UCLouvain
Institute of Mechanics,
Materials and Civil Engineering

Robust optimisation of the pathway towards a sustainable whole-energy system

A hierarchical multi-objective reinforcement-learning-based approach

Doctoral dissertation presented by
Xavier RIXHON

in partial fulfillment of the requirements for
the degree of Doctor in Engineering Sciences

September 2024

Thesis committee
Pr. Francesco CONTINO (UCLouvain, Supervisor)
Pr. Hervé JEANMART (UCLouvain, Supervisor)
Pr. Paul FISETTE (UCLouvain, President)
Pr. Christophe DE VLEESCHOUWER (UCLouvain, Secretary)
Pr. Sylvain QUOILIN (ULiège)
Dr. Stefano MORET (ETH Zurich)
Pr. Stefan PFENNINGER (TU Delft)

Contents

Symbols **iii**

1 Reinforcement Learning CO_2-policy investigation **1**

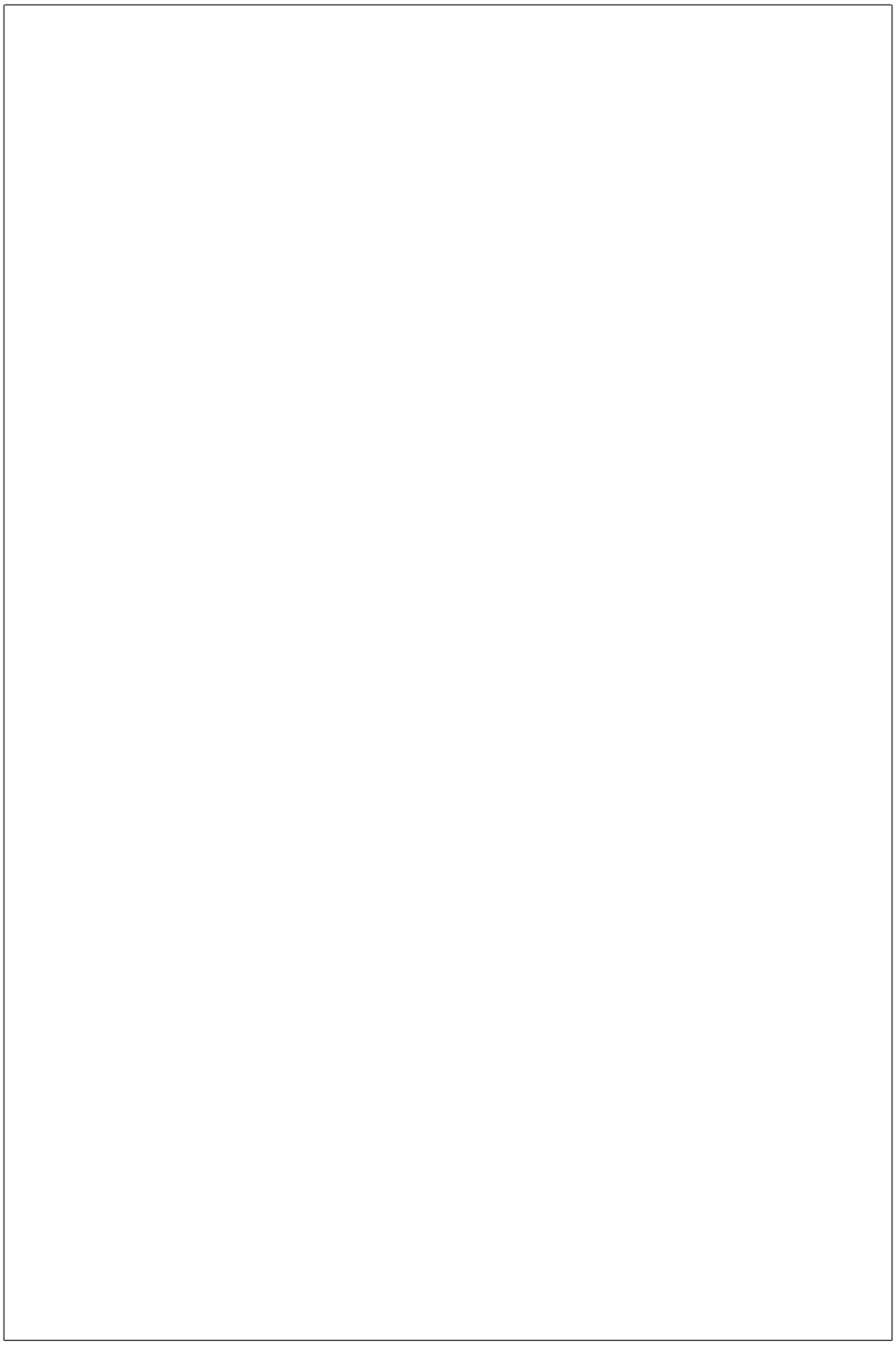
1.1 Definition of the actions, reward and states 3

1.2 Convergence and learning process 7

1.3 Testing and comparison with references 13

1.4 Discussion 13

Bibliography **17**



Symbols

Acronyms

API	Application Programming Interface
BECCS	Bioenergy with Carbon Capture and Storage
BEMS	Building Energy Management System
BEV	Battery Electric Vehicle
BTX	Benzene, Toluene and Xylene
CAPEX	Capital Expenditure
CCGT	Combined Cycle Gas Turbine
CCS	Carbon Capture and Storage
CHP	Combined Heat and Power
CNG	Compressed Natural Gas
DC	Direct Current
DHN	District Heating Network
DNN	Deep Neural Network
DRL	Deep Reinforcement Learning
ESOMs	Energy System Optimisation Models
ESTD	EnergyScope Typical Days
EUD	End-Use Demand
FC	Fuel Cell
FEC	Final Energy Consumed
GDP	Gross Domestic Product
GHG	Greenhouse Gases
GSA	Global Sensitivity Analysis
GWP	Global Warming Potential
HP	Heat Pump
HT	High-Temperature

HVC	High-Value Chemicals
IAMs	Integrated Assessment Models
ICE	Internal Combustion Engine
IEA	International Energy Agency
IPCC	Intergovernmental Panel on Climate Change
IQR	Interquatile Range
LCA	Life Cycle Assessment
LCOE	Levelised Cost of Energy
LFO	Light Fuel Oil
LOO	Leave-One-Out
LP	Linear Programming
LPG	Liquefied Petroleum Gas
LT	Low-Temperature
MDP	Markov Decision Process
MMSA	Methanol Market Services Asia
MTBE	Methyl Tert-butyl Ether
MTO	Methanol-to-olefins
NED	Non-energy Demand
NG	Fossil Gas
NN	Neural Network
NRE	Non-renewable Energy
NSC	Naphtha Steam Cracker
OPEX	Operational Expenditure
PC	Principal Component
PCs	Principal Components
PCA	Principal Component Analysis
PCE	Polynomial Chaos Expansion
PDF	Probability Density Function
PV	Photovoltaic
RE	Renewable Energy
RL	Reinforcement Learning
SAC	Soft Actor Critic
SMR	Small Modular Reactor
SVD	Singular Value Decomposition
TRL	Technology Readiness Level
UQ	Uncertainty Quantification
VRES	Variable Renewable Energy Sources

List of publications

Limpens, G., **Rixhon, X.**, Contino, F., & Jeanmart, H. (2024). "*EnergyScope Pathway: An open-source model to optimise the energy transition pathways of a regional whole-energy system.*" In Applied Energy, (Vol. 358). URL: `https://doi.org/10.1016/j.apenergy.2023.122501`

Rixhon, X., Limpens, G., Coppitters, D., Jeanmart, H.,& Contino, F.(2022). "*The role of electrofuels under uncertainties for the Belgian energy transition.*" In Energies (Vol. 14). URL: `https://doi.org/10.3390/en14134027`

Rixhon, X., Tonelli, D., Colla, M., Verleysen, K., Limpens, G., Jeanmart, H. ,& Contino, F.(2022). "*Integration of non-energy among the end-use demands of bottom-up whole-energy system models.*" In Frontiers in Energy Research, Sec. Process and Energy Systems Engineering, (Vol. 10). URL: `https://doi.org/10.3389/fenrg.2022.904777`

Rixhon, X., Colla, M., Tonelli, D., Verleysen, K., Limpens, G., Jeanmart, H., & Contino, F.(2021). "*Comprehensive integration of the non-energy demand within a whole-energy system: Towards a defossilisation of the chemical industry in Belgium.*" In proceedings of ECOS 2021 conference (Vol. 34, p. 154).

Rixhon, X., Limpens, G., Contino, F., & Jeanmart, H. (2021). "*Taxonomy of the fuels in a whole-energy system.*" In Frontiers in Energy Research, Sec. Sustainable Energy Systems, (Vol. 9). URL: `https://doi.org/10.3389/fenrg.2021.660073`

Limpens, G., Coppitters, D., **Rixhon, X.**, Contino, F., & Jeanmart, H. (2020). "*The impact of uncertainties on the Belgian energy system: application of the Polynomial Chaos Expansion to the EnergyScope model.*" In proceedings of ECOS 2020 conference (Vol. 33, p. 711).

Chapter 1

Reinforcement Learning CO_2-policy investigation

"For the things we have to learn before we can do them, we learn by doing them."
Aristotle, in *The Nicomachean Ethics*, IV$^{\text{th}}$ century BC

Uncertainties about the future along with a large variety of Integrated Assessment Models (IAMs) yield to an even larger variety of Greenhouse Gases (GHG) emissions reduction pathways [1]. For instance, several studies [2, 3] advocate for actions to take in the near future, especially to keep on track with the 1.5°C (if not, 2°C) increase of global temperature by the end of the century. On the contrary, using their top-down model DICE, Nordhaus [4] state that immediate and drastic actions are not compulsory to meet the ambition of climate change mitigation. This is even more valid when models assess a myopic transition pathway subject, with limited foresight through the future with progressively unveiled uncertainties.

To address this issue, several approaches have been used. Among them, multi-stage stochastic programming is often put forward as a promising method. Stochastic programming formulates the problem as a mathematical program with probabilistic constraints or objective functions. These models explicitly consider the uncertainty by incorporating probability distributions for the uncertain parameters. The goal is to find an optimal decision that minimizes/maximizes the expected value of the objective function while satisfying the probabilistic constraints, modelled as a scenario tree. At each stage of the problem, here the transition pathway of a whole-energy system, the model has the possibility of recourse, i.e. to adapt the decisions made at earlier stages, in the response to unveiled uncertainties [5]. Using MARKAL model [6], Kanudia and

Loulou [7] assessed a multi-stage stochastic optimisation of the 5-year steps transition of Quebec between 1995 and 2035 accounting for high/low mitigation action plan and high/low growth scenarios. The authors found that hedging strategies, adapting with the future uncertainties, were outperforming the perfect foresight and deterministic optimisation of the different scenarios. However, stochastic programming is usually applied to limited number of uncertainties, i.e. up to 10,and relies on probability distribution that are often difficult to define properly. Increasing the number of these uncertainties in stochastic programming usually leads to a computational burden that limits the use of such a method in IAMs [1]. Based on the approach of Bertsimas and Sim [8], and similarly to Moret [9], Nicolas et al. [1] rather opted for the robust optimisation of the global pathway up to 2200 given different temperature deviation targets, i.e. 2 or 3°C by 2200 via the use of uncertainty budget, Γ, in the TIAM-World model [10]. Considering 9 climate parameters and their respective lower and upper bounds, the idea behind the uncertainty budget stems from the improbability of all parameters simultaneously reaching one of their two extreme values.

In the exploration of the myopic transition pathway under uncertainties, we decided to investigate the Reinforcement Learning (RL) approach to benefit from its policy optimisation mechanism. Indeed, policymaking for transitioning a whole-energy system can be viewed as an iterative process of learning from policy implementation efforts, involving ongoing analysis of energy policy challenges and experimenting with various solutions [11]. RL exhibits two main advantages: its effectiveness to handle uncertainties and the model-free approach where an accurate representation of the real world is not needed to optimize the policy [12].Besides the environment, i.e. the myopic transition pathway of the whole-energy system via EnergyScope Pathway (see Chapter **??**), the first part of this chapter presents the three key features of interaction between the agent, optimizing its policy, and the environment: actions, states and reward. Then, the results of this policy optimisation point out strategies to follow, i.e. *sweet spots*, in the transitions under uncertainties as well as *no-go zones* where the chances of succeeding the transition, i.e. respecting the CO_2-budget, are very limited. Finally, these results are compared with references, i.e. the perfect foresight and the myopic optimisation of the transition under the same uncertainties but without the trained RL-agent that can support this transition thanks to its learned policy.

Contributions

Applying the RL approach to the optimization of the myopic transition pathway of a whole-energy system presents several novelties. First of all, as introduced in Sec-

tion **??**, when applied to energy systems, RL is more dedicated either to smaller scale systems (e.g. Building Energy Management System (BEMS), vehicles and energy devices) or to sector-specific, often the power sector, problems (e.g. dispatch problems, energy markets and grid) [12]. In our case, the sector-coupling, the long-term goal at the end of a multiple-steps transition and the number of uncertain parameters make this application new for RL.

Usually, applications of RL focus more on the result of the learning process, the optimised policy, for optimising the control of system [12]. On top of it, this thesis investigate the learning episodes themselves to explore the field of possibilities to succeed the transition.

Then, applying to this optimisation environment, i.e. EnergyScope myopic Pathway, rather than a simulation environment, allows building a hierarchical multi-objective optimisation framework. In this agent, while the objective of the environment remains the minimisation of the total "transition" cost (on the concerned limited time window), the agent optimises its strategy to respect the CO_2-budget.

Finally, comparing the RL-based results with more conventional approaches, i.e. perfect foresight, if not myopic, optimisation without learning process, highlights the added-value brought by the optimised policy.

1.1 Definition of the actions, reward and states

As already introduced in Section **??**, the environment with which the RL-agent interacts is the optimisation of the transition pathway whole-energy system on a specific time window, e.g. 2020-2030 then 2025-2035 and so on, until 2040-2050 (see Figure **??**). 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. 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 he took it. Based on these two pieces of information, i.e. the new state and the reward, the agent takes another set of actions and the window 2025-2035 is optimised. This goes on until eventually reaching 2050. The main purpose of this section is to define the shape of the reward as well as the sets of actions and states.

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 po-

litical and socio-technical aspects [13]. In our work, focusing only on the techno-economic aspect, we assume that the actions taken by the agent are directly implemented and impacting the environment, without "misfire". 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, $\text{act}_{\text{gwp}} \in [0, 1]$, aims at limiting the emissions at the representative year ending the concerned time window, $\mathbf{GWP_{tot}}(y_{\text{end of the window}})$, between the level of emissions in 2020, i.e. $\mathbf{GWP_{tot}}(2020) = 123\,\text{Mt}_{\text{CO}_2,\text{eq}}$, and carbon-neutrality:

$$\mathbf{GWP_{tot}}(y_{\text{end of the window}}) \leq \text{act}_{\text{gwp}} \cdot \mathbf{GWP_{tot}}(2020) \tag{1.1}$$

Out the total GHG emissions in Belgium in 2020,oil (i.e. so-called Light Fuel Oil (LFO) in the model), on the one hand and, on the other hand, fossil gas, account for roughly 40% and 31%, respectively. Then, even though its use in 2020 is much more limited compared to the two formers, i.e. 28 TWh of solid fossil fuels (i.e. so-called COAL in the model) versus 159 and 142 TWh for oil and fossil gas, respectively, coal is a cheap, 17€/MWh, and highly-emitting resource, $0.40\,\text{kt}_{\text{CO}_2,\text{eq}}$/GWh. For these reasons, three additional actions support the strict limitation of overall emissions of the first action: limiting the consumption of these three fossil resources up to the level of consumption in 2020, $\mathbf{Cons_{fossil\ gas}}(2020)$, $\mathbf{Cons_{LFO}}(2020)$ and $\mathbf{Cons_{coal}}(2020)$, 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:

$$\mathbf{Cons_{fossil\ gas}}(y) \leq \text{act}_{\text{fossil gas}} \cdot \mathbf{Cons_{fossil\ gas}}(2020) \qquad \forall y \in \text{time window} \tag{1.2}$$

$$\mathbf{Cons_{LFO}}(y) \leq \text{act}_{\text{LFO}} \cdot \mathbf{Cons_{LFO}}(2020) \qquad \forall y \in \text{time window} \tag{1.3}$$

$$\mathbf{Cons_{coal}}(y) \leq \text{act}_{\text{coal}} \cdot \mathbf{Cons_{coal}}(2020) \qquad \forall y \in \text{time window} \tag{1.4}$$

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

Reward

Properly defined the reward fed by the environment to the agent is crucial in RL for several reasons.If the reward is not properly defined, the agent may optimize its policy for an unintended objective, leading to undesired or subotpimal behavior, i.e. the so-called misalignment of the learning objective [14]. Even worse, it can lead to reward hacking (or reward tampering) where the agent exploits loopholes in the reward func-

tion to achieve higher rewards without actually performing the desired task [15]. On the contrary, a proper definition of the reward function increases the sample efficiency, i.e. requiring less episode to converge to the optimal policy. It also makes the policy more stable and able to withstand variations and uncertainties in the environment [16].

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 the undesirable behaviours [17]. In our case, we have decided to combine these two approaches (see Figure 1.1).

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. $\text{act}_{\text{gwp}} = 0$, and forbidding the use of the three aforementioned fossil fuels, i.e. $\text{act}_{\text{fossil gas}} = \text{act}_{\text{LFO}} = \text{act}_{\text{coal}} = 0$, 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 EnergyScope is able to provide a solution to the time window to optimise and the end of the transition, i.e. 2050, is not reached, a test on the cumulative emissions so far. On the one hand, if these cumulative emissions exceed the CO_2-budget, $1.2\,\text{Gt}_{\text{CO}_2,\text{eq}}$ (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, if reaching 2050, we decided to tweak the reward function in capping the share of the cumulative emissions and integrating the transition cost. Given the main objective of the agent to respect the CO_2-budget and not to be more ambitious "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. Moreover, 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 1.1 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 are more penalising than an overshooting transition cost, i.e. weight of 200 for the emissions versus 100 for the cost. The values of these weights are the results of a trial and error. This way, we observed that the agent first targeted the respect of

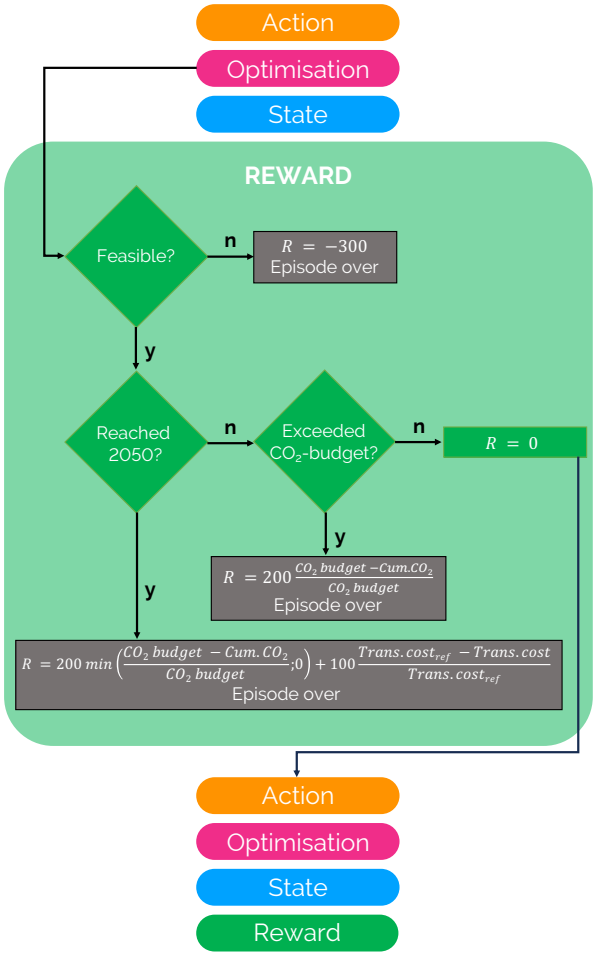


Figure 1.1. Reward function, R. Before reaching 2050, the episode is prematurely ended and a negative reward is given if the optimisation is infeasible or if the CO_2-budget is exceeded. If the optimisation provides a solution and the CO_2-budget is not exceeded, the episode continues. Finally, if the episode goes until 2050, the reward is a weighted sum between the capped cumulative emissions and the total transition cost.

the CO_2-budget and then, to a lesser scale, avoided reaching over-costly transitions.

States

Besides the reward, states are the other piece of information provided by the environment to the agent. In RL, the purpose of states are 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 [17]. 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 another process of trial and error, we have converged to a four-dimension 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 in 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 by in Renewable Energy Directive (RED) III of the European Commission [18]: 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 Consumed (FEC) and the primary energy mix, we decided to define this efficiency with a focus on the End-Use Demand (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.

1.2 Convergence and learning process

Before testing the optimal policy $\pi^* \left(a_n | o_n\right)$, the first step consists in assessing the learning of the Neural Network (NN), also called "training". For this, numerous episodes are played through the myopic optimisation of the transition pathway of Belgium. At the beginning of each episode, the agent starts with the actual Belgian energy

system of 2020 (see Appendix **??**). Then, a sample of values, drawn for the uncertain parameters, affects the model for the 2020-2030 time window. This sample will remain valid for the following time windows. In other words, there is only one sample draw per episode (see Figure **??**). Then, the agent takes a set of actions, affecting the environment that feeds back the agent with a new state and a reward. This goes on until the end of the episode. For the new episode, similarly to the Uncertainty Quantification (UQ) analysis (see Section **??**), the new sample of uncertain parameters for the new episode is drawn following the quasi-random Sobol' sampling technique [19].

To reduce the computational burden and maximise the exploration of the transition pathways, the learning phase of the agent has been done on the monthly model. Even though averaging time series of end-use demands and renewable productions brings some discrepancies (i.e. faster emergence of local Variable Renewable Energy Sources (VRES) and smaller electrification of the system), the main advantage of the monthly approach is its computational time, i.e. couple of seconds versus 15 min for the hourly model (see Appendix **??**).

Reward and success

The learning phase has been split into batches of 500 steps. For the first 100 steps of each batch, the RL-model collects transitions before learning starts. This makes sure replay buffer is full enough for useful updates. At the end of each batch, the up-to-date policy, i.e. the NN, is saved. This way, we can assess the progress in the learning process and its convergence (see Figure 1.2). 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. Right hand side of Figure 1.2 shows the success rate as the share of transitions meeting the CO_2-budget (see Section **??**) over the total number of attempted transitions, i.e. episodes. Even though this success rate is not what drives the agent's optimisation, it shows that the shape of the reward (see Figure 1.1) leads towards more and more successes.

During the learning process, the algorithm explores numerous transition pathways: 2037 successful transitions out of 10,751 attempts. Each pathway provides valuable insight into the best course of actions — the primary goal of reinforcement learning. As a side benefit, the collection of all explored pathways also identifies the intermediate milestones to reach and the range of actions that must be avoided or must be taken. Yet, the exploration during this learning process is not exhaustive. The trends provided below are therefore not proven. The randomness of the process and the number of explored transition pathways still give us high confidence.

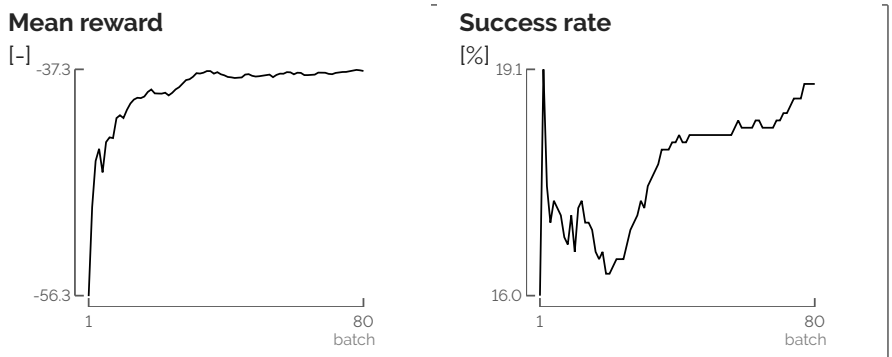


Figure 1.2. Mean reward and success rate of the different learning batches. The stabilisation of the reward curve shows a convergence of the learning process for the agent's point of view. The evolution of the success rate also shows the shape of the reward aims at more and more successful transitions.

There is a range in the reward where failures and successes overlap (see Figure 1.3). 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 1.1) or successful transitions that are more expensive. Besides this overlap, we observe that successes account for the majority of the cases where the reward is positive. This is another indication that the shape of the reward is appropriate in this exploration of transition pathways.

Considering the end of the time-window where the CO_2-budget is exceeded, the right hand side of Figure 1.3 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, mid-term actions are necessary to hope succeeding the transition.

States: Renewables and efficiency

In line with Rixhon et al. [20],uranium is considered as a non-renewable resource.

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 1.5). 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

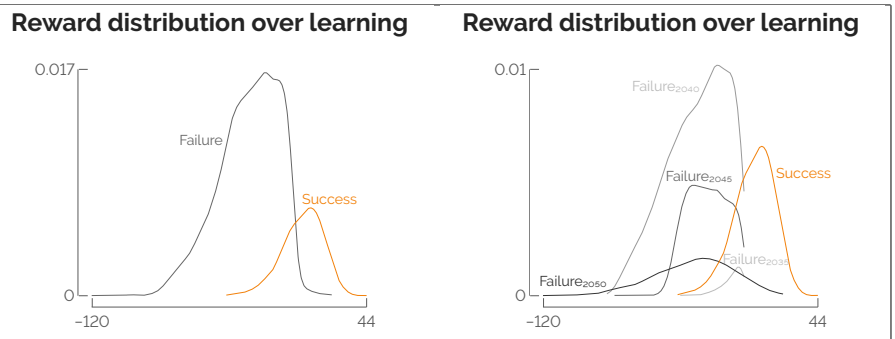


Figure 1.3. Reward distribution between successes and failures. Right hand side details at the end of which time-window the failure occurred. The "tipping year" is 2040 as failing the transition by 2040 represents 57% of all the failures. Beyond this point, through this learning process, succeeding the transition represents 38% of the cases.

different timing seems to lead to more successes. In other words, there is no clear set of actions that would support more effectively the transition.

To identify the actions that have an actual impact on the environment, we can check if they were 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 1.6). 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 as binding, actually limiting the objective function to be more optimal. In other words, binding constraints, when tightened, aggravate the objective value function.

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 1.7). 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 an effective impact on the environment only when it forces the system to be close to carbon-neutrality. The range over which limiting the use of fossil gas binds the

[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.

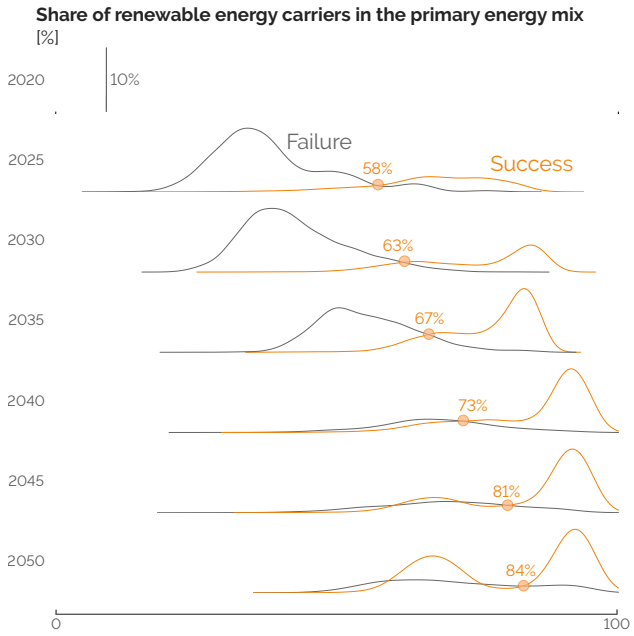


Figure 1.4. Exploration of the state space over the learning process with a focus on the share of renewable energy carriers in the primary energy mix. Integration of local VRES at early stages then massive import of electrofuels later are needed to secure successful transitions. Below a near-term threshold (about 60%), the chances of success are limited, i.e. no-go zones.

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).

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 159TWh of LFO are consumed by naphtha-cracker (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-optimum removes it of 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

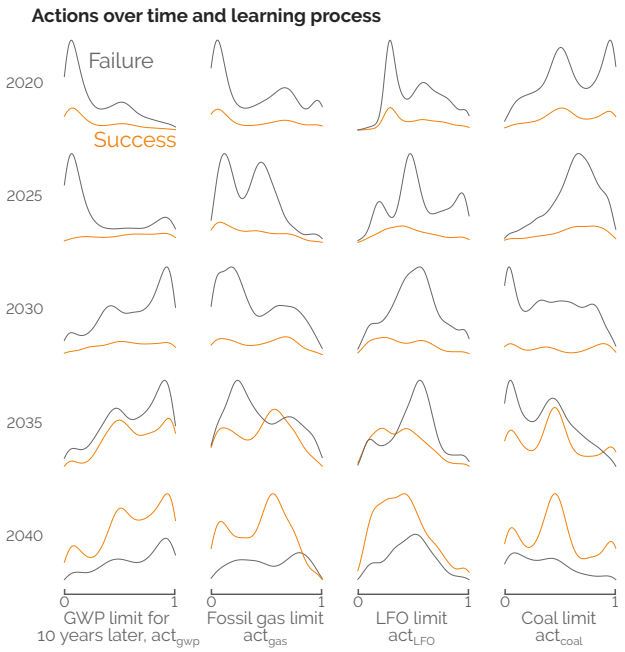


Figure 1.5. Over the whole learning process, actions taken by the agent. Besides the wide exploration of the action space, it does not seem to be any clear set of actions to take to support successful transitions.

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.

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 value of act$_{coal}$ still represents small consumption of coal.

- Assess the severity of learning on Monthly model, versus Hourly model. Compare the policies: one versus the other versus one followed by the other

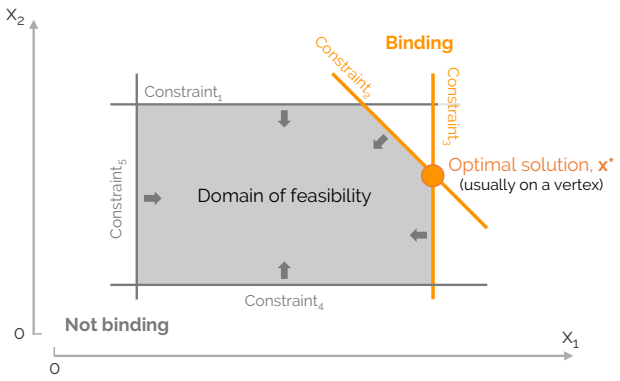


Figure 1.6. Binding versus non-binding constraints. In LP where the feasibility domain is non-empty and bounded, the constraints defined a convex feasibility domain in the space of variables (here, x_1 and x_2). The optimal solution usually locates on a vertex of this domain, i.e. the intersection of several constraints (here, constraints 2 and 3) limiting the solution. These constraints are considered as binding, i.e. having a limiting impact on the optimal solution.

1.3 Testing and comparison with references

How does it react to potential shocks (e.g. no fuels from 2040) compared to classic myopic?

Herman et al. [22] defined robustness as "… the fraction of sampled states of the world in which a solution satisfies all performance requirements"

Confirm here with what is said in [23]: Importantly, our results also show that near-term policy stringency is an important driver of cumulative Res-FFI-CO2 in climate change mitigation scenarios. If strengthening of NDCs fails, Res-FFI-CO2 will be even higher, not only because of additional near-term emissions, but also due to a decrease of economic mitigation potentials in the longer term caused by further carbon lock-in. Delaying the strengthening of mitigation action will increase the world's dependence on CDR for holding warming to well below 2°C, and is likely to push the 1.5°C target out of reach for this century.

1.4 Discussion

Potentially interesting to implement reward shaping to accelerate learning or guide the agent towards achieving the desired behaviour more efficiently. In our case, the agent

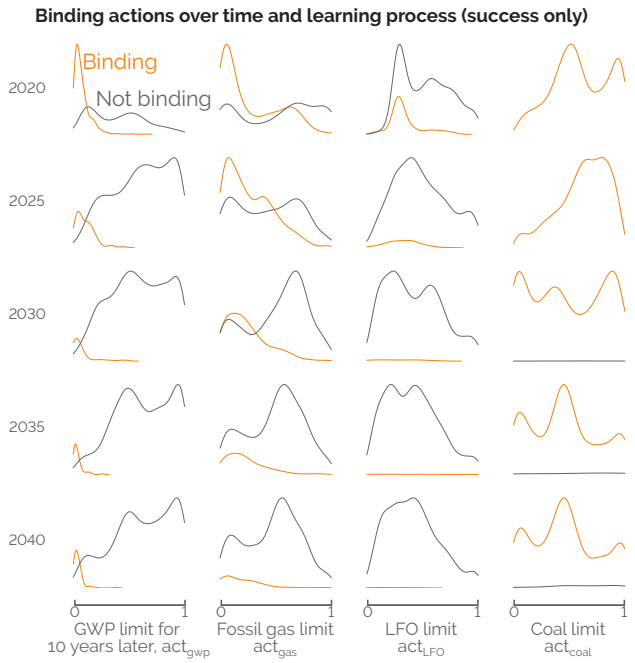
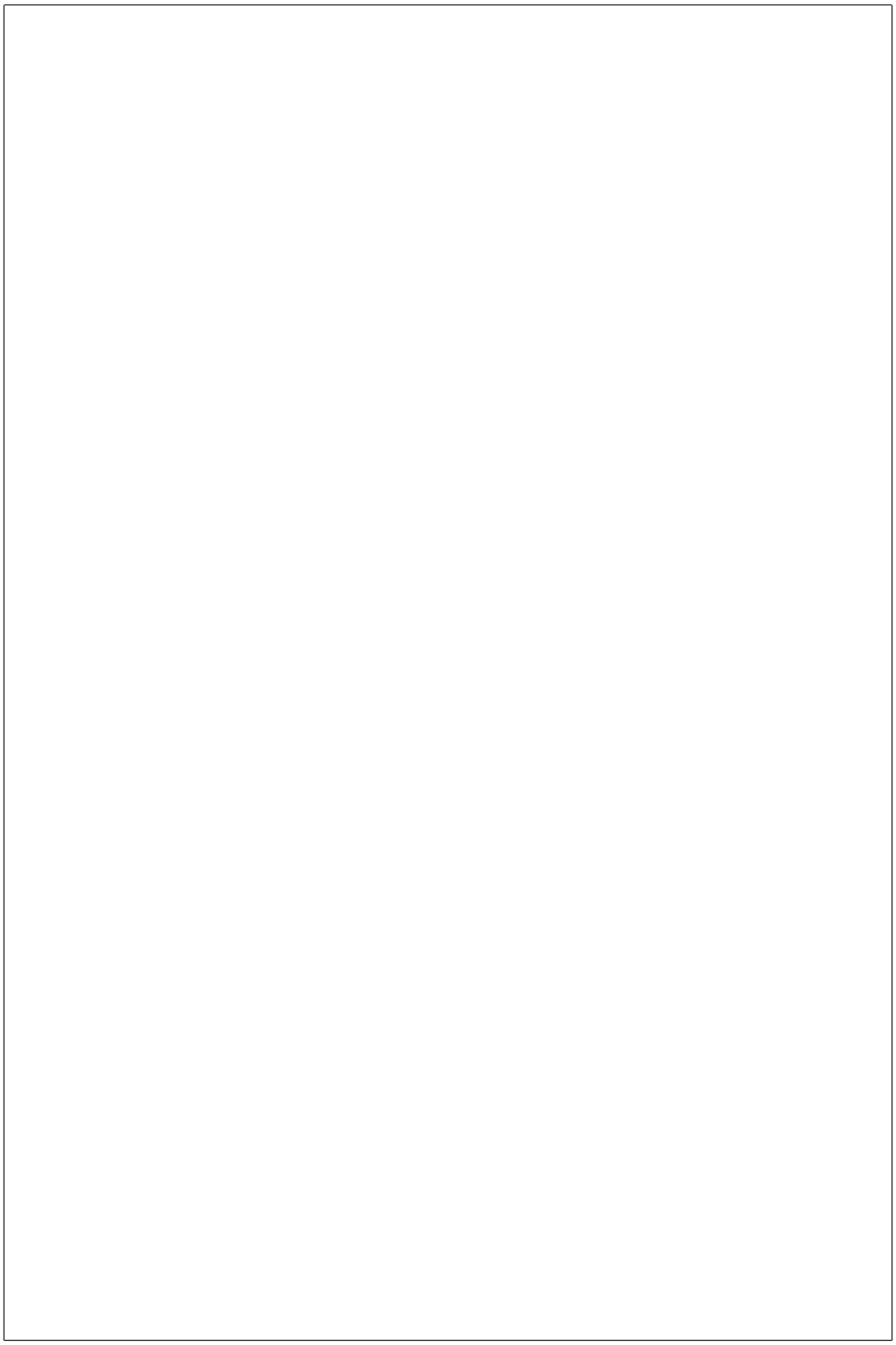


Figure 1.7. Out of the successful transitions, binding and not binding actions taken by the agent. Depending on the action and its timing, it is actually constraining the optimisation through EnergyScope Pathway or not. Sweet spots can be identified when considering the limits of GWP and fossil gas consumption. Limiting coal consumption is always constraining, unlike LFO that is "naturally" substituted by EnergyScope Pathway in the near-term.

receives a sparse reward signal, indicating success or failure at the end of an episode. However, this sparse reward signal may not provide enough information for the agent to learn effectively, leading to slow convergence or difficulty in learning the optimal policy. Reward shaping addresses this issue by providing additional, intermediate rewards during the learning process based on various heuristics, domain knowledge, or problem-specific insights. These intermediate rewards can help guide the agent towards desirable states or actions, making the learning process more efficient and effective. However, reward shaping should be applied with caution, as poorly designed reward functions can lead to unintended consequences such as suboptimal policies, re-

ward hacking, or overfitting to the shaped rewards rather than learning the underlying task (see Section 1.1).

Talk here about potential improvement of the approach like the Multi-Fidelity Reinforcement Learning by Cutler et al. [24], rather than just unidirectional transfer from low to high-fidelity



Bibliography

[1] C. Nicolas, S. Tchung-Ming, O. Bahn, E. Delage, Robust Enough? Exploring Temperature-Constrained Energy Transition Pathways under Climate Uncertainty, Energies 14 (2021) 8595.

[2] Intergovernmental Pael on Climate Change (IPCC), Global Warming of 1.5°C. An IPCC Special Report on the impacts of global warming of 1.5°C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change, sustainable development, and efforts to eradicate poverty, Technical Report, IPCC, 2018.

[3] W. Steffen, J. Rockström, K. Richardson, T. M. Lenton, C. Folke, D. Liverman, C. P. Summerhayes, A. D. Barnosky, S. E. Cornell, M. Crucifix, et al., Trajectories of the earth system in the anthropocene, Proceedings of the National Academy of Sciences 115 (2018) 8252–8259.

[4] W. Nordhaus, A question of balance: Weighing the options on global warming policies, Yale University Press, 2014.

[5] I. E. Grossmann, R. M. Apap, B. A. Calfa, P. García-Herreros, Q. Zhang, Recent advances in mathematical programming techniques for the optimization of process systems under uncertainty, Computers & Chemical Engineering 91 (2016) 3–14.

[6] L. G. Fishbone, H. Abilock, Markal, a linear-programming model for energy systems analysis: Technical description of the bnl version, International journal of Energy research 5 (1981) 353–375.

[7] A. Kanudia, R. Loulou, Robust responses to climate change via stochastic MARKAL: The case of Québec, European Journal of Operational Research 106 (1998) 15–30.

[8] D. Bertsimas, M. Sim, The price of robustness, Operations research 52 (2004) 35–53.

[9] S. Moret, Strategic energy planning under uncertainty, Ph.D. thesis, EPFL, 2017.

[10] R. Loulou, U. Remme, A. Kanudia, A. Lehtila, G. Goldstein, Documentation for the times model part ii, Energy technology systems analysis programme (2005).

[11] M. Howlett, M. Ramesh, A. Perl, et al., Studying public policy: Policy cycles and policy subsystems, volume 3, Oxford university press Toronto, 1995.

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

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

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

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

[16] P. Henderson, R. Islam, P. Bachman, J. Pineau, D. Precup, D. Meger, Deep reinforcement learning that matters, in: Proceedings of the AAAI conference on artificial intelligence, volume 32, 2018.

[17] R. S. Sutton, A. G. Barto, Reinforcement learning: An introduction, MIT press, 2018.

[18] 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, Technical Report, European Parliament, 2023. Official Journal of the European Union 2413, 1-77.

[19] P. Bratley, B. Fox, Implementing sobols quasirandom sequence generator (algorithm 659), ACM Transactions on Mathematical Software 29 (2003) 49–57.

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

[21] D. Bertsimas, J. N. Tsitsiklis, Introduction to linear optimization, volume 6, Athena Scientific Belmont, MA, 1997.

[22] J. D. Herman, H. B. Zeff, P. M. Reed, G. W. Characklis, Beyond optimality: Multistakeholder robustness tradeoffs for regional water portfolio planning under deep uncertainty, Water Resources Research 50 (2014) 7692–7713.

[23] G. Luderer, Z. Vrontisi, C. Bertram, O. Y. Edelenbosch, R. C. Pietzcker, J. Rogelj, H. S. De Boer, L. Drouet, J. Emmerling, O. Fricko, et al., Residual fossil CO2 emissions in 1.5–2 C pathways, Nature Climate Change 8 (2018) 626–633.

[24] M. Cutler, T. J. Walsh, J. P. How, Reinforcement learning with multi-fidelity simulators, in: 2014 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2014, pp. 3888–3895.