the transition.

Looking at the total transition cost, the combination of the agent's actions and favourable economic conditions (see Section 1) make the myopic transitions cheaper, on average, than the PF cases (see right side of Figure 11). This is also because the perfect foresight approach always finds a solution even in the worst conditions such as the high cost of purchasing resources and high EUD. This explains the wider variability of the PF results too. However, with the same sample of uncertain parameters, given the assumed full knowledge over the whole time horizon, PF naturally results in a cheaper transition than its myopic equivalent.

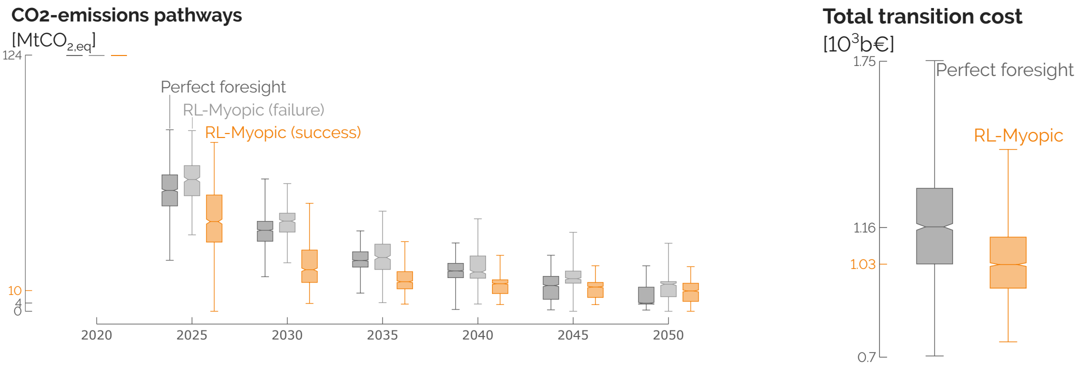

Figure 11: Comparison of CO_2 emissions pathways (left) and total transition cost (right) from the perfect foresight optimisation under uncertainties and the RL-based myopic optimisation. Myopic transitions succeed with a more drastic reduction of emissions in the short term and, on average, more favourable economic conditions.

The analysis of the cumulative costs shows that the Operational Expenditure (OPEX) is the main difference between myopic and perfect foresight transitions (see Figure 12). Supported by the agent's actions, successful myopic transitions have a lower OPEX than the perfect foresight ones.

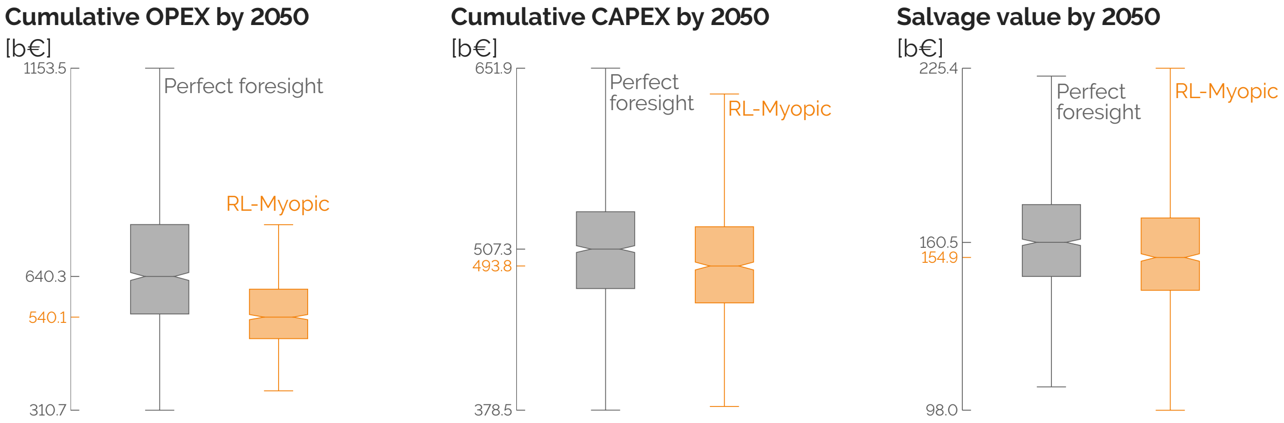

Figure 12: Comparison of cumulative OPEX (left), CAPEX (centre) and salvage value (right) in 2050 from the perfect foresight optimisation under uncertainties and the RL-based myopic optimisation.

The cost of purchasing the energy carriers represents about 70% of the total cumulative OPEX. The assessment of the primary energy mix by 2050 highlights that the difference in OPEX between the perfect foresight and the myopic pathways comes from the import of electrofuels, and especially of e-ammonia (see Figure 13). In the majority of the cases, e-ammonia is more than two times more imported in the myopic transitions. Being cheaper than e-methane (see Chapter **??**), e-ammonia brings flexibility in the production of electricity via Combined Cycle Gas Turbine (CCGT) (see Chapter **??**). Besides the slightly favourable economic conditions (see Table 2), the myopic optimisations opt to invest massively into importing renewable molecules because of the limited knowledge of the future, and, among others, the availability of SMR. This explains why 50% of the successful transitions reached cumulative emissions below 900 $Mt_{CO_2,eq}$ (see Figure 4).

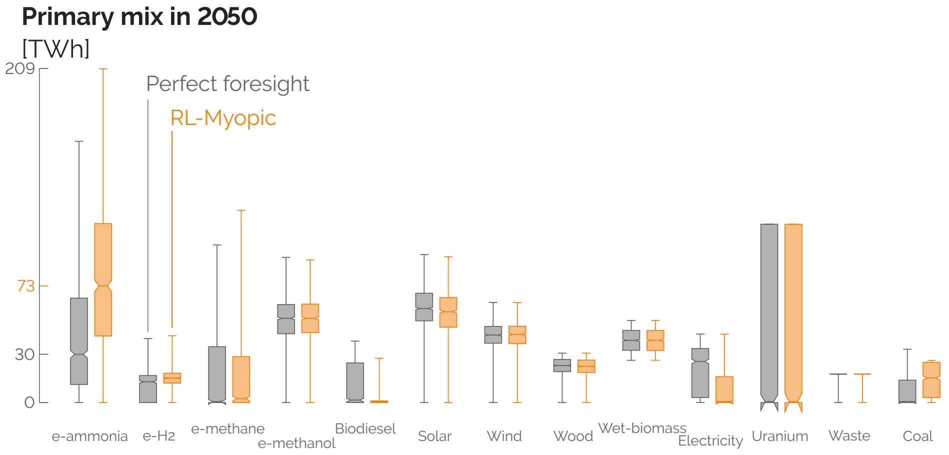

Figure 13: Comparison of the primary energy mix in 2050 from the perfect foresight optimisation under uncertainties and the RL-based myopic optimisation. The biggest difference is about e-ammonia to supply CCGT.

To bind or not to bind, that is the question

To identify the actions that have an actual impact on the environment, we can check if they are binding or not. In a Linear Programming (LP) problem, constraints represent hyperplanes in the domain of variables. In a two-dimension space, these are straight lines (see Figure 14). When the problem is bounded and feasible, these lines are the edges of a convex polygon: the domain of feasibility. The optimal solution, $\mathbf{x}^*$, is the combination of variables leading to the optimal value of the objective function. Besides being within the domain of feasibility, it is proven that this optimal solution, when unique[1], locates on a vertex of the domain[21]. The constraints intersecting at this vertex are considered binding, actually limiting the objective function to be more optimal. In other words, binding constraints, when tightened, aggravate the objective value function. If these are inequality constraints, as represented in Figure 14, it means that the left and right sides of the equations are equal.

References

1. Poncelet, K., Delarue, E., Six, D., and D'haeseleer, W. (2016). Myopic optimization models for simulation of investment decisions in the electric power sector. In 2016 13th International Conference on the European Energy Market (EEM). IEEE pp. 1–9.

2. Cao, D., Hu, W., Zhao, J., Zhang, G., Zhang, B., Liu, Z., Chen, Z., and Blaabjerg, F. (2020). Reinforcement learning and its applications in modern power and energy systems: A review. Journal of modern power systems and clean energy *8*, 1029–1042.

3. Perera, A., and Kamalaruban, P. (2021). Applications of reinforcement learning in energy systems. Renewable and Sustainable Energy Reviews *137*, 110618.

4. Thiran, P. Exploring options for a fossil-free european energy system: The role of renewable fuels. Ph.D. thesis UCL-Université Catholique de Louvain (2024).

[1] There are cases where the objective function has the same optimal value along an entire edge. In this case, there is an infinity of solutions and the problem is indeterminate.

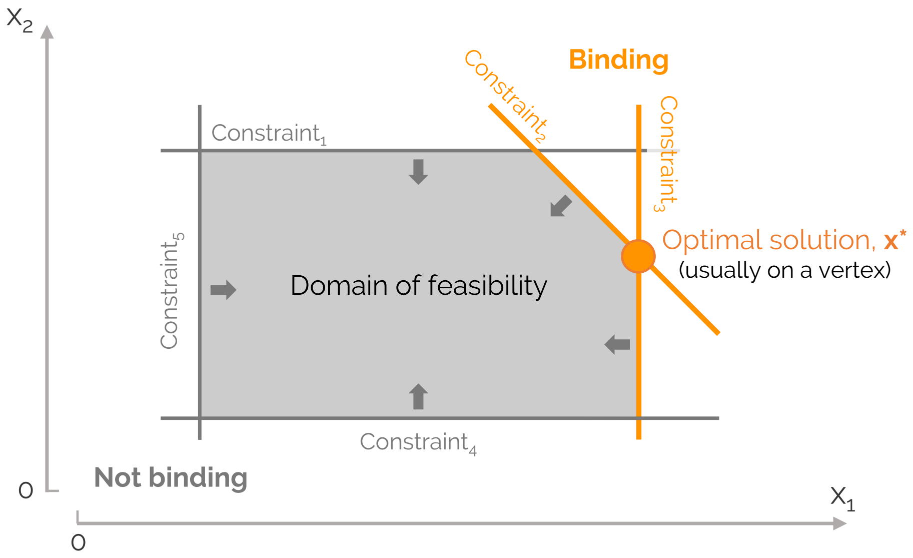

Figure 14: Binding versus non-binding constraints. In LP where the feasibility domain is non-empty and bounded, the constraints defined a convex feasibility domain in the space of variables (here, x_1 and x_2). The optimal solution usually locates on a vertex of this domain, i.e., the intersection of several constraints (here, constraints 2 and 3) limiting the solution. These constraints are considered binding, i.e., having a limiting impact on the optimal solution.

5. Sun, T., Qin, M., Su, C.W., and Zhang, W. (2024). The indispensable role of energy import: does its price really matter for german employment? Energy Strategy Reviews *55*, 101495.

6. Rixhon, X., Limpens, G., Contino, F., and Jeanmart, H. (2021). Taxonomy of the fuels in a whole-energy system. Frontiers in Energy Research - Sustainable Energy Systems and Policies. doi: `10.3389/fenrg.2021.660073`.

7. Vogt-Schilb, A., Meunier, G., and Hallegatte, S. (2018). When starting with the most expensive option makes sense: Optimal timing, cost and sectoral allocation of abatement investment. Journal of Environmental Economics and Management *88*, 210–233.

8. Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. (2018). Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In International conference on machine learning. PMLR pp. 1861–1870.

9. Pickering, B., Lombardi, F., and Pfenninger, S. (2022). Diversity of options to eliminate fossil fuels and reach carbon neutrality across the entire european energy system. Joule *6*, 1253–1276.

10. Castrejon-Campos, O., Aye, L., and Hui, F.K.P. (2020). Making policy mixes more robust: An integrative and interdisciplinary approach for clean energy transitions. Energy Research & Social Science *64*, 101425.

11. Christiano, P.F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. (2017). Deep reinforcement learning from human preferences. Advances in neural information processing systems *30*.

12. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., and Mané, D. (2016). Concrete problems in ai safety. arXiv preprint arXiv:1606.06565.

13. Henderson, P., Islam, R., Bachman, P., Pineau, J., Precup, D., and Meger, D. (2018). Deep reinforcement learning that matters. In Proceedings of the AAAI conference on artificial intelligence vol. 32.

14. Sutton, R.S., and Barto, A.G. (2018). Reinforcement learning: An introduction. MIT press.

15. European Parliament. Directive (EU) 2023/2413 of the European Parliament and of the Council of 18 October 2023 amending Directive (EU) 2018/2001, Regulation (EU) 2018/1999 and Directive 98/70/EC as regards the promotion of energy from renewable sources, and repealing Council Directive (EU) 2015/652. Tech. Rep. European Parliament (2023). Official Journal of the European Union 2413, 1-77.

16. David Silver (2016). RL Course by David Silver. https://www.youtube.com/watch?v=2pWv7GOvuf0&list=PLzuuYNsE1EZAXYR4FJ75jcJseBmo4KQ9-..

17. Haarnoja, T., Tang, H., Abbeel, P., and Levine, S. (2017). Reinforcement learning with deep energy-based policies. In International conference on machine learning. PMLR pp. 1352–1361.

18. Ziebart, B.D. (2010). Modeling purposeful adaptive behavior with the principle of maximum causal entropy. Carnegie Mellon University.

19. Raffin, A., Hill, A., Gleave, A., Kanervisto, A., Ernestus, M., and Dormann, N. (2021). Stable-baselines3: Reliable reinforcement learning implementations. The Journal of Machine Learning Research *22*, 12348–12355.

20. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M. et al. (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:1603.04467.

21. Bertsimas, D., and Tsitsiklis, J.N. (1997). Introduction to linear optimization vol. 6. Athena Scientific Belmont, MA.