Developing and Deploying Sim-to-Real Reinforcement Learning Techniques with Applications in Energy Systems

Lucas Spangher

Electrical Engineering and Computer Sciences
University of California, Berkeley

Technical Report No. UCB/EECS-2023-4
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2023/EECS-2023-4.html

January 13, 2023

Developing and Deploying Sim-to-Real Reinforcement Learning Techniques with Applications in Energy Systems

by

Lucas Spangher

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Computer Science

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Costas Spanos, Chair
Professor Stuart Russell
Professor Duncan Callaway
Professor Stefano Schiavon

Fall 2022

Developing and Deploying Sim-to-Real Reinforcement Learning Techniques with Applications in Energy Systems

Abstract

Developing and Deploying Sim-to-Real Reinforcement Learning Techniques with Applications in Energy Systems

by

Lucas Spangher

Doctor of Philosophy in Computer Science

University of California, Berkeley

Professor Costas Spanos, Chair

Climate change requires a radical and complex transition in the way our energy sector generates and uses energy. Solar and wind energy are leading carbon-free power sources, but they are non-dispatchable, meaning that an operator cannot control when their generation occurs. As a result, their growth may be stymied unless supporting technologies can change how the rest of our energy system responds to their generation. Demand response is an important tool in a suite of supporting technologies that help smooth the introduction of non-dispatchable energy into the grid. The more effective demand response signals (prices) are on responses (deferring of energy by appliances and building systems), the more quickly solar and wind may decarbonize our energy system.

Demand response signals may be amplified by advanced controls, of which reinforcement learning is a prime example. We present two different environments for testing demand response price-setting at different levels of the grid. The first environment is MicrogridLearn, an environment that transmits prices to a collection of buildings and learns to set prices in a way that better shapes the energy and lowers energy costs for all buildings in its purview. The second environment is OfficeLearn, an environment that simulates behavioral response to prices. Using these two environments, we identify six key challenges in moving RL from simulation to reality, and present RL strategies, some novel, to overcome them. First, RL controllers implemented in the real world need to be data efficient: we present Offline-Online RL, Surprise-Minimizing RL, and extrinsic and intrinsic planning as solutions. Second, RL in the real world should be robust and guarantee safe action: we present a novel method, the *guardrails* planning model, and demonstrate that using a conservative decision process with a distributional prediction can help learning. Third, RL may get stuck in local optima: we present meta-learning over domain randomization as a technique to ensure agent robustness. Fourth, agents may be attacked: we demonstrate a novel adversarial attack on RL and present a defense. Fifth, energy applications may require real-world RL to protect privacy and generalize to new subdomains easily; we present the first ever application of Personal Federated Hypernetworks (PFH) to RL to accomplish both tasks. Finally, hyperparameter sweeps may entail large data consumption; we present a regression analysis of hyperparameter

sweep values to give a sense of hyperparameter-parameter strength.

Finally, we discuss how RL may be implemented in experiment. We give a prospective experiment plan. We present an API to connect the RL controller to the real world. We then discuss two prior experiments run in the same office setting: an A/B test of two different energy visualizations, and a measure of the persistence of the effects of energy reduction after the experimental period ended.

Our work contributes to societal knowledge in the following ways. We are the first to propose the use of RL for price-setting in energy systems, and we are the first to propose a Social Game as a mechanism to incentivize price sensitivity in an office setting. Of our RL methods, we are the first to propose adversarial poisoning during train time of algorithms. We are also the first to propose the use of personal federated hypernetworks for training new RL agents. Our other methods have been inspired by similar implementations in other fields, but are novel to the communities and application spaces in which we operate.

We hope that, from our work, the community may continue to iterate on RL architectures for price-setting, and may implement these techniques in experiment.

To my brother Alex for continually inspiring me with a clear-headed, kind, precise, and funny quality of thought, for keeping me striving for being a better person and scholar. Thanks for believing in me more than I believe in myself at times!

To my mom and dad for setting me up for the person I am today, for investing time, effort, and love in my education, and specifically for bribing me with a bicycle to leave the Department of Energy early to enroll at UC Berkeley. Thanks!

To Gustavo, for your love, patience, and giving, for your structure, and for holding me steady and helping me reach the end.

To COVID-19 for forcing me to do a lot more modeling and simulation than I otherwise would have done, for pushing me away from aspects of life that may have been incompatible with academic success. I will never forgive you for all you have taken away, but at least I was able to make the best of what you gave.

To UC Berkeley (generally) for the role it has played in the development of ideologically progressive thought and for being a shining light on a hill.

To a general scientific platitude that states that often those with less supervision during a Ph.D. will learn how to exert a executive, organizational, and methodological discipline on their own.

To UC Berkeley (specifically) for a hearty masterclass in learning executive, organizational, and methodological discipline on my own, and for instilling in me a deep awareness of and appreciation for the value of regular and disciplined mentorship.

To HiP, my student co-op, for endless amounts of beauty and color throughout the four years I lived there. Thank you for getting me through the pandemic, for teaching me to live cooperatively and to have patience with housemates and to appreciate and enjoy many tribes of people I would have otherwise not interacted with. Apart from generalized lessons, thank you for each of the dozens of housemates I had the opportunity to individually grow to love and appreciate. Thank you to Berkeley for the radical past that allowed for the development of such a special place.

To San Francisco, the sparkling gem of a city that continually takes my breath away. Thank you for never ceasing to amaze in your sights, your splendor, your vibrant social spaces, your warmth, and your community. Thank you for giving what you gave to the world and for continuing to create, all on the basis of a collective love. Thank you for creating the Radical Faeries and other wonderful communities. Thank you for irrevocably shaping my twenties and turning me into the person I am today. I will forever read your name and see in it the intricate beauty of hope and of a gentle light on the horizon: San Francisco, San Francisco, San Francisco.

And finally and more abstractly, to all those who dedicate their lives to addressing the climate crisis:

For those who came before me, who have struggled to recognize climate change for the dangers it pose, who laid the foundations for the technologies that we refine and expand on, who labored in green technology long before it seemed feasible or transformative, and who risked life and livelihood to fight for a better environment;

For those who will come after me, who dream of a new and more fair energy and economic system to base human civilization, who will expand the technologies we work on into systems of interlocking effectiveness, who may continue to carry the torch of civilization into a less forgiving climate;

Above all to those who may never be, as a result of our failure to act more quickly in the face of crisis.

Contents

Extended Contents

Acronyms

ACCEL Adversarially Compounding Complexity by Editing Levels. ix, 135, 136

ADMM Alternating Direction Method of Multipliers. xv, 30, 31

AEO Annual Energy Outlook. 5, 6

Ah Amp-hours. xvi, xvii

AHU Air Handling Unit. 10

AI Artificial Intelligence. 23, 94, 131

API Application Programming Interface. viii, xxii, 27, 46, 103–105

ARPA-E US Department of Energy's Advanced Research Projects Agency. 151

ASG After Social Game. 125

AutoML Automatic Machine Learning. 58, 95

BCE Before Common Era. 152

BMS Building Management System. 12

BOS Building Operating System. 12

BSG Before Social Game. 125

CE Common Era. 152

CMP Controlled Markov Process. 53

COVID-19 SARS-CoV-2. viii, 46, 106

CPU Central Processing Unit. 32

CREATE Campus for Research Excellence and Tech Enterprise. 108

DAggr Dataset Aggregation. vi, xxi, 48–53

DER Distributed Energy Resources. 28

DR Demand Response. v, xvi, 6–9, 23, 25–27, 34–37

DSG During Social Game. 125

EG&E Eastern Gas and Electric. 82

EIA US Energy Information Agency. 5

EV Electric Vehicle. 7, 8

FGSM Fast Gradient Sign Method. 79

GHG Greenhouse Gas. 2

GMM Gaussian Mixture Model. 121, 123

GPyOpt Gaussian Optimization using GPy. 58

GW gigawatt. xx, 6, 8, 152

HMM Hidden Markov Model. xxiii, 110, 111

HVAC Heating, Ventilation, Air-conditioning and Cooling. v, xx, 9–12, 20, 25, 34

IoT Internet of Things. 34

kW kilowatt. 9

kWh kilowatt hour. 118, 120, 125, 152

LC Login Count. 116

LCOE Levelized Cost of Energy. 2–4, 150

LSTM Long Short Term Memory network. xxi, 13, 58–60

MAML Model Agnostic Meta-Learning. xxi, 72–77, 130

MDP Markov Decision Process. ix, xviii, 13, 14, 20, 39, 86, 130, 131

ML Machine Learning. 1, 15, 16

MPC Model Predictive Control. 12, 13

MW megawatt. 153

MWh megawatt Hour. 118

NCCS National Climate Change Secretariat of Singapore. 112

NN Neural Net. 13, 14, 58, 59, 64, 65, 72, 89

NREL National Renewable Energy Laboratory. xx, 3, 5, 7, 8

OLS Ordinary Least Squares. xxi, 58–60, 116

OO Out of Office. 116, 124, 125

PFH Personal Federated Hypernetwork. xvi, xxii, xxv, 23, 85, 86, 88–95, 100, 138, 155

PG&E Pacific Gas and Electric. 38, 153

PID Proportional Integral Derivative. 10, 11, 20

PLR Positive Level Replay. 136

POMDP Partially Observed Markov Decision Processes. 14, 39

PPO Proximal Policy Optimization. v, vi, xxi, 17–19, 55–57, 62, 72–77, 89, 96

PV Photovoltaic. xx, 3, 4, 6, 32

RA-SAC Risk-Aware Soft Actor Critic. xxi, 65–67, 69, 70

REINFORCE REward Increment = Nonnegative Factor x Offset Reinforcement x Characteristic Eligibility. 18

RESTful REpresentative State Transfer. 104

RL Reinforcement Learning. iii, v–ix, xvii, xx–xxii, xxiv, 1, 12–18, 20–23, 25, 27, 30–34, 36, 38, 41, 44–49, 51, 53–55, 60, 61, 63–69, 72, 76–82, 84–92, 95, 99–101, 103–105, 130, 131, 134, 135, 137, 138, 153, 155

RLLib Reinforcement Learning Library. 68, 104

RTP Real Time Pricing. xxi, 36, 38, 59, 60

SAC Soft Actor-Critic. xxi, 17, 19, 31, 41, 48, 50–52, 61, 63, 65–70, 154

SB Smart Buildings. v, xx, 9, 20, 21, 23, 153

SGD Stochastic Gradient Descent. 73, 86

Sim-to-Real Simulation-to-Reality. iii, vi, viii, 23, 46, 99, 101, 136

SinBerBEST Singapore Berkeley Building Energy in the Tropics. 106

SMiRL Surprise Minimizing Reinforcement Learning. vi, xxi, 15, 53–57

SQL Structured Query Language. 104

SVM Support Vector Machine. 14, 122, 154

TI Treatment Indicator. 116

TN Treatment Number. 116

TOU Time of Use. xxi, 27, 29, 35, 36, 38, 41–43, 48, 51, 52, 59, 63, 65, 67–70, 133, 153

TT Treatment Time. 116

VAV Variable Air Valve. 10

VCG Vickrey-Clark-Groves. 102

WI Weekend Indicator. 115, 116, 124, 125

Math Used

Symbol	Definition
$\vec{x}$	An indication that some variable x is a vector
$\overline{x}$	The mean of some variable x
$\hat{x}$	The observed value of some unknown variable x
x^*	Denotes the optimality of an x^* relative to all other members x of the set X; i.e. $f(x^*) \geq f(x)\ \forall x \in X$
$\vec{x}_*$	Also denotes the optimality of a vector, $\vec{x}$, used sparingly and only when the arrow and star interfere with each other in typesetting

Table 0.1: List of symbol modifiers used

Symbol	Definition
$\|\|\vec{x}\|\|_i$	The i^{th} norm of some vector $\vec{x}$
e^x	The exponential function (2.718^x), raised to the x power
ln	The natural log
$\mathbb{1}(x)$	An indicator function of x where x is a boolean expression. It evaluates to 1 if x is true and 0 otherwise.
$\mathbb{E}_{\Xi}(f(\Xi))$	The expected value of an expression f with respect to some distribution Ξ
∇_x	The gradient with respect to a multivariate variable x
$\frac{\partial f(x,y)}{\partial x}$	The partial derivative of a possibly multivariate function, f, with respect to one variable, x
$f(x)$	Used as a generic function, when specifying is not necessary
g	Used as a generic objective, when specifying is not necessary
$X^\intercal$	The transpose of some matrix X
$(a, b]$	A range between a and b where a parenthesis indicates exclusion and bracket indicates inclusion
$==$	A test for equality
$:=$	Equal by definition

Table 0.2: List of functions used.

Symbol	Definition
α, λ	Used as generic scaling or weighting hyperparameters; i.e. SMiRL weight or entropy weight
β	Exponential decay parameter
β_1, β_2	First and second moments of ADMM
γ	Discount factor
ι	An adversarially poisoned object
δ	Temporal difference error
ϵ	A noise term
η	A task, i.e. $\eta \in \mathcal{T}$ (see below)

θ, ϕ, ψ	Symbols used to designate parameters
μ	A sample mean
ν	A ratio of the probability of an action under policy parameterized by new parameters to the probability to action under a policy parameterized by older parameters
ξ	A hypernetwork as in Personal Federated Hypernetwork (PFH)
π	Pi, i.e. 3.14159...
π_θ	A policy function parameterized with respect to θ
Π	A set of policies
ρ	An adversarial perturbation bound
τ	A trajectory, i.e. sequential observations of (s, a). If τ is subscripted, i.e. τ_{π_θ}, then the trajectory is generated using policy
a, a_t	An instantiated action and an action specifically taken at time t, respectively; used somewhat interchangeably
A	The Advantage function, defined as the difference in value between a specific action and the average action at a state
b	The baseline counterfactual demand, i.e. the demand that would have occurred if some intervention like demand response or battery discharge had not happened (see d and e.)
B	Battery capacity in Amp-hours (Ah)
c	Cost, possibly a function of some other metrics.
d	Gross energy demand, after some behavioral Demand Response (DR) intervention but before battery discharge (see b and e)
D	A dataset
e	The total net demand after behavioral energy demand response might reduce or defer energy and after battery discharge or solar generation might offset energy. I.e., first some counterfactual baseline b exists, then some d is generated based on a DR intervention, and then some net e is observed
E	Total energy
g	Gross local generation
G	A Guardrail rule
h	A specific hour in a set of hours
H	A total number of hours considered (i.e. a workday, or full day)

i, j	Denotes indices, freely defined locally depending on context, and many times used in tandem in the same equation
I	An incentive
J	An RL objective; may be expressed as a function of θ
k	May be used to express a small number of tries or categories, i.e. k-shot learning
K	Population level categories
L	A loss, often as a function of model parameters θ and observed data y
M	Environmental steps
n, m	Counts, freely redefined locally, usually representing sample sizes, vector length, etc
N	A population level count
O	Order, as in order of operations
p	Price, often used with subscript to denote who is setting the price
P	A probability. Includes transition probabilities, action probabilities, and planning model probabilities. Can be used as a function or a singleton
q, Q	An observed value and an unobserved or functional form of an (s, a) pair, respectively
r	A reward function in RL
s, s_t	A specific state and a specific state at a specific time t, respectively
T	max time step or steps
u	Battery behavior (Ah discharged or charged.)
U	Number of workers
v	Vector embeddings
V	expected discounted reward from a state, s
w	Weights, often freely and locally defined
W	World
X	Generally used to denote independent variables or distributions
Y	Generally used to denote dependent variables or distributions
z, Z	Latent variables and distributions, respectively
$\mathcal{A}$	The set of all possible actions in an environment; the action-space
$\mathcal{B}$	The Bernoulli distribution

$\mathcal{E}$	Set of Environments
$\mathcal{F}$	A planning model
$\mathcal{G}$	The Geometric distribution
$\mathcal{H}$	Entropy
$\mathbb{H}$	Honest values (in terms of adversarial training.)
$\mathcal{M}$	An MDP
$\mathcal{N}$	The normal distribution, often as a function of μ and σ, i.e. $\mathcal{N}(\mu, \sigma)$
$\mathbb{R}$	The set of real numbers
$\mathcal{R}$	The space of rewards, used in general definitions of MDPs
$\mathcal{P}$	The set of transition probabilities in an MDP
$\mathbb{P}$	The set of possible prices, implying bounds on each element.
$\mathcal{S}$	The set of all possible states in an environment, i.e. the state-space
$\mathcal{T}$	A set of tasks, i.e. $\eta \in \mathcal{T}$
$\mathbb{U}$	Regret measure
$\mathbb{W}$	An indicator function of whether an action is accepted in the real world (guardrails)
$\mathbb{Z}$	The set of integers

Table 0.4: List of symbols used.

Symbol	Definition
$\forall$	"for all" members, generally of a set, i.e. $\forall x$
$\in$	A member is inside a set, i.e. $x \in X$
$\subset$	A set is a subset of another set, i.e. $A \subset B$
$\simeq$	Is almost equal, i.e. $x \simeq y$
$\propto$	A value is proportional to another, i.e. $x \propto y$
$\geq$	Greater or equal to
$\leq$	Less than or equal to
$\equiv$	Equivalent to

Table 0.3: List of (set) relations used.

List of Figures

List of Tables

Acknowledgments

- I gratefully acknowledge Costas Spanos, my advisor, for allowing me to pursue interests as I developed them. The direction that Costas put on my studies was always one of carrots – i.e., studying in the buildings space could lead to interesting experiments using SinBerBEST's assets – rather than sticks – i.e. I cannot study X, – and for that, I am very grateful. Money and support was never an issue, especially during the pandemic, and Costas' work behind the scenes to accomplish this is commendable. I gratefully acknowledge the time and frequency of meeting that Costas gave me, and I am grateful for his wisdom in navigating the twists and turns of life in grad school.

- I gratefully acknowledge all of my undergrads for their time, energy, enthusiasm, and kindness. Specifically, Austin Jang has taught me so much in engineering as well as friendship and it has been an honor to watch him grow and help publish some of his findings. The work in Personal Federated Hypernetwork (PFH), guardrails, metalearning, and adversarial poisoning may not have existed without his input. Akaash Tawade was my first undergrad, and I will always remember and be grateful for his showing me how much fun it could be to lead a team on a project. I owe him recognition in the persistence and clustering work. Akash Gokul will always amaze me with his brilliance and self-driven technical accomplishments, and I acknowledge him and Joseph Palakapilly for their efforts in first getting RL to work on OfficeLearn. I will be grateful Manan Khattar for his kind, mature, and thoughtful participation in my group. I owe him the development of the API. Will Arnold is a first class mind and talent, and Tarang Srivistava is wonderfully competent and an honest scholar. Thanks to both for the SMiRL work. Josh Lorincz is extremely bright, self-motivated, and persistent, and Japjot Singh is exceptionally talented in his development of normalizing flows. Larry Yan has a bright future ahead of him, and the work he continues to do with Austin is going to land well.

- I gratefully acknowledge Selvaprabu Nadarajah for sticking through it with Austin and I to get a beautiful journal paper out of the work we did in guardrails. Selva offered the most engaged and insightful mentorship during our time, and I will be grateful for the intensity with which he interrogated our data.

- I gratefully acknowledge Utkarsha Agwan for an incredibly fun and rewarding few weeks putting together and testing MicrogridLearn. Our collaboration showed me the possibilities of what intense and interesting collaboration could bring, and formed a foundation of interesting science that I am proud to have contributed to.

- I gratefully acknowledge Orr Paradise and Sam Gunn for another intensely fun and productive period producing the adversarial poisoning paper. I learned so much, and was so grateful to be a part of and witness to the interesting theoretical work that was done with you and Austin.

- I gratefully acknowledge the whole RAISE lab – including Hari Das, Alex Devonport, Wendy Lin, Utkarsha Agwan, and Ming Jin, Ruoxi Jia and Ioannis Konstantakopoulos, for all of the support and camaraderie throughout the years, for understanding the

particular setting we found ourselves in, for working together to produce our group meetings. Thanks for being allies and friendly faces. Thank you Alex for carrying me through Reinforcement Learning in Fall 2018 when the department continually rescheduled my preliminary exams: I am sure you are as surprised as I am that I would one day produce a dissertation on RL.

- It is with enormous and immense gratitude that I thank the Advanced Research Projects Agency at the US Department of Energy (ARPA-E) for providing a cornerstone of community and professional guidance that launched me to where I am today. It was an incredible year that I spent in 955 L'enfant plaza among people who were truly motivated to make a difference in the world of green energy and climate change. I got to indulge the beauty and wonder of advanced technologies every day and was inspired by all that was possible. Thank you to Ryan Umstattd, Jenny Gerbi, Daniel Northrup, Ellen Williams, Sue Babinec, Dawson Cagle, David Brown, Adrienne Little, Joe Rollins, Kirsten Brown, Ann Xu, Pat McGrath, Christopher Atkinson, and many many others for the incredibly formative professional experience. To date, this is the closest professional community that I have held, and I am deeply appreciative for the wealth of technical experience it gave me. I am also deeply grateful to everyone who worked at ARPA-E and gave up financial earnings to pursue careers in green energy.

- A brief thanks to my dissertation committee (Duncan Callaway, Stefano Schiavon, and Stuart Russell) and my quals committee (Anca Dragan, Pieter Abbeel, and Sergey Levine.)

- A huge thanks to Shirley Salanio for her incredible competence, prompt and professional replies, and kindness.

- Thanks to the small queer engineering team who was around me throughout the first three semesters of grad school: Kelly Fernandez, Zoe Cohen, Alex Devonport, and ally Hari Das. I appreciate your friendship and support. It was great company to start grad school in. Thank you to Zoe for continuing to carry the torch in QUICSE, the Queers in Computer Science and Engineering, who briefly, on a bank statement, were the Queens in Computer Science and Engineering.

Chapter 1

Introduction

1.1 Motivation in Energy and Climate Change

Climate change threatens myriad processes that power our modern world and its various lifeforms: be it direct, obvious, and significant impacts like ecosystem collapse (Nadeau et al., 2022; Canadell and Jackson, 2021), food security (Gregory et al., 2005), disease spread (Rohr et al., 2011), war (Zhang et al., 2007), or the vast movement of people[1] (Masson-Delmotte et al., 2021; Piguet et al., 2011), or indirect and non-obvious impacts such as the probability of fights during a baseball game (Salehyan, 2014; Mooney, 2014) or the increased probability of volcanic activity[2] (Cooper et al., 2018). The phenomenon is complex and varied, and it is possible to learn new things about it every time one addresses the subject[I3]. I start my dissertation assuming some personal knowledge and relationship from the reader to the breadth of the threat that climate change poses to society[II].

What may be less well understood by the casual reader of this dissertation[4] is the physical basis by which climate change occurs. For an important discussion on the physical basis for climate change and on the development and trustworthiness of climate models, please see Appendix I.

1.1.1 The Energy Transition

Swift and radical transition away from carbon emitting fossil fuels is generally seen as a way to avoid the worst effects of climate change (Delina, 2017; Creutzig et al., 2014). Energy

[1]The most significant direct impacts that are not noted in the main text are: desertification of previously fertile lands (Le Houérou, 1996), rising sea levels due to melting glaciers (Lindsey, 2020), more severe and intense storms due to warming oceans (Woodward and Samet, 2018), and changing rain patterns (Hendrix and Salehyan, 2012).

[2]Volcanic activity may result from a decrease in downward pressure from land-based glaciers as they shrink and lose mass.

[3]For this dissertation, I will use roman numerals to indicate chapter endnotes and arabic numerals to indicate footnotes. I will keep footnotes limited to strict factual or meta-asides that may be helpful for contextualizing direct writing intentions. I have put climate change chapters into their own appendix to give them special importance.

[4]I assume that some readers may come from Machine Learning (ML) or Reinforcement Learning (RL) background and may enjoy learning about the history and physical basis of climate change.

Transition requires a very complex coordination of different technologies (Smil, 2010), which we will discuss. In framing the goals of the Energy Transition, we will for the body of this dissertation note only that the Energy Transition is best understood as providing a system of alternatives to a fossil-based energy system which preserves the freedoms that fossil energy provides: freedom of movement, freedom of material accumulation, and freedom of time (Williams, [n. d.]). For further foray into the Energy Transition's philosophical goals[III], constraints[IV], and ethics[V], and please see the chapter endnotes.

The Energy Transition may face material bottlenecks[VI], but because the technology of the Energy Transition *supports* the generation of energy rather than *composes* the fuel necessary for generation (Dresselhaus and Thomas, 2001), our current market structure is well designed to facilitate substitutions [VII] to bottlenecks[5]. An alternative system of electricity will be composed of a coordination of many changing parts (Markard, 2018). My dissertation will focus on facilitating one of these parts.

1.1.2 Leading Carbon-Free Electricity Generators are Non-Dispatchable.

For the purposes of this writing, we will define *power* as the rate at which *energy*, or the capacity to do work, is generated. We define *alternative power* generation as any power generation that does not meaningfully produce Greenhouse Gas (GHG) emissions during operation[6](Michaelides, 2012). The main types of alternative power, currently, are *hydropower, nuclear, wind generation* and *solar generation.* Here we define the *dispatchability* of a generator to be the ease with which an operater may choose the level of power that it is producing (please see (Mudumbai et al., 2012) for a definition of the economic dispatch problem.) We define *volatility* to be the uncontrolled variability with which a generator produces power, whether or not it is predictable.

1.1.2.1 Dispatchable Carbon-free Generation

Hydropower The most significant current form[VIII] of *hydropower* is large-scale and dam-based, which generates energy by harnessing the potential energy of water (Cernea, 1997). These generators are dispatchable because an operator may directly control the dam, letting more or less water go through the turbines in a dam and continue downstream (Carvalho et al., 2011). Dam-based hydropower, while featuring a low Levelized Cost of Energy (LCOE) and dispatchable (Killingtveit, 2019), is unlikely to comprise a significant part of new generation: it is location limited and it may be increasingly difficult to create new dams as water politics become more fraught (Scudder, 2012). LCOE takes into account manufacturing and operating costs (Ueckerdt et al., 2013); for a thorough definition of LCOE and a discussion on the "tyranny of LCOE", please see the chapter endnotes[IX]

[5]Thus, we believe that it is inappropriate to fully criticize the Energy Transition for only being a continuation of the current material status quo.

[6]There is of course life-cycle carbon dioxide emissions in the manufacturing and installation of the equipment, and small carbon emissions in the operation of the equipment (i.e., imagine the carbon released from boats that drive around offshore wind turbines to maintain them.) We ignore these in our definition.

Nuclear Power *Nuclear* power refers almost exclusively to *nuclear fission*, i.e. the breaking apart of large atomic nuclei in order to capture the resulting energy from broken atomic bonds (Cameron, 2012). For a brief discussion distinguishing nuclear energy from chemical energy[X] and the mechanics[XI], and newer generations of fission reactors[XII], please see the chapter endnotes. Nuclear fission is not exactly dispatchable, in that an operator cannot easily turn down a reaction without destroying the ability for those specific cores to be turned back on; however, as fission is an intensely controlled reaction, it is non-volatile, and operators can guarantee a baseline supply of energy (Guoqiang et al., 2021). However, despite nuclear's reliability and cost-effectiveness (Schwarz and Cochran, 2013), a large-scale energy transition strategy dependent on nuclear fission alone would be difficult given political resistance to nuclear due to perception of safety (Schmidt et al., 2015). For brief discussions of the safety of current reactors[XIII] and nuclear fusion[XIV], please see the endnotes.

1.1.2.2 Non-Dispatchable Carbon-Free Generation

Wind Power *Wind power* harvests energy from the motion of air across the earth's surface. The recognizable horizontal axis wind turbine[7] has become one of the more exciting developments in green power development, consistently decreasing in Levelized Cost of Energy (LCOE) from 1980-2020 until it has among the lowest LCOE of any power source, including fossil fuels (Beiter et al., 2021). Wind power's decrease in LCOE has in part to do with improving scales of technology[8] as well as increasing scales of production. (For a short history[XV] long aside on technological scaling[XVI], please see the chapter endnotes.

Wind has a major vulnerability: it is both volatile, non-dispatchable, and tough to predict (Tian, 2021). In fact, according to the National Renewable Energy Laboratory (NREL), "for any particular generator, there is an 80% chance that wind output will change less than 10% in an hour and a 40% chance that it will change 10% or more in 5 hours" (NREL, 2012). Non-dispatchability is improved by offshore wind technologies, which harvest from more constant wind bases (Brunner et al., 2020). However, alone, wind power cannot meet the promises of modernity as it constrains when resulting energy can be used.

Solar Power *Solar power* shares some similarities with wind[XVII]. It encompasses a broad range of technologies, broadly taxonomized as technologies that harvest solar thermal energy[9], and technologies that directly convert solar energy to electicity[10], called *photovoltaics*. Photovoltaic (PV) generally function by stacking layers of one-way electron-impermeable

[7] Of course, "wind power" is an umbrella term covering many types of technologies, the most notable of which are horizontal three pronged turbines, vertical axis turbines, and offshore floating turbines of both types.

[8] One large technological trend has been the discovery that larger wind turbines tend to reduce the LCOE as there tend to be some fixed costs to manufacturing and installation of each turbine that are independent to its size.

[9] Concentrated solar energy generally takes the form of a large-scale plant that focuses rings of concentric mirrors towards a central heat collector. It used to be one of the most promising solar technologies given its LCOE was the lowest within solar. However, the LCOE of concentrated solar was overtaken by the LCOE of photovoltaics (Dowling et al., 2017).

[10] The author also has a tattoo of concentrated solar thermal energy. If the reader reads this footnote, they are welcome to ask for a picture.

material on top of each other, with doping materials like boron that help light permeate and absorb the energy to lift electrons to higher layers of material. Prevented from returning directly to the source atom, the electrons accumulate and eventually force themselves to "go around the long way", passing through wires that harvest their current (El Chaar et al., 2011). PV are the dominant form of solar energy production, as they are modular and have low LCOE. Unfortunately, although solar power is more predictable than wind power, it is perhaps even more non-dispatchable (Emmanuel et al., 2020). Solar systems produce most of their power towards the middle of the day when the sun is directly overhead, and level off to zero by the time night falls.

1.1.3 Energy System Cost

A discussion of the various features of energy brings around a simple but powerful observation (Koningstein, 2022; Koningstein and Fork, 2014):

> *The viability of alternative energy should not be based on alternative energy cost alone, but rather, by alternative energy* ***system*** *cost.*

We will now discuss exactly what this means.

1.1.3.1 Problems with Volatility

Electrical power is distributed from generators to consumers by way of a grid, an incredibly complex collection of physical wires. Electrical power is bid onto the grid based on demand predictions that fluctuate daily. Thus, electricity markets must be controlled to an exquisite degree. Indeed, given the interchangeability of electrons generated, it is amazing to think that when one turns on an electric water boiler for their morning coffee, a generator many miles away is likely harvesting electrons from coal burning. The orchestration of such a complex "organism" is truly an accomplishment of humanity.

Where energy is produced in the grid effects whether it can be sold at all. If a wind farm is generating many miles away from a city center, it may be capacity limited by some outgoing transmission wire. Generation that is larger than the capacity of the wires connecting it to demand is sent into ground, i.e. thrown away, as it would damage the wire and various transformers that compose the system (Fork and Koningstein, 2021; Bird et al., 2016).

When energy is produced also effects whether it can be sold. Even supposing high capacity transmission lines connect a wind farm to demand, if the wind farm were generating at some hour during the night when energy demand is low, then the energy would be sent into ground as well. Too much energy causes an increase in electrical frequency that can damage appliances across the network. Thus, energy is often *curtailed*, or sent to ground (Bird et al., 2016). In 2020, 3% of all California's solar produced was sent into ground, and on some days that number reached 20-25% (O'Shaughnessy et al., 2020). Here, the speed at which other generators in the network can react is important: some peaker plants of natural gas may spin generators up and down in a matter of seconds, but other generators, such as coal or nuclear, require hours to spin up and down. One may not rely on flexibility from other generating sources to aid in high solar and wind penetration, especially as we desire the reservoir of flexible generators to shrink as we reach higher decarbonization goals.

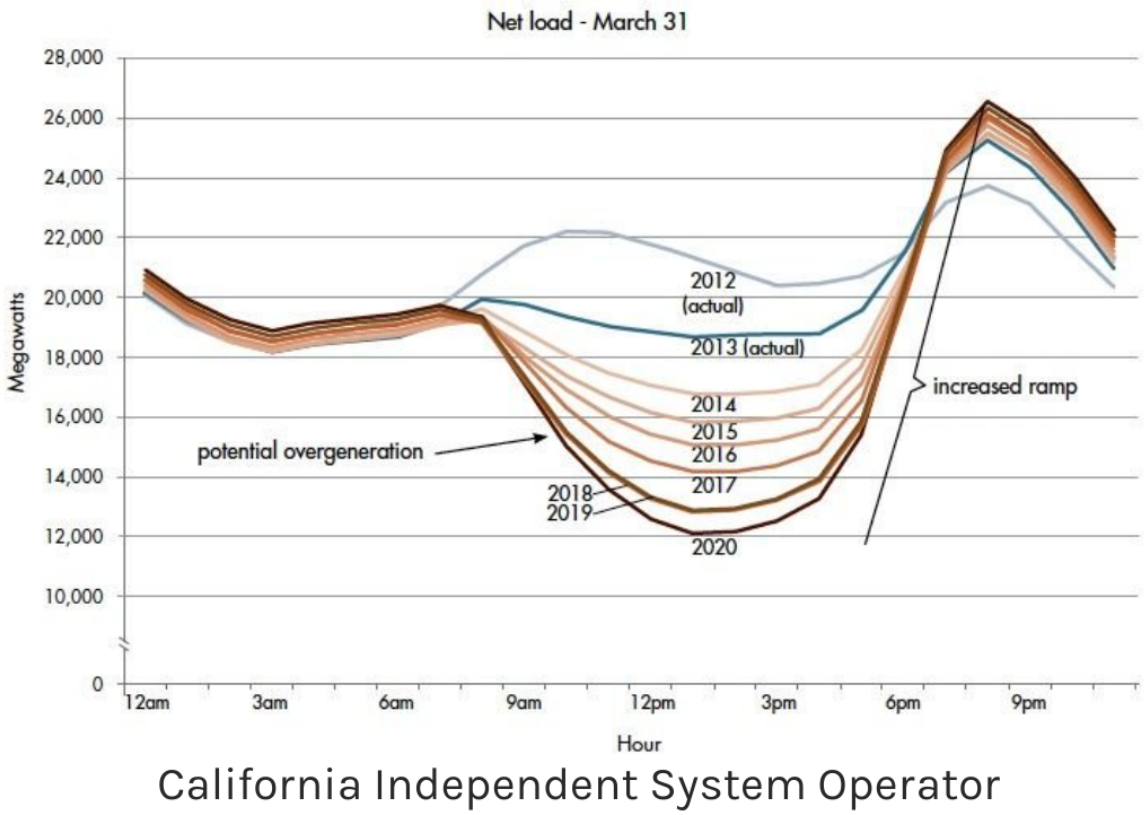


Figure 1.1: One day of net demand (gross demand minus solar production) in the grid, heavily smoothed. Reproduced from NREL (Denholm et al., 2015).

Figure 1.1 is a famous and extremely influential chart produced by the Department of Energy and affectionately nicknamed the "Duck Curve" due to the resemblance to the outline of a duck[11] (Denholm et al., 2015). The curve shows a heavily smoothed net demand on the Californian grid on the same day over multiple years. Due to the statistical averaging of lines on the Duck Curve, it is difficult to see the danger inherent in approaching zero: fluctuations around the mean are likely to cause negative net load and thus risk energy curtailment.

It is important to note that the problem of volatile energy is starting to rear its head *now*, when renewable energy still makes up a small portion of the overall grid (Zhou et al., 2016). Indeed, the statistics quoted about California's grid described grid dynamics where only 17% of all energy is solar and 6% of all energy is wind (O'Shaughnessy et al., 2020): how might these change in the future when these resources push 50-60% each?

The Annual Energy Outlook (AEO) model, produced by the US Energy Information Agency (EIA) (Center, 2020), has helped National Renewable Energy Laboratory (NREL) estimates the value of additional photovoltaics on the grid (please see Figure 1.2 (Stoll et al., 2017).) Under a scenario with little flexibility (see below for technologies that create

[11]Here we speculate that the US Department of Energy, despite the boundless extents of its whimsical creativity, did not in fact come up with this name.

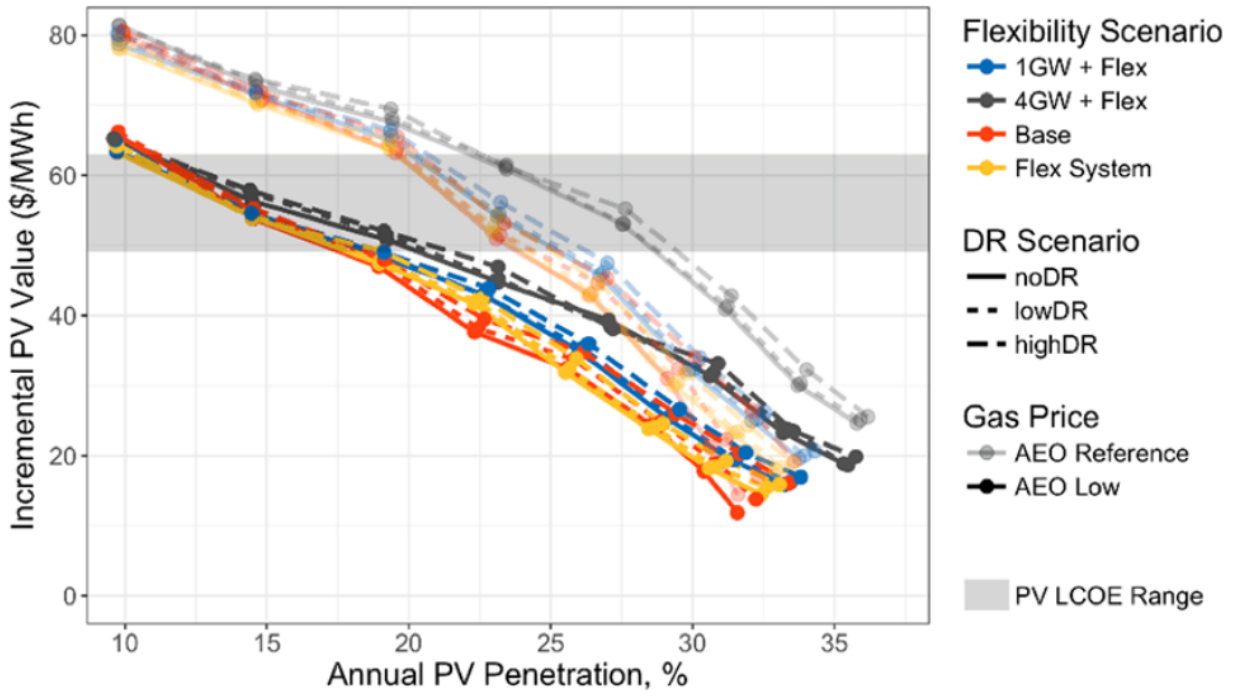


Figure 1.2: Cost of each PV installed decreases as the total penetration of PV increases, different scenarios considered by gigawatt (GW). Reproduced from (Stoll et al., 2017).

flexibility), additional photovoltaics dropped below the price of the photovoltaics; i.e., their value does not justify their cost.

However, in the AEO, if flexibility is built into the grid, photovoltaics are able to retain their value. Thus, it is absolutely essential *now* to develop solutions that ease the addition of volatile energy sources onto the grid.

1.1.3.2 Technologies to Address Non-Dispatchability

Thankfully, there exist some methods to address energy sources that are non-dispatchable: geographic averaging, energy storage, and energy Demand Response (DR). We will cover each now.

Geographic Averaging Geographic Averaging is an approach that aims for larger, more connected grids, so that generators that can service the grid are spread out over larger areas. In the case of wind, geographic averaging essentially harnesses the *Central Limit Theorem* of probability: if a distribution made up of a few random samples per hour of the day has high variability, taking more samples will result in a less variable, more normal distribution. The power output of a larger geographic area essentially will average out volatility in generation (Kwak and Kim, 2017). In the case of solar, geographic averaging increases the longitudinal and latitudinal surface area over which the sun may shine, which makes the daily cycle over a narrower latitude and longitude range less pronounced and more matched to energy demands,

which tend to peak between 7am - 10am and 4pm - 9pm (Gyamfi et al., 2013). In the US, the main way to enable geographic averaging is through grid interconnects of main grid areas[12] Of course, there are drawbacks that geographic averaging cannot fully solve: solar cannot be averaged into the grid in when the United States' East Coast experiences its 7am, and there are seasonal variations in wind that affect the entire US (Bloom et al., 2021; Brinkman et al., 2021).

Energy Storage Energy storage may be thought of as moving energy throughout time. "Energy storage" encompasses perhaps the largest diversity of technologies of any alternative energy umbrella term used thus far. Energy storage may be broken down into chemical storage (i.e., batteries) and physical storage (i.e., pumped hydropower, flywheels, gravitational storage). In chemical storage, most notable at time of this writing is Lithium-Ion battery technology, as it commands the storage market with enormous factories and scales of production (Kim et al., 2019). In the future, large flow-based batteries[XVIII] may be emerging for stationary grid applications, and solid-state batteries[XIX] may be emerging for Electric Vehicle (EV) applications[13] (Rahardian et al., 2019). In physical storage, *pumped hydro* is the most notable (Rehman et al., 2015). Pumped hydro is a system consisting of tanks or pools of water at different elevations that intakes energy in order to move water to higher stages and then release energy by allowing water to flow through turbines to lower pools. It remains the most efficient form of energy storage writ large, with over 90-95% of energy recoverable[XX]. However, pumped hydro is limited by geographic availability (Yang and Jackson, 2011).

In general, energy storage turns non-dispatchable energy into dispatchable energy, thereby making it a natural companion to wind and solar. However, it requires material infrastructure, time, and thoughtful construction to implement on a large scale. These limitations open the way for DR to contribute a third option.

Demand Response (DR) DR, which is a main subject of this dissertation, is a mechanism by which energy consumers are incentivized to shift demand away from times when generation is scarce to times when generation is plentiful (Siano, 2014). As DR does not need physical infrastructure (Radovanovic et al., 2021) and can oftentimes fully recoup the costs of its own incentives, it has distinct advantages to energy storage and long-distance transmission, both of which require fairly substantial material investment. We argue here that DR can play an immediate role in easing the effect of volatile renewable energy, and over time will continue to play a complementary role (Albadi and El-Saadany, 2007). It is being considered internationally, with businesses in the US having earned over $2.2 billion in revenue through DR in 2013 and the European Commission having voiced strong support of DR (Coalition,

[12]One large and influential study, the NREL Seam Interconnect study, demonstrates that under any future scenario of grid technological makeup, connecting the Eastern and Western grids could significantly increase the ability for the grids to share generation resources and lower costs across the board. In fact, according to NREL, "expanding international transmission would provide up to $30 billion (2018 $US) of net value to the continental power system between 2020 and 2050—increasing power system reliability and enabling exchange of load and renewable generation diversity between regions." (Bloom et al., 2021).

[13]For an example of how swiftly the field is moving, at the time of writing, Honda is investing $300 million into a solid state battery plant in Togichi, Japan, and aims to introduce the batteries in EVs by 2025 (Patel, 2018).

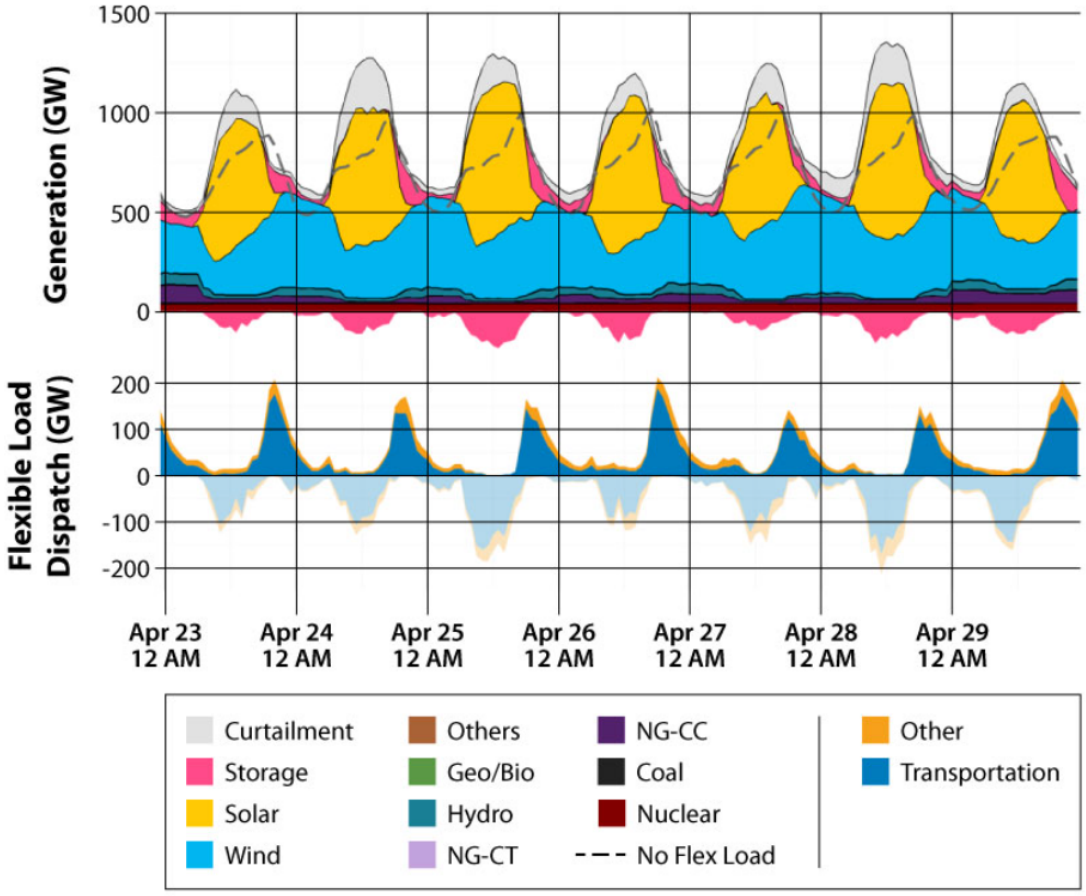


Figure 1.3: Grid generation makeup by generation type. Reproduced from NREL (Mai et al., 2018).

2014). DR in the Asia-Pacific region is rapidly growing as well, with South Korea leading the pack at 3.8 GW (Sullivan, [n. d.]). Indeed, an influential study by NREL called "Electrification Futures" (Mai et al., 2018), which was "designed to quantify potential energy, economic, and environmental impacts to the U.S. power system and broader economy", estimated the need for flexible dispatch in the future. A scenario likely to reduce a medium amount of carbon emissions, shown in Figure 1.3 shows flexible load on the same order of magnitude as the generation as it is able to offset the massive amount of non-dispatchable energy generated by wind. Most of this energy is made up of EVs.

1.1.4 What is Special about DR?

In addition to being a no-material, quickly implementable way to facilitate energy dispatchability, DR may be thought of as generation in its own right. A concept pioneered by Lawrence Berkeley National Lab's Art Rosenfeld, the "negawatt", is an amount of energy "avoided" due to energy efficiency or DR signals (Gillingham et al., 2006). It has famously been monetized as actual generation by Google Nest through its fleet of smart thermostats (Payne, 2018).

There is, of course, merit to this concept: a utility, who must contractually meet the total demand of its ratepayers, would consider an additional kilowatt (kW) avoided equivalent to an additional kilowatt generated, and pays as such.

The ability to produce a DR gives democratization over energy supply by allowing all consumers to take part in "generation". Now, any home may perform some degree of dispatchability with its appliances. The scope of this power is potentially vast (Albadi and El-Saadany, 2007).

1.1.5 Controls in Energy Systems

The way in DR is determined (i.e. how a utility quantifies the amount of DR to ask for), signalled (i.e. how the needs of the utility are transmitted to end users), and responded to (i.e. how end users react to DR signals) will necessitate a series of conscious and automatic decisions. Here, controllers are important for optimizing automatic decisions (Pal and Chaudhuri, 2006).

1.1.5.1 Controller Types, with HVAC Control as an Example

In order to implement all of these technologies - price-setting, home appliance reaction, and everything in between - control is needed. We will describe several different types of controllers, and, using taking Building Heating, Ventilation, Air-conditioning and Cooling (HVAC) as an motivating example, we will demonstrate why the development of advanced control may be necessary.

Controlling the HVAC system of a building can be, and for many decades has been, accomplished by classical control techniques. However, recent research has expanded the sophistication of controllers as well as expanded the types of systems that controllers may control. Some work focuses exclusively on window (Chen et al., 2018), others on shading Han et al. (2020), and others exclusively on HVAC setpoints (Esrafilian-Najafabadi and Haghighat, 2021). Still others attempt to unify these systems into single controllers, including the OCTOPUS controller (Ding et al., 2019a). For reviews of control research in building systems, please see (Wang and Hong, 2020).

Heating, Ventilation, Air-conditioning and Cooling (HVAC) As a note, the author had the pleasure of working on a project within Google Research called "Smart Buildings (SB)", a system for real-world control within office buildings[14], and so we will intersperse our discussion of HVAC with examples from this project. We will describe the aims and goals of the project here for the sake of demonstrating what types of building control are possible and to demonstrate that corporate entities are taking advances in building control seriously, as they can result in serious financial benefits (Sipple, 2022).

The HVAC of buildings are broken up into intricate networks of components. A central controller may propagate signals to adjust the boiler or air chiller setpoint, which then blows hot air through a series of Air Handling Units (AHUs) to Variable Air Valves (VAVs). This

[14]Unfortunately, the specific technology that the author worked on has not been cleared for publication, so we will talk about the motivation and setup for the SB project in the introduction and not later sections where the author's direct work will feature prominently.

(a) Depiction of AHU, Boiler, Meters, and Chillers.

(b) Depiction of a VAV.

Figure 1.4: Two images depicting the location of various HVAC components in a building. Images originally created by Google Research and are reproduced with permission (Sipple, 2022).

system enervates the building, and is accompanied by meters reading the temperature in different building zones. One large room may have multiple zones (VAV domains), and conversely one zone may cover several small rooms (Yun et al., 2021; Jiménez-Raboso et al., 2021). Please see Figure 1.4 for a depiction of HVAC components on a building.

Simple Control Controlling the HVAC system of a building can be, and for many decades has been, accomplished by classical control techniques, if our design goal is to keep temperatures somewhat close to a setpoint. Two simple controller types have proven to achieve this goal for HVAC systems: threshold-based "bang-bang"controllers, and Proportional Integral Derivative (PID) controllers.

Bang-bang controllers are ones that exert binary control over a system: they turn on when they hit a low setpoint, and off when the hit a high setpoint. In other words, they applies the greatest allowable heating or cooling when the measured temperature leaves a

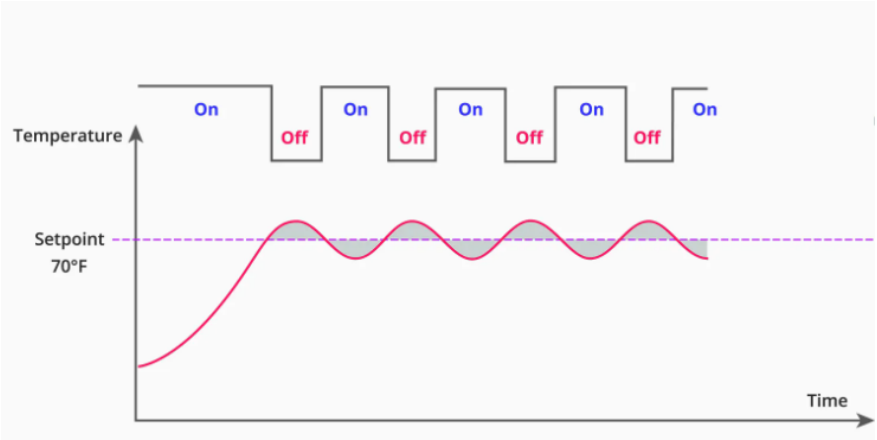


Figure 1.5: An idealized Bang-bang controller in action. Image from (Mortenson, 2022).

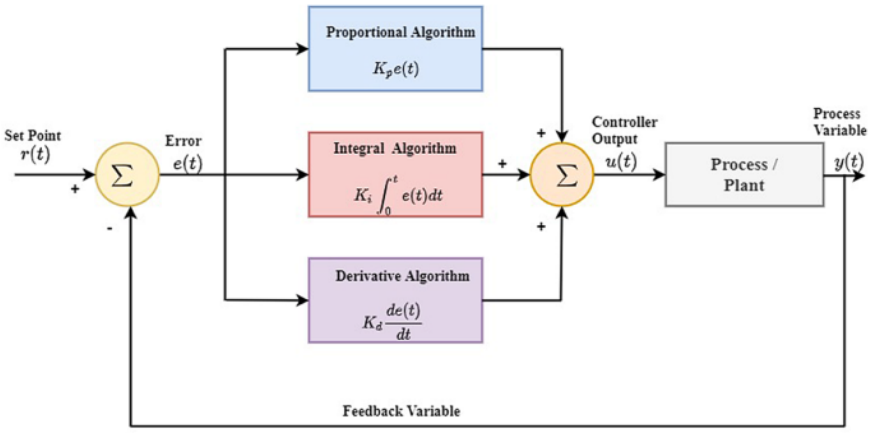


Figure 1.6: A schematic of the control inputs in a PID controller. Image from (Borase et al., 2021).

certain band around the target temperature, but otherwise remains inactive (Seyde et al., 2021). Please see Figure 1.5.

Proportional Integral Derivative PID controllers use a linear feedback strategy, which decides a control action using a linear function of previous measurements. PID denotes a sum of Proportional, Integral, and Derivative functions formed as the sum of the current measurement (the proportional component) and approximations of the time derivative and integral of the temperature (Borase et al., 2021; Geng and Geary, 1993).

These simple control policies even achieve optimality in certain cases and for certain metrics: for example, the bang-bang strategies are optimal for minimizing the total time that the HVAC system is active, and PID controllers can be tuned to minimize the sensitivity of the controller to exogenous disturbances (Seyde et al., 2021). However, if we ask for a strategy that is optimal with respect to delivering thermal comfort, carbon emissions, or even a combination of several objectives, to act predictively of energy load profiles throughout the

day, to take into account occupancy when deciding on HVAC setpoints, or other desiderata of intelligent controls, we approach the limit of what is possible for simple controllers. What may be considered the output of the system is also the only input that it can consider: the output of the HVAC system being temperature means that this is the only input that it considers, and even modern Building Management Systems (BMSs) do not feed descriptive input like occupancy into controllers. Given this lack of info, bang-bang or PID controllers cannot form rich functions combining different metrics, and thus act only to immediately adjust temperature above or below manually set setpoints Kasahara et al. (1999). In this way, the controller is not only unidimensional in terms of objective but also purely memoryless.

If we wish to account simultaneously for the many input-output subsystems that the building comprises, or if we wish to take detailed account of human behavior in the feedback loop, we leave the frontier well behind us. Simple control cannot consider multiple objectives, or different objectives than the direct input/output. It cannot consider a state space richer than a single variable. It also cannot incorporate or plan for future desiderata. In the case of HVAC, this means that it cannot incorporate occupancy as a state space consideration nor occupancy prediction as a way to plan actions. It cannot act to reduce energy consumption as well as carbon emissions, and it cannot plan for a 9am setpoint change by pre-cooling or pre-warming the space (Attaran et al., 2014). To find controllers that satisfy these new objectives, we must turn to modern BMS frameworks that collect, synthesize, and communicate data called Building Operating Systems (BOSs). These allow for more complex methods in control theory and RL. For systems that we can model, even if the model is only partly specified and possibly very complicated, modern control theory can synthesize safe and optimal controllers; when the system cannot reasonably modeled, such as where human decisions play a direct role, we shall find that RL can develop very effective controllers.

Model Predictive Control (MPC) MPC takes a large step towards allowing for multiple objectives and predictive control. MPC improves these aims by incorporating optimization directly into the control policy. MPC relies on an ancillary dynamics model, and solves a prediction problem T steps into the future at each step. Generally, if the dynamics model is f, the action a, the state s, the objective g, and the current timestep t, then MPC solves:

$$\begin{aligned} \underset{a}{\text{maximize}} & \sum_{t=t'}^{t'+T} g(s,a) \\ \text{subject to: } & s(t+1) = f(s(t), a(t)); t = t', \ldots, T-1 \\ & s(t'+T) = 0; t = t', \ldots, T-1 \\ & s(t) \in S; t = t', \ldots, T-1 \\ & a(t) \in A; t = t', \ldots, T-1 \end{aligned} \tag{1.1}$$

The condition $s(t) \in S$ is a state space constraint, which can be used to ensure safety by attempting to keep the state vector in a region of the state space that is known *a priori* not to be dangerous.

When considering a linear dynamics model, f becomes:

$$s(t+1) = Bs(t) + Ca(t) \tag{1.2}$$

i.e., series of states that are modelled in such a way to be linearly related (via matrices B and C) to the prior states (Afram and Janabi-Sharifi, 2014).

Limitations with MPC The state space S can be multi-model and expressive relative to the bang-bang controller, but practitioners may find that the state space of RL models can be far larger, containing hundreds to thousands of variables. MPC also cannot handle complex datatypes such as images, or complex and long-term time series data which policy networks of RL can digest and use effectively through convolutional or Long Short Term Memory network (LSTM) embeddings, respectively.

MPC is fast, computationally efficient, and crucially, it lends itself to safety critical applications better than RL would, as it constrains itself to a narrower, known state space. There are times, however, when the safety constraints of the state space are either not important or under-determined. For instance, if assumptions of linearity in state transition are incorrect, or if there are important exogenous variables not measured (e.g. occupancy metrics, behavioral preferences, or even sensor dysfunction), a linear model may be a very poor fit to the data and an MPC may optimize for the wrong predictions. Using a Neural Net (NN) to fit to the dynamics function can go a decent way towards approximating non-linear or complex dynamics, but may require significantly more data (Afram and Janabi-Sharifi, 2014).

In this dissertation, we lean whole-heartedly into RL in order to accomplish these goals. We will define RL technically and thoroughly.

1.2 Reinforcement Learning

1.2.1 Markov Decision Processes

Markov Decision Processes (MDPs) are discrete-time stochastic control processes defined by the tuple:

$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{P}) \tag{1.3}$$

Where $\mathcal{S}$ are a set of *states*, $\mathcal{A}$ a set of *actions*, $\mathcal{R} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ as a *reward* signal[15], and $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ as a set of *state transition probabilities* for each state, action, and subsequent state an agent[16] may attempt to reach (or not reach.) They provide a generic modeling framework that underpins RL (Puterman, 2014; Sutton and Barto, 2018).

For example, a controller might take a discrete-valued action $a_t \in \mathbb{Z}^{|\mathcal{A}|}$ in an MDP for T days indexed by $t = 1, 2, \ldots, T$. The discrete action may be: "0: turn on lights, 1: turn off lights"[17] The pertinent data available at a period t to make decisions would be represented using a state vector s_t. These data can be temperature readings (Sipple, 2022), image

[15]Please note that some reward functions may not factor in the next state, and so may be more succinctly defined as $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$.

[16]The terms "agent", "controller", and "policy" are used relatively interchangeably in the various fields that touch the general application space of controls. All indicate that there is some automatic decision maker who acts in an environment which gives it some feedback.

[17]If the action space is continuous, it would be: $a_t \in \mathbb{R}^{|\mathcal{A}|}$, and an example of the action may be a price.

embeddings (Bandara et al., 2022), audio waves (Latif et al., 2022); anything that can be quantified. Defining an agent's state space is often done by a human designer and thus are a possible source of bias.

The MDP cannot be solved directly since the state transition probabilities $P(s_t, a_t)$ are unknown *a priori*. RL is useful to address this issue by implicitly estimating the state transition probabilities. RL can also encode extensions of MDPs (Partially Observed Markov Decision Processes (POMDPs), (Krishnamurthy, 2016), multi-step MDPs, i.e. semi-MDPs (Sutton et al., 1999b).)

1.2.2 RL Definitions

An RL controller's actions seek to optimize the "objective" J (notation defined fully in the paragraph below), a discounted expected sum of rewards r for actions (a_t) and states (s_t), i.e.,

$$J(\theta) = \mathbb{E} \sum_{s_t, a_t \sim \tau_{\pi_\theta}} [\gamma^t r(s_t, a_t)] \tag{1.4}$$

Where $\gamma :\rightarrow [0, 1]$ is some discount factor privileging more recent actions, and θ are the parameters that define a *policy*,[18] denoted $\pi(\theta)$ or π_θ. $\pi(\theta) : \mathbb{R}^{\mathcal{S}} \rightarrow \mathbb{R}^{\mathcal{A}}$ is a surjective function mapping states to actions, and π_θ is shorthand for a policy parameterized by θ. With a fixed policy, one can collect a *rollout* or *trajectory* of data relative to that policy $\tau_\pi(\theta)$. RL training loops change θ in order to increase the performance of its policy. An optimal policy is defined as one that always chooses the best possible action, and is denoted π^*.

An additional metric, the value function $V(s)$ may be thought of as an instantiation of the objective given a certain state. I.e., we can create and sum a trajectory of states and actions (thus summing the discounted rewards) from each state given a policy:

$$V^\pi(s_t) = \sum_{t'=t}^{T} \mathbb{E}_{\pi_\theta}[\gamma^{t'-t} r(s_{t'}, a_{t'}) | s_t] \tag{1.5}$$

We define here also an optimal Q-function, Q^* which may be thought of as an instantiation of the objective given both a state and an action, and functions as a useful "what-if" for possible actions:

$$Q_r^*(s_t, a_t) = r(s_t, a_t) + \gamma \mathbb{E}_{P(s_{t+1}|s_t, a_t)} \arg\max_{a_{t+1}} Q^*(s_{t+1}, a_{t+1}) \tag{1.6}$$

Where $P(s_{t+1}|s_t, a_t)$ is a specific probability. The optimal policy and the optimal Q-function are related according to $\pi_\theta^*(s) = \arg\max_a Q^*(s, a)$. Please see Figure 1.7 for a simplified pathway of RL actions.

[18]Here, the parameters θ can refer to any parameters that make up the policy function, and it is the tendency of the field to try to be agnostic to which model this is. Thus, they could be linear model weights, random forest parameters, or Support Vector Machine (SVM) parameters, but they are almost invariably neural net weights and biases. RL is dominated by agent architectures composed exclusively of deep NNs (Mahesh, 2020)

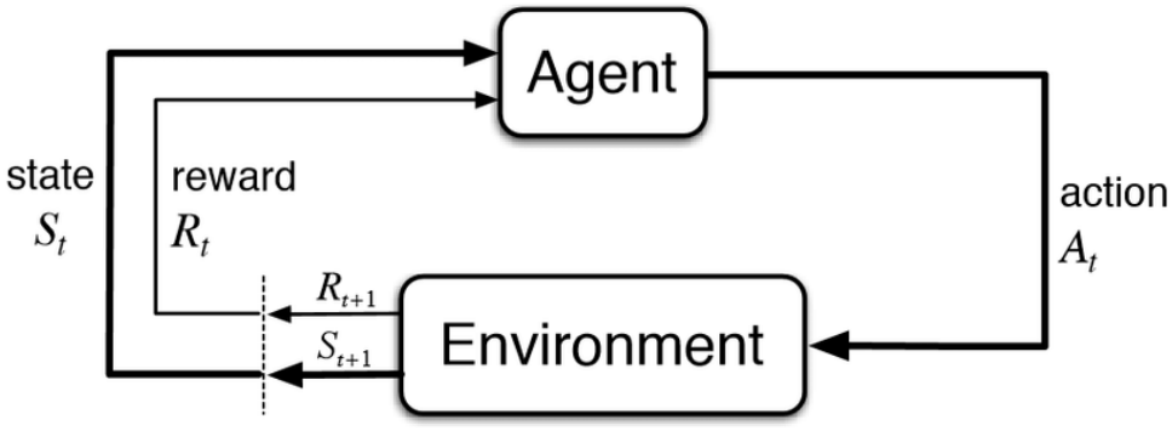


Figure 1.7: A simplified diagram of a generic RL pipeline.

In most cases, an RL agent is comprised of a combination of neural nets that help the agent perform various tasks such as helping it both predict long-term effects of its actions as well as choose the correct subsequent actions. As a brief aside, the structure and interaction of these neural nets can vary widely, with groups using them for such functions as embedding inputs, modeling other agents, and predicting long term reward. Please see Figure 1.8 for Richard Sutton's[XXI], generally considered a grandfather of the field, depiction of the strictly necessary elements of an agent that will aspire for general intelligence (Sutton, 2022). RL has classically been applied to contexts where actions and environments are simple or data is plentiful, with early use cases involving the control of backgammon (Tesauro, 1994), the cart-pole problem, and Atari (Mnih et al., 2013). Much recent work has been done to extend RL to other situations where data is less plentiful, rewards are more sparse, and required actions are more complex using Surprise Minimizing Reinforcement Learning (SMiRL) (Arnold et al., 2021b) and offline RL (Jang et al., 2021c). When we speak about "data" in the context of RL, we are generally referring to tuples of (s_t, a_t, s_{t+1}, r_t), as tuples of this data are sufficient to calculate the prior metrics (Equations 1.5, 1.6, 1.4) Of course, calculations of the metrics will change as more data comes in.

1.2.3 Gradients: How Models Train

It is useful to briefly touch on a guiding technique used throughout RL: the creation and propagation of *gradients*. Gradients are defined as the derivative of a multi-dimensional function with respect to each dimension. If $f(x) : x \in \mathbb{R}^n \to \mathbb{R}$, then $\nabla_x f : \mathbb{R}^n \to \mathbb{R}^n$ and is defined by:

$$\nabla_x f = \left[\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, ..., \frac{\partial f}{\partial x_n}\right]^\intercal \tag{1.7}$$

A gradient may more intuitively thought of as multi-directional direction of change in some manifold (an extension of high-school intuition of a derivative.) In ML, the shape of the manifold of interest is defined by functions that relate the model's current predictions

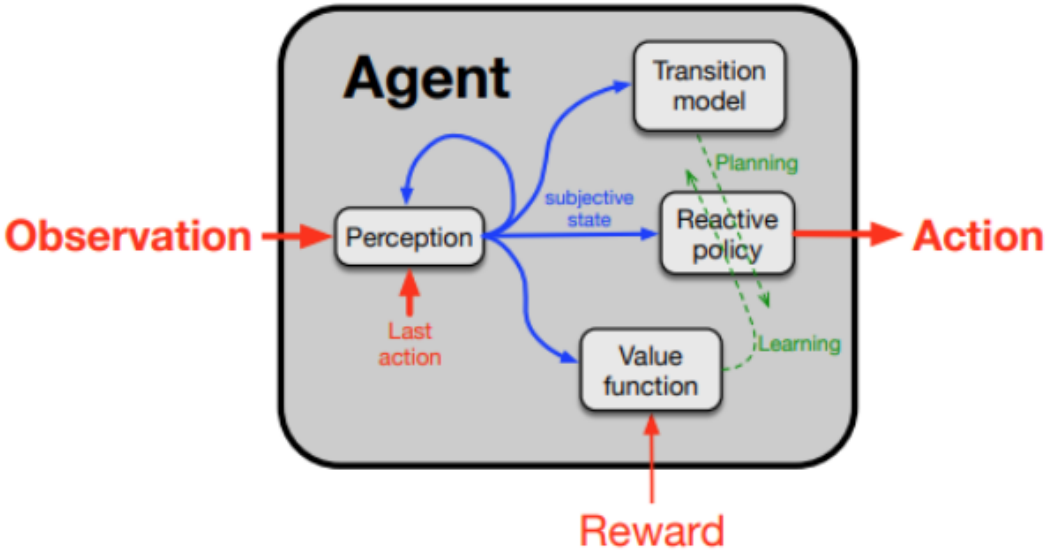


Figure 1.8: Sutton's depiction of a general intelligence agent (Sutton, 2022).

to observable data, such as the *loss* in supervised learning or Q or V predictions in RL, or the policy surface more directly. We will look at one common loss as it's simplest, the Mean-Squared Error Loss, which is defined

$$L_{MSE}(y, \phi) = \frac{1}{2}\mathbb{E}(y - f_\phi)^2 = \frac{1}{2n}\left[\sum_{i=1}^{n}(y_i - f_{\phi,i})\right]^2 \tag{1.8}$$

using $f_{\phi,i}$ as a shorthand for the ϕ parameterized model's prediction at dim i, $\hat{y}_i$. It changes based on each batch of data (the true y_i observed will be different) and the gradient is a vector of the following:

$$\begin{aligned}\nabla_\phi L_{MSE} &= \left[\frac{\partial L_{MSE}}{\partial \phi_1}, ..., \frac{\partial L_{MSE}}{\partial \phi_m}\right]^\intercal \\ &= \left[\frac{1}{n}\sum_{i=1}^{n}(y_i - f_{\phi,i})\frac{\partial f_{\phi,i}}{\partial \phi_1}, ..., \frac{1}{n}\sum_{i=1}^{n}(y_i - f_{\phi,i})\frac{\partial f_{\phi,i}}{\partial \phi_n}\right]^\intercal\end{aligned} \tag{1.9}$$

In sum[19], parametric ML models like neural nets are all trained with high-dimensional manifolds in the space of the parameters. Computing the gradient of these manifolds forms the basis of *model training*, whose goal is to adjust the parameters so that their predictions may consistently achieve a better loss or objective. Much of training consists of determining the appropriate direction in the parameter space to move all parameters in. The direction chosen is the gradient, and distance the parameter vector travels along this direction is often the *learning rate* (although there are much more sophisticated ways to determine distance –

[19]No pun intended.

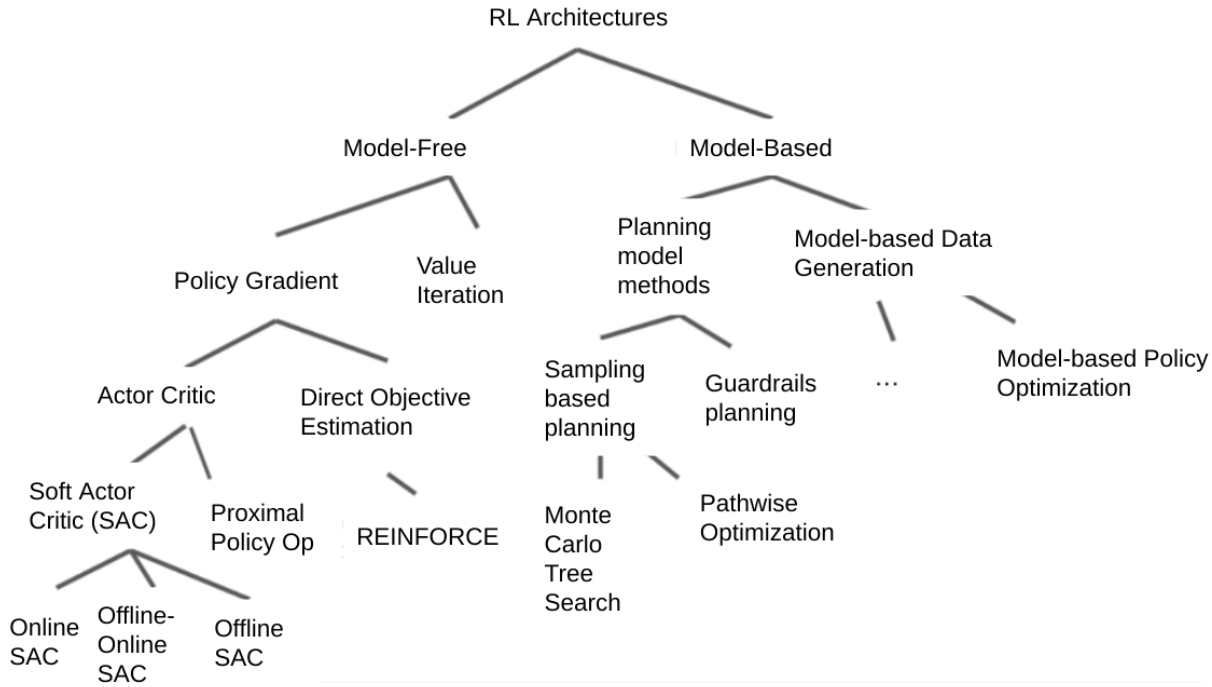


Figure 1.9: A taxonomy of prominent RL architectures.

please see momentum (Ding et al., 2019b) or other adaptive methods (Wang et al., 2019b) for more information.)

1.2.4 Common Architectures Used

We present a brief overview of common RL architectures. Please see Figure 1.9 which, tailored to our group's needs, represents a specific understanding of the important groupings in RL taxonomy. In this figure, an ellipses ("...") represents many additional architectures under each grouping that we note. We will now briefly discuss two Policy Gradients, Actor-Critic, Soft Actor-Critic (SAC), and Proximal Policy Optimization (PPO) methods, common architectures that we use throughout the paper, but all other methods listed can be found in (Sutton and Barto, 2018) save Pathwise Optimization (Yang et al., 2021).

1.2.4.1 Policy Gradient and PPO

General Definition At a very broad level, policy gradient methods seek to nudge the policy along estimates of $\nabla_\theta J(\theta)$ computed from data collected at each step. Typically, a policy would be set while it gathers data from a trajectory of length T, and then the objective computed as above. The parameters are updated at the end of the trajectory as per the following:

$$\theta_{T+1} = \theta_T + \alpha_T \nabla_{\theta_T} \hat{J}(\theta_T) \tag{1.10}$$

However, the ways in which policy gradient algorithms compute the objective differ widely, and thus there is a rich and varied literature in different types of policy gradients. Often, estimates of $\nabla_\theta J(\theta)$ are accomplished through different weighting of the gradient of the policy's action probabilities per state. One example estimates the objective's gradient by weighting action probabilities by the Q estimate, i.e.:

$$\nabla_\theta \hat{J}(\theta) \simeq \mathbb{E}_\pi \sum_a q_\pi(a_t, s_t) \nabla_\theta \pi(a_t|s_t) \tag{1.11}$$

Which forms the basis behind the REward Increment = Nonnegative Factor x Offset Reinforcement x Characteristic Eligibility (REINFORCE) algorithm (Williams, 1992). Intuitively, this weighting steps along this direction steer the parameter set towards action-probability surfaces that place higher probability on actions that produce higher Q values.

1.2.4.2 Actor-Critic Architectures

General Definition : Actor-critic architectures have long been a feature of RL (Sutton and Barto, 2018). Here, one network acts as the policy, and another estimates the value V of a given state. The policy acts directly in the world, and so at each step the policy computes an action and the critic computes what may be thought of as a separate estimate of the longer-term reward (Sutton and Barto, 2018). The critic's evaluation is the temporal difference error:

$$\delta_t = r_{t+1} + \gamma V(s_t + 1) - V(s_t) \tag{1.12}$$

Where V is estimated from the critic (and thus changes at each step as both networks learn.) The actor's parameters are changed at each step by input from the critic:

$$\theta(s_t, a_t) \leftarrow \theta(s_t, a_t) + \alpha \delta_t \tag{1.13}$$

Where α is some hyperparameter defining step-size. Temporal difference learning is among the simplest ways of updating the critic (Sutton and Barto, 2018), and other methods exist.

Please see Figure 1.10 for a graphical depiction.

Proximal Policy Optimization (PPO) PPO methods vary slightly, but all seek to constrain the actor's gradient step made by specifying some region around the objective space beyond which the training cannot step. This generally leads to more stable learning. The clip method we use modifies the objective function in the following way:

$$J^{\text{CLIP}}(\theta) = \mathbb{E}\left[\min\left[\nu_t(\theta)\mathbb{A}_t(a,s),\ \text{clip}(\nu_t(\theta),\ 1-\epsilon,\ 1+\epsilon)\mathbb{A}_t(a,s)\right]\right] \tag{1.14}$$

Where ϵ is a hyperparameter, usually $\in (0, .2]$, $\mathbb{A}_t(a,s)$ is the estimated "advantage" function (Baird, 1994), defined as the difference in Q values between the action chosen and the average action, i.e.

$$\mathbb{A}(a,s) = Q(a,s) - V(s) \tag{1.15}$$

and ν_t is the ratio of probability under the new and old policies, i.e.

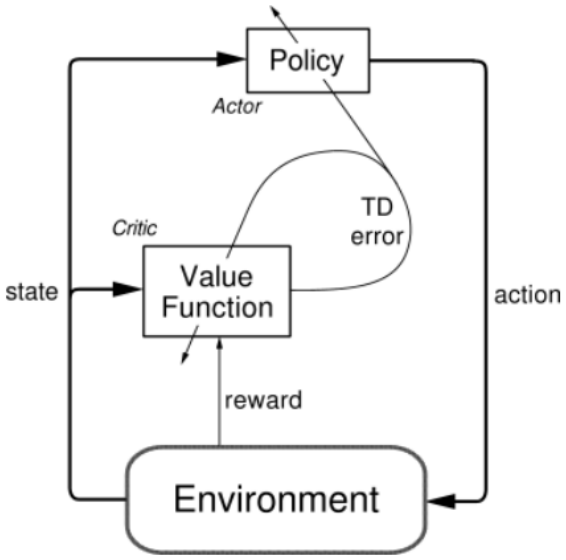


Figure 1.10: Actor Critic architectures (Sutton et al., 1999a)

$$\nu_t = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)} \tag{1.16}$$

and "clip" is a function taking three arguments, (x, y, z) and if necessary clips x such that $x \in [y, z]$.

We use PPO extensively throughout our work. For more info, please see (Schulman et al., 2017).

Soft Actor-Critic (SAC) SAC (Haarnoja et al., 2018), performs well out of the box on a range of tasks. It adds an entropy term to the objective function:

$$J(\theta) = \sum_{t=0}^{T} \mathbb{E}_{(s_t,a_t)\sim\tau_{\pi_\theta}}[r(s_t, a_t) + \alpha\mathcal{H}(\pi(.|s_t))] \tag{1.17}$$

Where $\mathcal{H}$ is the entropy calculation and α is a weighting that changes per step as the model becomes more confident in its predictions. It also modifies the value function:

$$V(s_t) = \mathbb{E}[Q(s_t, a_t) - \log(\pi(a_t|s_t))] \tag{1.18}$$

Which ensures that the Value is never 0. These modifications bias the model to choose actions that have higher entropy given the same reward, leading it to explore more effectively. We use SAC extensively throughout this work.

1.2.5 Model Based RL

Model-based RL is arguably most common in energy planning applications. Here useful functions like estimates of state transition dynamics are learned from offline data before optimizing the policy. Such auxiliary models are referred to as a "planning model" in RL because queries of the real world can instead be replaced in the MDP by queries of a fitted model. The pricing policy obtained by solving an MDP supervised by a planning model will depend on how well the planning model reflects the true environment.

Model-free RL forgoes a planning model by learning directly from the true environment and optimizing the policy. A central problem in this approach is balancing the trade-off between "exploration" and "exploitation". "Exploration" may be thought of as emitting actions for the purpose of understanding more of the environment's dynamics, *not* to achieve the lowest immediate cost. A simple exploration strategy could be to sample randomly in a neighborhood around the RL agent's chosen price. "Exploitation", may be thought of as emitting actions for the purpose of optimizing short or mid term cost. However, unlike model-based RL, there is no specific planning model. Rather, a stochastic policy implicitly encodes knowledge of the state space.

1.2.6 Case Study: Google's SB Project RL Results

We will know explain how Google's SB HVAC agent is structured so as to give an example of how the RL theory described above may be applied.

The SB RL controller's state space S is defined as a tuple of intake air flow, temperature, and CO_2 unique to each building's sensors, and action space A is a tuple of water and air supply setpoint temperatures. Please see Figure 1.11 for a graphic representing the flow of information in the SB MDP. The software design was modular in that different RL agents could be tested in addition to different environments (shown in Figure 1.11 is the real building, a (data-driven) emulation of the building, and a physical simulation of the building), different rewards could be tested, and different RL architectures (Sipple, 2022).

The reward structure was called the "three C's" reward, and was a weighted sum of carbon emissions, cost, and comfort[XXII]:

$$R_{\text{SB}} = w_1 * \text{carbon} + w_2 * \text{cost} + f(t) \tag{1.19}$$

Where w_i are weights and comfort is represented by $f(t)$, some negative quadratic weighting on a temperature regime outside the range of average comfort. Weighting was determined manually[XXIII] based on simulation behavior.

The SB agent produced, in simulation, a strong outcome when compared to a baseline PID controller. After training in simulation, the agent is able to measurably improve functioning in the building's HVAC. Please see Figure 1.12 and Figure 1.13 for a demonstration of the building functioning and aggregate metrics after convergence. Preliminary results show that the agent is able to learn and improve efficiency of buildings in simulation (Sipple, 2022).

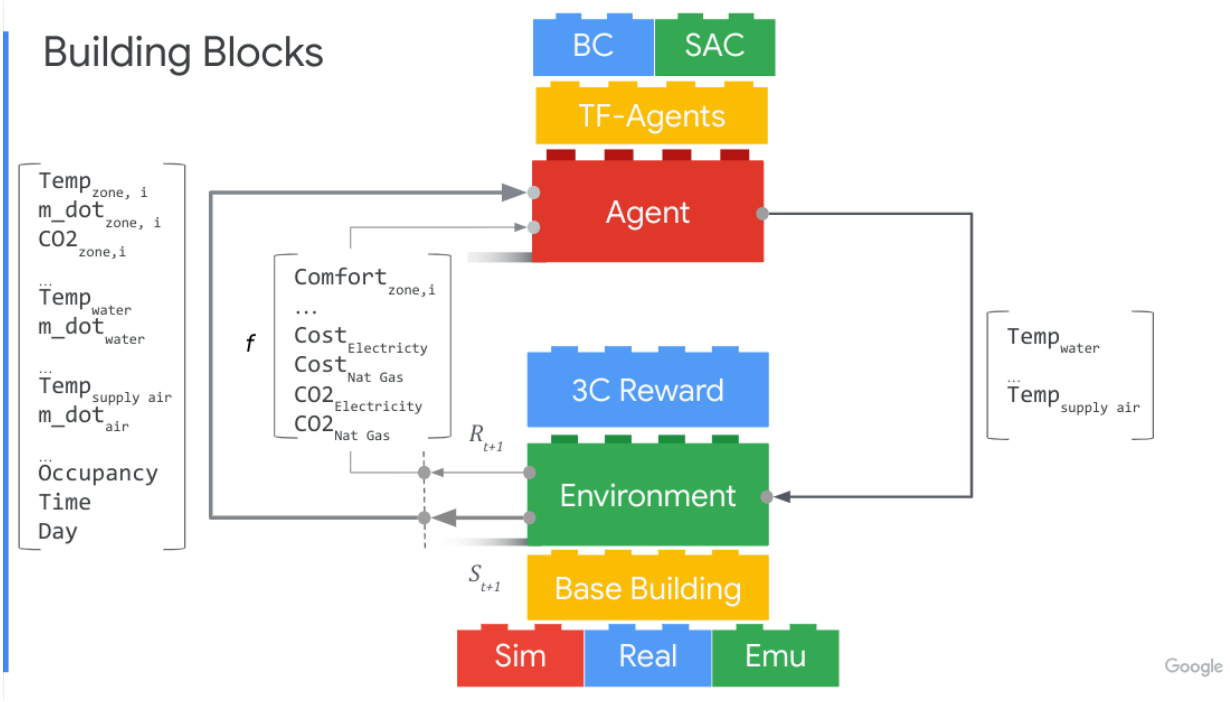


Figure 1.11: SB RL flow. Image reproduced with permission. (Sipple, 2022).

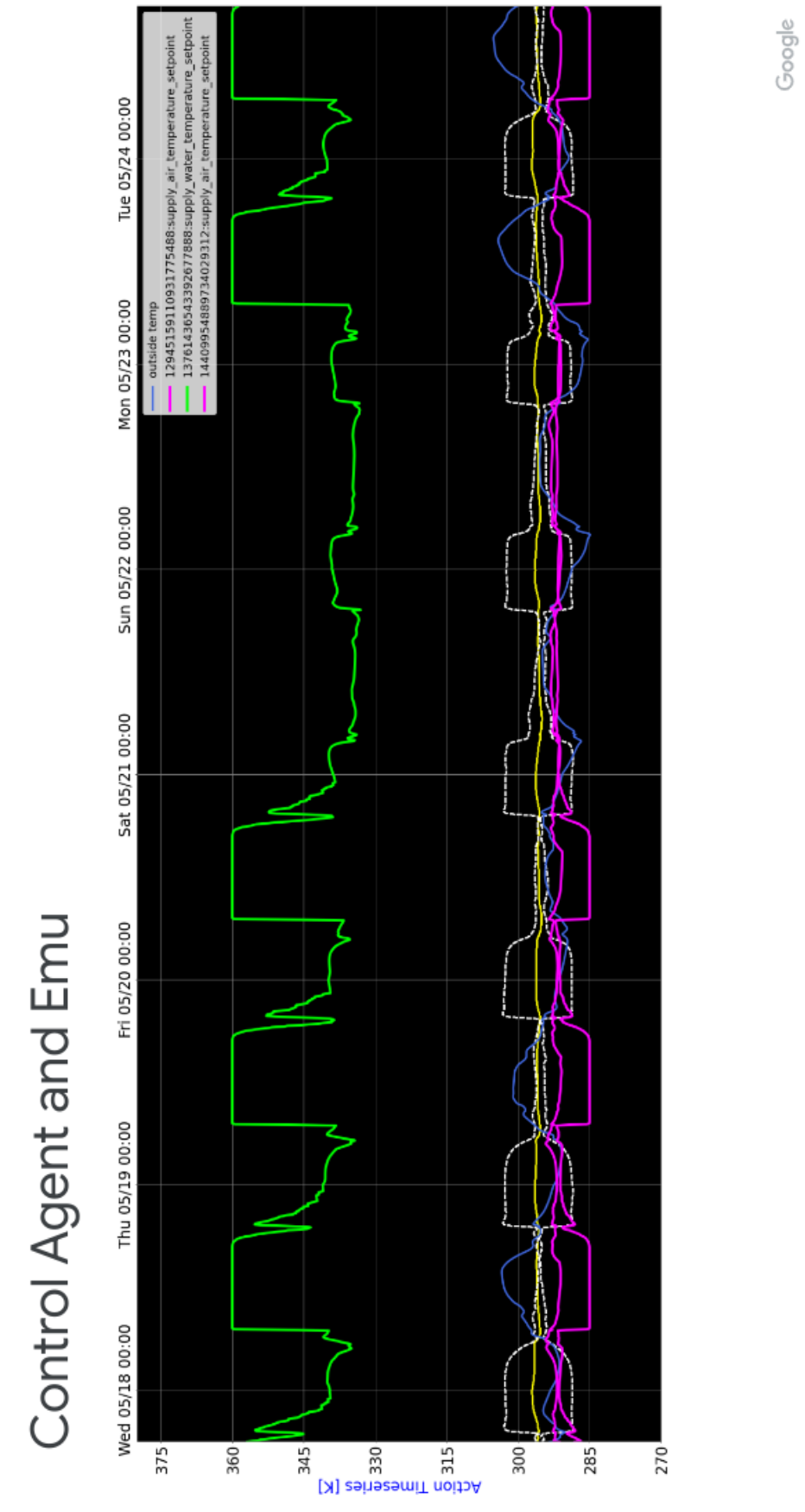


Figure 1.12: Example of operating conditions of the RL agent in the simulation., (Sipple, 2022)

The SB project is one among many examples of increasing sophistication of controls within buildings. Such a controller may in the future receive an input of grid prices during the day and coordinate setpoints without a human interfering in the process. The relative beauty of this example is that advances may occur independently of each other: the SB's controller optimizes a building's cost and carbon under any price regime that is handed to it, including the current flat price regime, thus it is financially prudent for an entity like Google to develop now (Sipple, 2022). However, a utility-scale dynamic price scheme may be developed sometime in the future and integrate smoothly with the same controller architecture.

1.3 A Roadmap to this Dissertation

1.3.1 Aims of this Dissertation

Our goals in this dissertation are multifold.

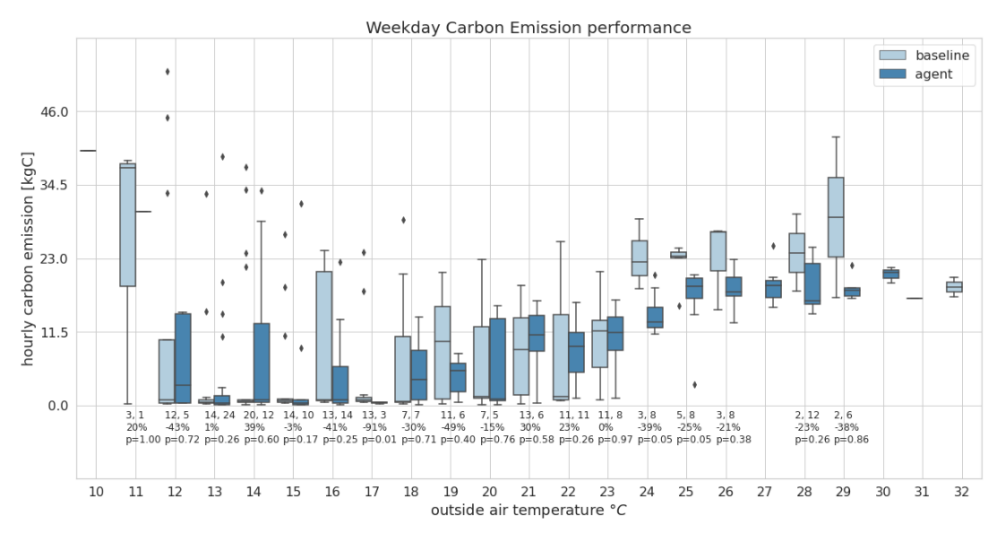
Two New RL Benchmarks We wish to surface and highlight two new RL benchmarks in energy DR: OfficeLearn and MicrogridLearn, discussed in Chapter 2. Both are *transactive control settings where agents set prices in order to influence behavioral response.* These are two unique and interesting environments that we use to test RL agents. We are the first to propose the use of RL in price-setting of this nature.

Articulation of Problems in Sim-to-Real RL We explore six problems in Simulation-to-Reality (Sim-to-Real) RL: data inefficiency, lack of robustness and safety, suboptimal training in simulation, potential for adversarial attacks, privacy preservation and generalization to new subdomains, and hyperparameter optimization. We hope that the articulation of these problems is interesting and compelling, and may lead to future work.

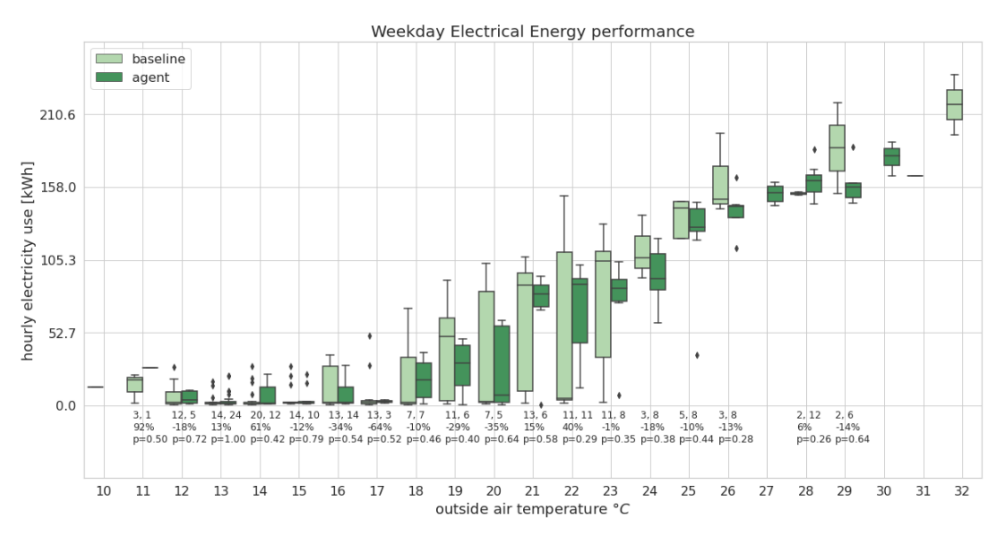
Methodological Innovations Of the many techniques that we investigate to address the problems above, we produce new and interesting methods in the space of RL. We wish to highlight the *guardrails* approach in planning models (Jang et al., 2021b), application of Personal Federated Hypernetwork (PFH) to RL (Jang et al., 2022b), and adversarial data poisoning (Gunn et al., [n. d.]) as novel advancements in RL Theory.

Methodological Recommendations Of the methods we try, we wish to recommend methods that were especially helpful to our work. Specifically, we recommend *guardrails*, Domain Randomization with Meta RL (Jang et al., 2021a), and subsequent study in Auto-curricula.

Education and Awareness We see this dissertation as one that bridges the gaps between two different communities: energy and Artificial Intelligence (AI). We wish to educate readers from the energy community about aspects of AI and RL that we find important. Similarly, we wish to educate members of the AI community on aspects of energy that we find most compelling and interesting.



(a) Carbon outputs in the simulation under the agent's policy versus baseline.



(b) Energy consumption in the simulation under the agent's policy versus baseline.

Figure 1.13: Results from the agent running in simulation versus the baseline. Image reproduced with permission. (Sipple, 2022).

1.3.2 Advice to Different Readers

Our goals in DR is to break ground on a technique we believe may strongly impact the field. We wish to challenge the assumptions that price-setting may only happen from top-down control, and suggest that price-setters can automatically incorporate observation and feedback in a way to optimize the prices set. If the reader is approaching this dissertation hoping to learn more about DR and transactive control, they may prefer to focus on Sections 1.1, 2.1, 2.2, and 2.3.

Our goals in RL are to provide a recommendation of, out of the many techniques that we have tried, a few that may be particularly helpful in guiding a transition between simulation and reality. They are clustered around specific problem types that come up in the area. If you are reading this for the purpose of understanding RL methods, I recommend that you glance over Chapter 2, giving some attention to Sections 2.2 (MicrogridLearn (Agwan et al., 2021a)) and the first parts of 2.3 (OfficeLearn (Spangher et al., [n. d.])) to understand how our environments are setup. You may choose to skip 2.4 and 2.5 entirely. You may then focus on whichever issue with sim-to-real RL that is most pertinent to your problem at hand. I strongly recommend that all RL readers read our proposed *guardrails* method and our proposed use of auto-curricula to guide HVAC development.

Chapter 2

Demand Response Problems as Reinforcement Learning Environments

2.1 Introduction

2.1.1 The Price Setting Problem

Consider a system composed minimally of price-respondent entities and a price-setter[20]. The price setting entity may set a different price $\vec{p} : [p_1, ..., p_H]$ for each hour of the day which is functionally composed of H hours. The price respondent has some energy DR function $d(p, b) : b \in \mathbb{R}^H \rightarrow \mathbb{R}^H$; i.e. the energy demand of one hour is dependent on not only prices of that hour but all other hours as well. The goal of the price-setter then is to pick the prices that optimize some objective T that it cares about:

$$\textbf{argmax}_{\vec{p}} \; g(d(\vec{p})) \tag{2.1}$$

Features of the environment influence the price-setter's actions, including whether there exists another price-setter, whether price-respondent entities are *prosumers* in that they both buy and sell a resource according to prices or just *consumers* who buy the resource, and how many price-respondent entities exist.

2.1.2 The Price-Setting Problem in Demand Response

2.1.2.1 Instantiating the Price-Setting Problem

When we instantiate the price-setting problem in DR, we assume that prices are an important way to signal the utilities' DR needs. The utility, or some *aggregator* below the utility, which may be thought of as a localized market clearing-house for energy, is thus the price-setter. The price-respondent entities are consumers of energy, be they abstract (microgrids) or individual (i.e. users within a building.)

[20] We extend this setting to include other entities as well.

2.1.2.2 Why Focus on Prices?

As previously discussed, DR has the potential to both democratize energy "production" through the concept of the negawatt, and coordinate a massive fleet of end appliances. However, harnessing the aggregate amount of DR requires a flexible way to signal to consumers. A complication is that the signal that may need to be individualized for localities or individuals: it must encourage shifting while not penalizing energy use when it is necessary. Such a signal must learn over time, as appliance bases and communities change. Individuated energy prices are thus a good signal for DR.

2.1.2.3 Current Practices

Currently, utilities do not signal to residential consumers at all: they pay a flat rate for energy consumed at any point of the day. On the vanguard of the movement is so-called Time of Use (TOU) pricing, which communicates two prices throughout the day. For example, Oakland City charges consumers one higher price from 4-9pm and a lower price for the rest of the day. However, this is clearly inadequate when considering that a grid's energy makeup (and cost to the utility) may change dramatically across both of those time intervals.

We argue that flexible prices throughout the day are a natural way to signal to consumers the value of energy consumed throughout the day. Under many utility jurisdictions, "net-metering" rules apply to energy produced, wherein consumers may sell net metering for the same price that they would buy it. Flexible prices combined with net metering truly democratizes energy generation and consumption by allowing consumers to decide when they may want to sell avoided their energy and when they may want to buy energy.

2.1.3 Common RL Environments: OpenAI Gym

As RL implementations become more standardized, it is important for the environments they operate in to adopt a common structure, or API. OpenAI, a non-profit very involved with the proliferation of RL, was one of the first to put out a common environment Application Programming Interface (API) called OpenAI Gym. OpenAI Gym environments are a series of standardized environments that provide platform for benchmarking the progress of learning agents. (Brockman et al., 2016). OpenAI Gym environments allow researchers to create custom environments that immediately allows deployment of a suite of out-of-the-box RL techniques.

We have coaxed the price-setting problem into one that is natural for RL: there is a regular time-period, actions (prices), and unknown environmental response with clear rewards. We will thus describe the software basis for implementation.

Popular, well-made, or particularly realistic Gym environments tend to concentrate focus and energy around a specific problem that they describe. We thus describe three different Gym environments that we have created and put into the open-source repo, hoping to lead work in the area.

2.2 Within a Microgrid

2.2.1 Background and Motivation

Many researchers have studied how to coordinate the energy of several buildings or agents with relative success: (Johnson et al., 2015) examines the control strategies of load deferral, thermostat adjustment, and load shedding for multiple simulated buildings, (Ma et al., 2015) explores a cooperative demand response scheme for multiple industrial refrigerated warehouses. We describe these works as fitting in some level in between individual building level or microgird level; we may imagine that one company that owns a few buildings within a larger microgrid may have reason to coordinate their buildings under the one agent's hood. These works have generally examined how to control appliances or buildings *in response* to some price or other supply signal.

Why control pricing in microgrids? Traditional consumers like buildings are increasingly investing in distributed energy resources such as solar panels and battery backups, and electrifying loads like vehicles. These resources can be used to supply the building's own demand, shave the peak load to reduce demand charges, or to increase resiliency in the face of grid failure or power shutoff and enable consumers to become *prosumers*, i.e. be electricity *producers* as well. However, such resources may remain underutilized, or be sized over-capacity to account for weather and load variability. Prosumers can profit from trading their surplus energy with other prosumers and improving resource utilization. Utilities have typically accommodated Distributed Energy Resources (DERs) through net-metering programs that compensate producers at the retail tariff. However, retail tariffs are much higher than wholesale energy market prices, and utilities have begun to phase out net metering programs in favor of direct market participation of DERs through aggregations. Prosumers can increase their profitability if they trade energy with other prosumers using social net-metering schemes implemented by utilities (Henriquez-Auba et al., 2018). In these schemes, communities can share energy resources while presenting the net consumption to the utility as a single entity.

Recent regulations have opened up multiple avenues for DERs to participate in energy markets. However, wholesale markets have minimum participation sizes and may require DERs to construct demand and supply bids. Virtual microgrids and DER aggregations offer a pathway for prosumers to trade energy locally instead of participating in energy markets.

2.2.2 Prosumer Aggregations

In this section, we imagine prosumer aggregations that facilitate trading between participants in the aggregation, and then balance the net load by purchasing from or selling to the utility. They can be formed with a variety of motives: private entities could manage aggregations for a fee or for a profit, and participants could form cooperative aggregations to maximize social welfare. Each aggregation has control the energy consumption and production of its participants, either through direct control or through some signal which can convey operational information to the participants. Realistically, prosumers will have independent cost minimization objectives, and will seek to optimize the operation of their resources for their own profit. Coordinating independent entities which are separately owned and managed is a difficult task, and *transactive control* is a strategy which uses the price of electricity

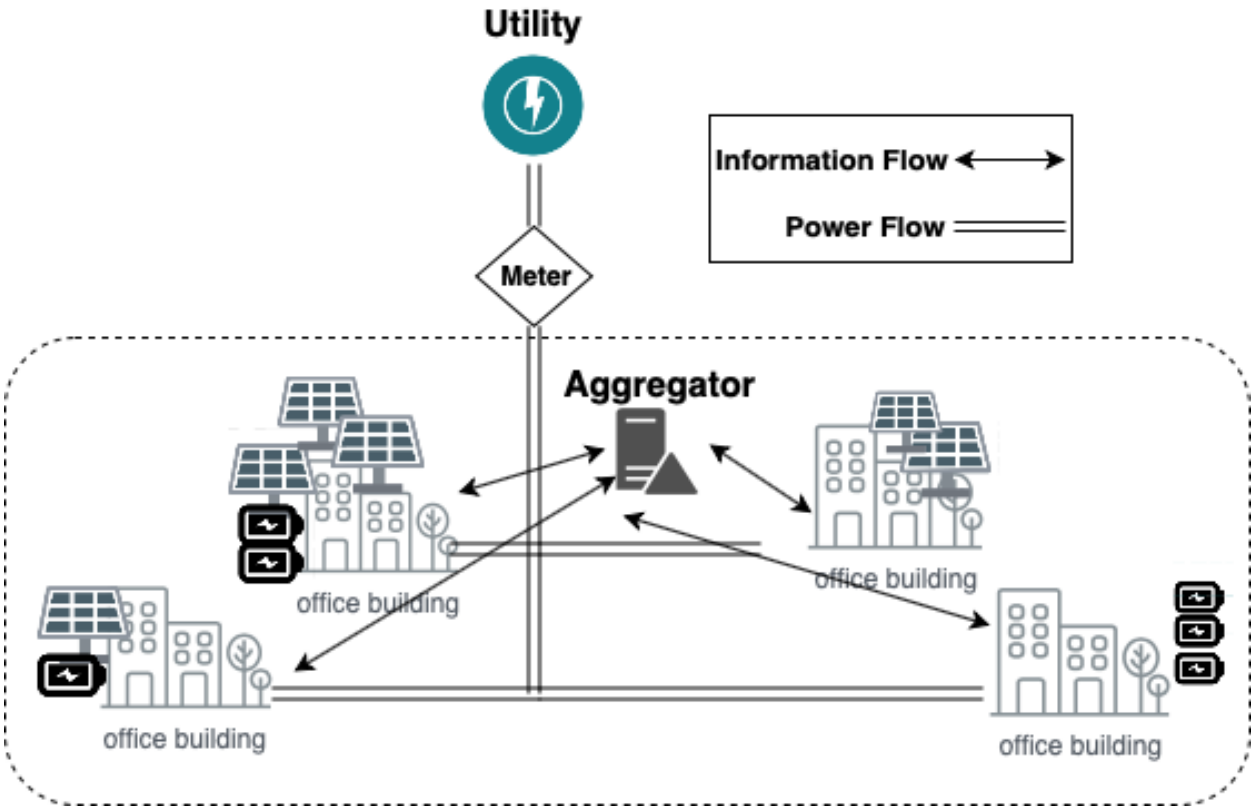


Figure 2.1: Net metering as a prosumer aggregation (Agwan et al., 2021a).

to influence the operation of prosumers. For a prosumer, responding to a day-ahead price is easier than estimating load/generation schedules required to participate in a market, or responding to real time prices. The aggregator can communicate TOU rates a day ahead to aggregation participants, who can then schedule their operation in response to energy prices similar to utility TOU rate plans. The aggregator has the task of designing prices that achieve the aggregation's objectives while dealing with an uncertain environment: *first*, the response of prosumers to energy prices is not known to the aggregator, and *second*, loads and generation are not perfectly predictable and have inherent occupancy and weather driven uncertainty. Plesae see Figure 2.1 for a visualization of what this may look like. Experts we conferred with, such as Duncan Callaway, have recognized controller-set hourly prices to be feasible in the real world (D Callaway, personal communication, 2022).

Researchers have worked on developing methods to control such aggregations. While microgrids have traditionally controlled distributed resources through a central authority dictating consumption/generation decisions, this can not be implemented in a situation where self-interested prosumers want to aggregate without ceding control of their operational decisions. (Kim and Giannakis, 2016; Wischik et al., 2008) considers a microgrid central controller trying to shape the load curves of participants by employing participant differentiated real time pricing. (Jia et al., 2013) studies the problem of minimizing deviation from day ahead estimates through pricing, and (Wang et al., 2015) models a hierarchical optimization problem to solve the aggregation control problem. We model a similar hierarchical optimization in

Section 2.2.3.

Aggregations commonly employ iterative pricing methods: (Wang and Huang, 2016) models prosumer trades as a Nash Bargaining problem, and solves it by decomposing it into two sequential problems which are solved iteratively using ADMM. This involves communicating price and energy consumption information back and forth between the aggregator and participants. Similarly, (Liu et al., 2017b) develops a pricing model for a prosumer aggregation but settles on a price in an iterative manner. These methods have a couple of disadvantages: first, they require the participants to communicate back-and-forth with the aggregator which requires two-way communication infrastructure; second, prosumers are required to develop demand forecasts, which can unnecessarily raise the computational barrier for entry.

2.2.3 RL Environment 1: MicrogridLearn

We model the prosumer and aggregator behavior as solutions to optimization problems, and then introduce the RL controller that we use to estimate prices in a day ahead manner.

Prosumer Model A prosumer typically has a combination of loads (flexible and inflexible), local generation and energy storage. We can denote the net energy consumption as $e^{(t)} = d^{(t)} - g^{(t)} + u^{(t)}$ where in any time period t, the prosumer has energy demand $d^{(t)}$, local generation $g^{(t)}$, and storage operation $u^{(t)}$. The prosumer purchases its net load at a time-of-use rate $p_b(t)$, and sells back any excess generation at $p_s(t)$. These prices are typically different (Henriquez-Auba et al., 2018), as utilities remove or disincentivize net-metering programs. The prosumer optimization problem (P-OPT) can be formulated as

$$\text{P-OPT}(\vec{p_b}, \vec{p_s}): \min_{\vec{u}} \quad \sum_{t=1}^{T} \left[p_b^{(t)} e_+^{(t)} + p_s^{(t)} e_-^{(t)} + p_{bat} |u^{(t)}| \right] \tag{2.2a}$$

$$= \vec{p_b}^\top \vec{e}_+ + \vec{p_s}^\top \vec{e}_- + p_{bat} \vec{1}_T^\top |\vec{u}| \tag{2.2b}$$

$$\text{s.t.} \qquad 0 \leq L\vec{u} \leq c \tag{2.2c}$$

where $\vec{u}$ represents the vector of battery charge/discharge over time with positive values denoting battery charging, $\vec{p_b}, \vec{p_s}$ represent the time vectors of most favorable buy and sell prices at each hour t for it between the aggregators's prices and the utilities prices, i.e.:

$$\vec{p_b} = \max(p_{b,\text{agg}},\ p_{b,\text{utility}})_t \ \forall t \in [1, T] \tag{2.3}$$

and

$$\vec{p_s} = \min(p_{s,\text{agg}},\ p_{s,\text{util},s})_h \ \forall t \in [1, T] \tag{2.4}$$

and $\vec{e_+}, \vec{e_-}$ represent the time vectors for net positive demand and net negative demand (net generation) respectively. The optimization objective (2.2a) incorporates the cost of procuring any net demand e_+ at the buy price, the revenue from selling any net surplus energy generation e_- at the sell price, as well as the cost of battery degradation evaluated with p_{bat}. The constraint (2.2c) encapsulates physical constraints on the state of charge for energy storage, charging speed constraints. More details, such as how a one-way battery efficiency is factored in, can be found in (Agwan, 2020).

Aggregator Model Aggregators can be operated as central clearing houses where energy trades are balanced, and the net consumption is procured from the utility which acts as the outside option. All prosumers purchase their net energy needs from the aggregator at a price set by the aggregator. The aggregator is constrained in its choice of prices: if it is worse then the outside option (the utility), prosumers will have no incentive to trade with it. This constraint is encapsulated as $\vec{p}_{s,\text{util}} \leq \vec{p}_{s,\text{agg}},\ \vec{p}_{b,\text{agg}} \leq \vec{p}_{b,\text{util}}$, where $\vec{p}_{s,\text{agg}},\ \vec{p}_{b,\text{agg}}$ represent the aggregator-set sell and buy prices respectively. Aggregations can be formed with multiple objectives, and we explore two particular objectives: profit maximization and market balancing.

For-Profit Aggregator Aggregators can aim to maximize the profit they earn for acting as a trade facilitator, and in a situation with perfect information they would solve the following bi-level optimization problem to set prices:

$$\max_{\vec{p}_{b,\text{agg}},\ \vec{p}_{s,\text{agg}}} \quad \left[\vec{p}_{b,\text{agg}}^{\top} \sum(\vec{e}_{*,+}) + \vec{p}_{s,\text{agg}}^{\top} \sum(\vec{e}_{*,-})\right] - \left[\vec{p}_{b,\text{util}}^{\top}(\sum \vec{e}_{*})_{+} + \vec{p}_{s,\text{util}}^{\top}(\sum \vec{e}_{*})_{-}\right] \tag{2.5a}$$

$$\text{s.t.} \qquad \vec{p}_{s,\text{util}} \leq \vec{p}_{s,\text{agg}},\ \vec{p}_{b,\text{agg}} \leq \vec{p}_{b,\text{util}} \tag{2.5b}$$

$$\vec{e}_{*} = \vec{d} - \vec{g} + \vec{u}_{*};\ \vec{u}_{*} = \arg\min_{\vec{u}} \text{P-OPT}(\vec{p}_{b},\ \vec{p}_{s}) \tag{2.5c}$$

where the objective in Eq. 2.5a represents the net profit for the aggregator, i.e. the revenue from sales to the prosumers minus the cost of procuring the net energy demand from the utility. $\sum(\vec{e}_{*,+}),\ \sum(\vec{e}_{*,-})$ are the sum of each prosumer's optimal demand and generation taken separately, while $(\sum \vec{e}_{*})_{+},\ (\sum \vec{e}_{*})_{-}$ are the net demand and generation of the aggregation once all internal trades have been balanced.

Transactive Control Using RL The problems modeled in Section 2.2.3 are hierarchical optimization problems, and do not have a closed-form solution without some form of information sharing between the aggregator and prosumers. As discussed in other papers which use ADMM and iterative pricing methods (Section 2.2.1), decentralized pricing methods require iterations to converge to a solution. We develop an RL controller that relies on a day-ahead price set using historical price information, generation forecasts, and prior energy consumption; i.e. the state space:

$$\mathcal{S}_t := [\vec{p}_{t-1},\ \vec{e}_{t-1},\ \vec{g}_t] \in \mathbb{R}^{3T} \tag{2.6}$$

The transactive controller does not iterate over prices, and instead learns to estimate future prices in a one-shot manner. Our action space $\mathcal{A}$ is the space of aggregator prices:

$$\vec{p}_{\text{agg}} \in \mathbb{R}^{T} \tag{2.7}$$

It is simple enough to be covered by an entropy maximizing agent, and we employ a SAC architecture to do so. The reward r for the for-profit aggregator is computed as the objective expressed in Eq. 2.5a. We simulate the behavior of our controller under uncertain generation forecasts and compare it to baseline iterative pricing algorithms in Section 3.5.2.5.

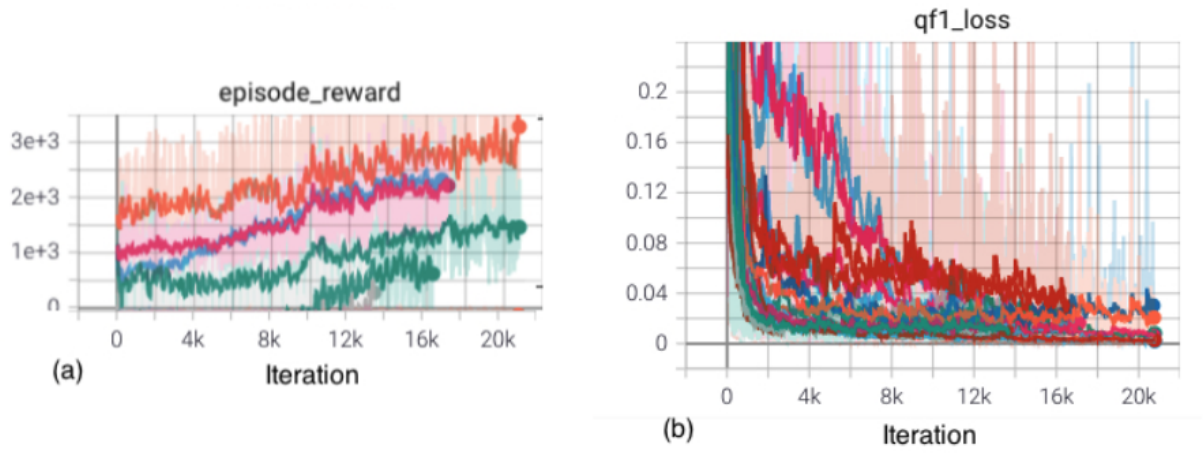


Figure 2.2: Training curves for RL agents under different instantiations of MicrogridLearn (Agwan et al., 2021a).

2.2.3.1 Results and Discussion from MicrogridLearn

We will now describe our results for each of the two objectives: market solving and profit maximization, prefaced by an explanation of our data, architecture and training process. Each step in the environment is a day where the RL controller broadcasts prices to the prosumers, who modify their energy storage and consumption to minimize their costs, and the controller uses their consumption data to calculate its reward.

Implementation We use the stable-baselines fork of OpenAI baselines (Hill et al., 2018), and our other implementation choices are detailed in our Github repository (Spangher et al., 2021). The final run presented here was distributed across 24 Central Processing Units (CPUs) for 12 hours each. The Q loss shown in Fig. 2.2 is one of many metrics that represents the neural network's training. For the simulations presented in this paper, the utility pricing is obtained from (OpenEI, 2017) and the prosumers considered are commercial office buildings modeled using load data taken from (Miller and Meggers, 2017) with additional details presented in (Agwan, 2020). A total of 10 commercial and university buildings in the Los Angeles region were chosen for simulation. Their power usage is of the order of $10 - 200$ kW, and the simulations use three levels of battery installations and PV array sizes for each of the buildings: small, medium and large.

Marginal Benefit of Aggregating We compare the system costs in the presence and absence of a profit-maximizing aggregator for two different levels of prosumer solar generation and battery capacity. As can be see in Fig. 2.3, the RL controller reduces the system costs and provides value which can be distributed among the aggregator and prosumers.

RL controllers are particularly well suited for dynamic data driven environments with unknown, complex, and changing system models, e.g. with changing prosumer resources. The novelty of our proposition is in using RL to preemptively price energy in local markets, but is accompanied by many challenges that will need to be addressed for practical applicability.

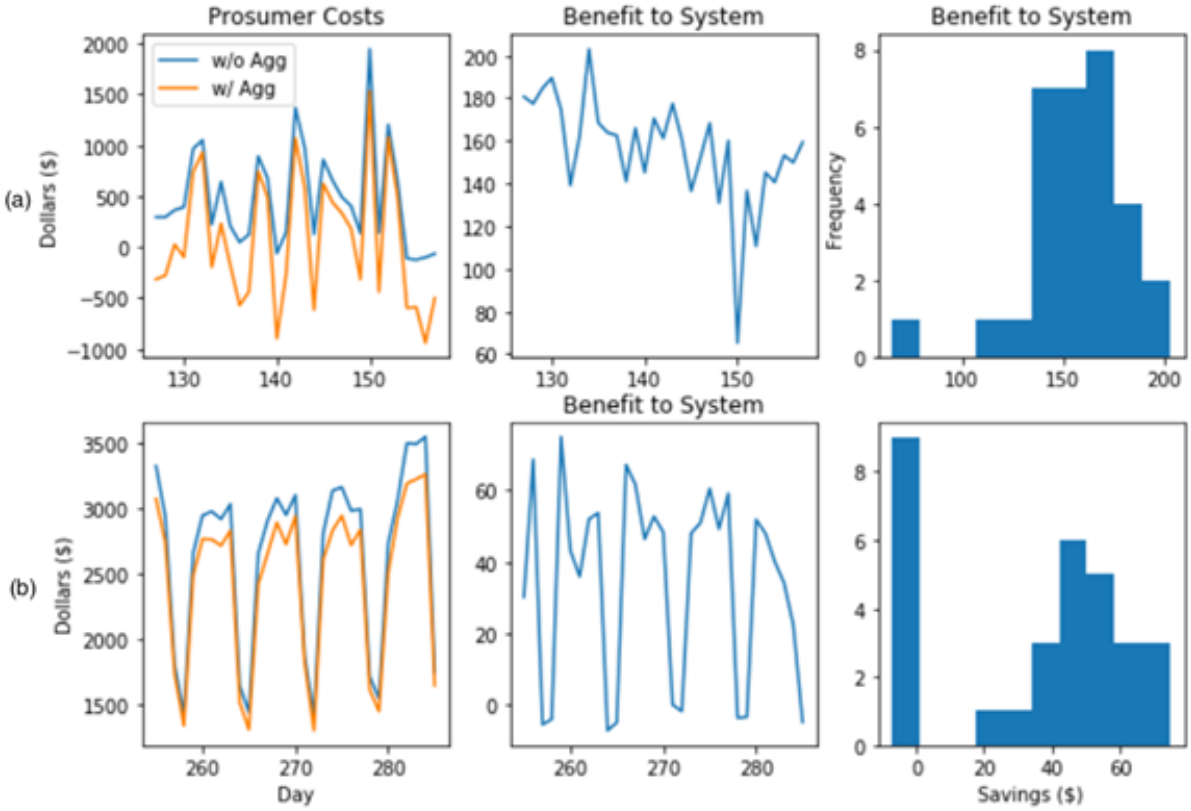


Figure 2.3: Comparing system costs, i.e. sum of aggregator and prosumer costs with and without a profit maximizing RL controller for two resource levels: a) Medium, and b) Small. (Agwan et al., 2021a)

2.3 Within a Building

2.3.1 Background and Motivation

We now wish to ask the question: what tools may a building manager leverage to respond to prices set at the aggregator or utility scale? We will describe the existing work in building-level DR , specifically focusing on behavioral energy deferral.

Buildings are a natural place to implement automated DR strategies. It has been shown that prices can affect DR in individual homes, be they appliance or system-based response. For instance, (Adelman and Uçkun, 2019) study the effect of passing a dynamic price profile to homes with smart meters and appliances, comparing this profile to flat and peak pricing and concluding that dynamic prices elicit changes in demand.

Many DR strategies have been explored to coordinate the action of appliances, such as linear programming for HVAC controls (Kim, 2018), rule-based HVAC controllers (Yoon et al., 2014b), and supervised machine learning for Internet of Things (IoT) device control (Kaur et al., 2021). For example, RL has been used for HVAC and window controllers (Chen et al., 2018), whole building HVAC controllers (Azuatalam et al., 2020), commercial building HVAC controllers (Kathirgamanathan et al., 2021), and home energy management (Lissa et al., 2021). (Gao et al., 2015) examines the control of appliances in the presence of load uncertainty using a Monte Carlo method, and (Peirelinck et al., 2022) looks into different approaches to the same problem by minimizing uncertainty through transfer learning approaches.

2.3.2 Behavioral Energy Shifting via Transactive Control

We will now dive into another type of load shifting that a building may seek to develop.

For the purposes of this work, we define *behavioral energy demands* as demands related to direct uses of energy by office workers. In a traditional office, desk-level resources may include computer use, refrigerator use, desk fan use, and cell phone charging. In newer offices, common energy demands may be behaviorally influenced by voting systems that allow users to submit input on what level heating, cooling, or light intensity should be at. One may consider these novel systems as ways to incorporate behavioral preferences into energy DR .

Studying behavioral DR is very relevant. Most DR research has focused on intelligent control of appliance signals ((Asadinejad et al., 2018), (Ma et al., 2015), (Li et al., 2018), (Yoon et al., 2014a), (Johnson et al., 2015), (Das et al., 2020), (Kim, 2018)). However, there is limited literature on the supply side of the problem; i.e. how to actually set those signals. There is even less literature on conducting *behavioral* DR to influence the energy consumption behavior at the level of office workers instead of at the level of markets composed of entire buildings. . The lack of office-centered price-setting studies is understandable considering that offices do not have a mechanism to pass energy prices to workers. Social games have been studied as a way to address this issue. A reversed Stackelberg social game is implemented in (Ratliff et al., 2014) with results from real office workers. The use of cheap IoT and mobile devices for social games is explored in (Papaioannou et al., 2018) and (Iria et al., 2020), respectively. These Social Games have been shown to motivate energy behavior in workers

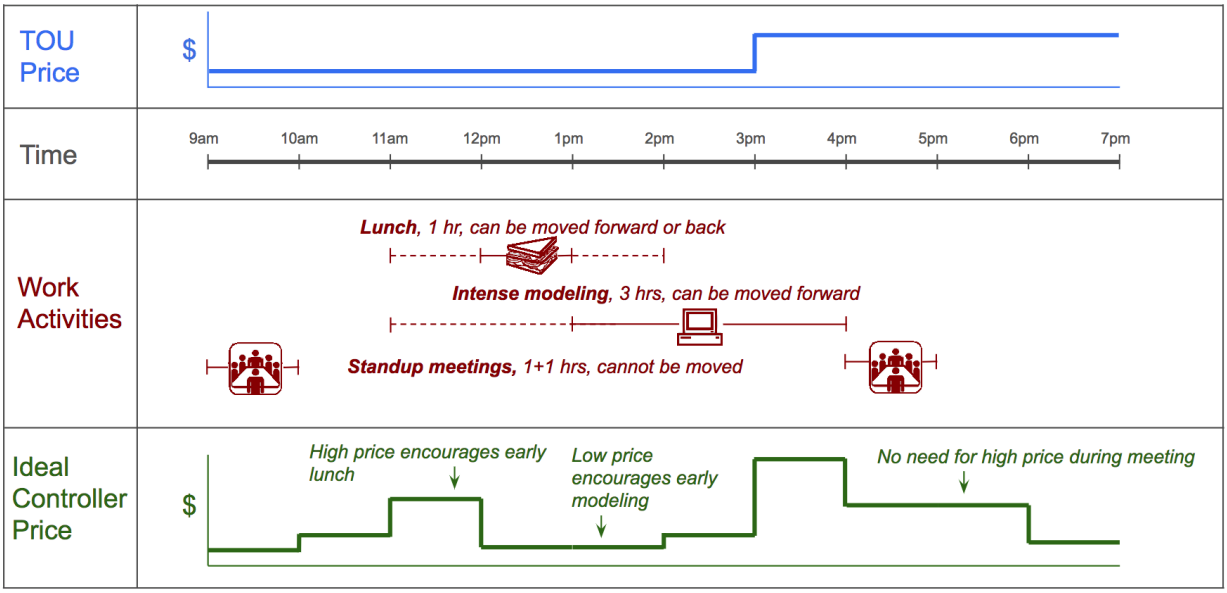


Figure 2.4: A schematic of the energy consumption in a hypothetical "Acme Energy office" that demonstrates the value in tailoring a price signal for a specific office schedule.

by sharing energy consumption information and rewards for energy savings (Spangher et al., 2019c)

In the presence of such a Social Game mechanism, "artificial" price signals could represent an exciting control for DR . To elaborate, consider an office building that receives energy from a utility via a TOU pricing program. DR amounts to the ability of workers to shift some of their energy usage across time or curtail energy consumption to reduce the energy cost with respect to the price signal shared with them. The simplest form of a price signal could be the TOU price schedule of the utility itself. This signal may not be efficient in eliciting the best DR from workers since it does not account for the nature of work being performed and the ability of specific workers to shift consumption. Therefore, "personalizing" a price signal to a building; specifically the energy responses of its workers, could lead to lower building energy costs than passing through the TOU price signal.

To illustrate, consider the example shown in Figure 2.4. In Acme Energy Office, the office manager pays for energy on a TOU pricing that is high from 3pm-7pm, which means that the office manager would like to shift energy use earlier in the day. The office workers generally take their lunch at noon, which is low-energy, but can move lunch an hour before or afterwards. Thus, a high price before they normally take lunch may encourage them to shut down their computers and take lunch early. The office generally collaborates on intensive computer modeling from 1pm-4pm, which should be an unbroken three hours but can be moved forward. If lunch is moved forward, then lower prices from 12pm-3pm and the highest price of the day from 3pm-7pm may encourage a shift of the modeling from 12pm-3pm. The possible ideal controller price may encourage a shift in demand to the middle of the day that allows the demand to coincide with a time period when solar generation affects overall grid energy composition.

Thus, the work we propose creates a mechanism to study behavioral response. In section 2.3.3 we will propose a simulation environment for testing RL for price-setting within buildings, and in 4.1 we will propose an experiment to validate the work.

2.3.3 RL Environment 2: OfficeLearn

Here, we formalize an OpenAI Gym environment for the testing of RL agents within a single office building to encourage exploration in occupant level building DR. At a high level, the office workers are incentivized with a fixed cash reward to save energy by either shifting some of their energy usage to parts of the day when price signals make energy cheaper, or curtailing some energy usage altogether. See Figure 2.5 for an overview of our overall setup.

At a high level, Figure 2.5 is a schematic showing the interplay between agent and office environment, and ensuing energy responses. The agent receives prices from the grid, then transforms it into "points" (called as such for differentiation.) Office workers engage with the points in the way an individual might be engaged with their home energy bill, which is reasonable assuming behavioral incentives detailed in (Spangher et al., 2020c). The office receives these points at the beginning of the "day". Workers proceed to use energy throughout the day and at night the system delivers a record of their energy consumption, which is reduced into a reward that trains the agent.

2.3.3.1 Formal Setup

We consider an office of U simulated people with H hour workdays [21]. The office buys energy from a grid-level utility that sells energy at different prices during different hours of the day. We assume the utility uses a TOU pricing scheme [22]; i.e. a pricing scheme with a base price for much of the day and a higher price for a few hours when grid demand is congested that is constant across the year[XXIV].

A building manager wishes to shape energy consumption to lower the building's current energy cost, which is the sum of energy cost incurred by each worker at each hour of the day: $\sum_{i=1}^{U} \vec{p}_{\text{util}}^{\intercal} \vec{b}_i$, where $\vec{b}_i = [b_{i,1}, ..., b_{i,H}]$ is the baseline energy consumption of the i^{th} worker when there is no DR, $\vec{p}_{\text{util},h}$ are the bounded, nonnegative TOU grid prices for each hour $h = 1, 2, \ldots, H$. In contrast, workers are unaware of energy prices and have no direct incentive to change behavior. Therefore, the building manager implements a DR system consisting of a price controller and a reward system. The price controller chooses a vector of bounded, nonnegative prices $\vec{p}_{\text{RL}} = [p_{\text{RL},1}, ..., p_{\text{RL},H}]$ for each hour of the day and broadcasts these prices to all workers. In addition, the building manager incurs a total fixed expense of I dollars ($) to incentivize workers to minimize their energy cost relative to this price signal. More

[21] We focus on homogeneous workdays for simplicity, i.e. ones in which every worker responds the same, but our model extends to handle heterogeneous workdays in a straightforward manner.

[22] As a reminder to the reader, other approaches to dynamic pricing also exist, such as a Real Time Pricing (RTP) scheme, in which every hour has a different price and every day features a different price curve, and the familiar "flat pricing", which features a single price per day. Time of Use (TOU) pricing has two prices in a day. We choose the unvarying TOU case for our simulation to reflect a common real-world middle ground, but our model extends to handle a fully varying RTP scenario in a simple manner

formally, when facing price $\vec{p}_{\text{RL}}$, the worker i deviates to net enegy $\vec{e}_i$ from his/her baseline consumption $\vec{b}_i$ to a deterministic consumption response function[23]

$$\vec{e}_i := d_i(\vec{b}_i, \vec{p}_{\text{RL}}) := [d_{i,1}(b_i, p_{\text{RL}}), d_{i,2}(b_i, p_{\text{RL}}), \ldots, d_{i,H}(b_i, p_{\text{RL}})] \tag{2.8}$$

The building manager's objective is to minimize the total worker energy cost by setting the prices $\vec{p}_{\text{RL}}$ to influence worker energy consumption $d_i(b_i, p_{\text{RL}})$:

$$\min_{\vec{p}_{\text{RL}} \in \mathbb{P}_{\text{RL}}} \sum_{i=1}^{U} \vec{p}_{\text{util}}^{\,\intercal}\ d_i(\vec{b}_i,\ \vec{p}_{\text{RL}}) + I. \tag{2.9}$$

Let $\vec{p}_{\text{RL},*}$ denote an optimal solution to this problem.

Note that the fixed amount I paid to incentivize workers affects the optimal $\vec{p}_{\text{RL},*}$ only in terms of guaranteeing the consumption response $d_i(\vec{b}_i, \vec{p}_{\text{RL},*})$ from the workers in Equation 2.9. From this objective formulation, we can see that implementation of DR is desirable to the building manager only if:

$$\left[\sum_{i=1}^{U} \vec{p}_{\text{util}}^{\,\intercal}\ d_i(\vec{b}_i,\ \vec{p_{\text{RL}}}) + I\right] < \sum_{i=1}^{U} \vec{p}_{\text{util}}^{\,\intercal}\ \vec{b}_i \tag{2.10}$$

that is, if the total energy cost under DR and the cost of the incentive is smaller than the total energy cost without DR.

2.3.3.2 Environment Mechanics Overview

In this section, we highlight a summary of the environment and the underlying Markov Decision Process. The flow of information is succinctly expressed in Figure 2.5.

The environment takes the format of the following Markov Decision Process (MDP), $(\mathcal{S}, \mathcal{A}, \mathcal{P}, r)$. They are defined as follows:

$$\mathcal{S}_t := [\vec{p}_{\text{util, t}}, \vec{d}_{t-1}, \vec{b}_t] \in \mathbb{R}^{3H} \tag{2.11}$$

Where H=10 to cover the workday.

$$\mathcal{A} : \vec{p}_{\text{RL}} \in \mathbb{R}^H \tag{2.12}$$

$$\mathcal{P} : \mathbb{1}[\vec{e}_i == d_i(\vec{p}_{\text{RL}})]\ \forall\ i \in [1, ...U] \text{ where } \vec{e}_i \in \mathbb{R}^H \tag{2.13}$$

The reward, r, defined in Section 2.3.3.6.

We describe out design choices and variants of the MDP below.

[23] A deterministic DR may be appropriate in certain building environments such as data centers or other largely human-free buildings. It is a mild assumption in our work as we will relax the assumption that we know this function in the next section.

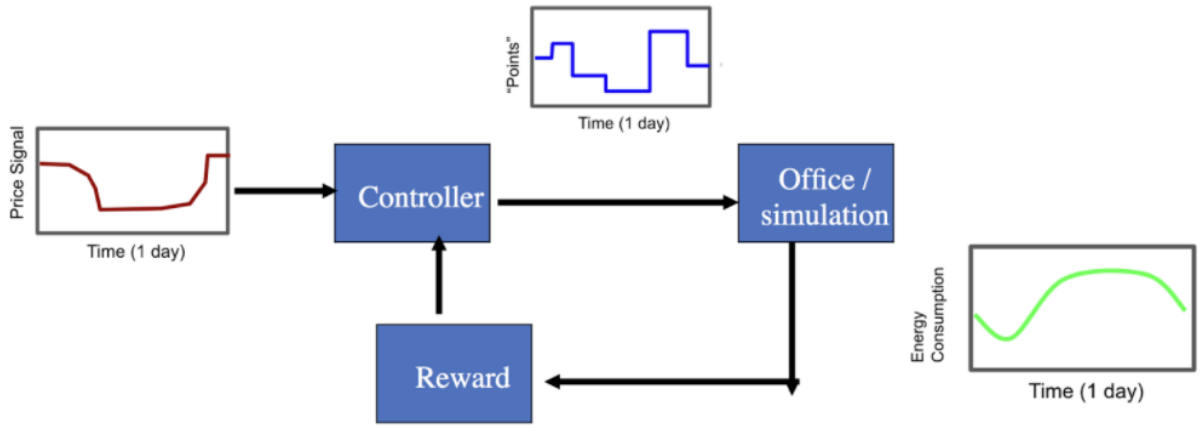


Figure 2.5: RL flow of the OfficeLearn environment.

2.3.3.3 State Space, $\mathcal{S}$

The steps of the agent are currently formulated day by day, with ten-hour working days considered. Therefore, while the state space has several different components (described below), each is of ten dimensions as each one is hourly in nature.

2.3.3.4 Grid Price Regimes

Utilities are increasingly moving towards time dependent energy pricing, especially for bigger consumers such as commercial office buildings with the capacity to shift their energy usage. TOU pricing involves is a simple, two-level daily price curve that changes seasonally and is declared ahead of time. We use PG&E's TOU price curves from 2019. Real Time Pricing (RTP), meanwhile, is dynamic for every hour and changes according to supply and demand in the energy market. We simulate it by subtracting the solar energy from demand of a sample building. There is significant seasonal variation in prices depending on geography, e.g. in warmer climates, the increased cooling load during summer can cause an increase in energy prices.

- **Energy of the prior steps**, $\vec{d}_{t-1}$: The default instantiation of the environment includes the energy use of office workers of the prior step. This allows the agent to directly consider a day-to-day time dependence. The U simulated office workers in this version are currently memory-less from day to day in their energy consumption, but a future simulation will allow for weekly deferable energy demands to simulate weekly work that can be deferred and then accomplished. The energy of the prior steps may be optionally excluded from the state space by those who use our environment.
- **Grid prices of the prior step**, $\vec{p}_{\text{util, t-1}}$: Users may optionally include the grid price from prior steps in the state space. This would allow the agent to directly consider the

behavioral hysteresis that past grid prices may have on a real office worker's energy consumption. Although this is a noted phenomenon in human psychology generally (Richards and Green, 2003), it is not well quantified and so we have not included it in how we calculate our simulated human agents.

- **Baseline energy**, $\vec{b}_t$: Baseline Energy may optionally be included in the state space. If the agent directly observes its own action and the baseline energy, it observes all of the information necessary to calculate certain simpler simulated office worker responses. Therefore, inclusion of this element will make the problem fully observable, and truly an MDP rather than a POMDP.

- **Action space**, $\mathcal{A}$: The agent's action space expresses the points that the agent delivers to the office. The action space is by default a continuous value between zero and ten, but may be optionally discretized to integer values if the learning algorithm outputs discrete values. The purpose of the action is to translate the grid price into one that optimizes for behavioral response to points. Therefore, the policy will learn over time how people respond to the points given and maximally shift their demand towards the prices that the grid gives.

2.3.3.5 Office Workers: Simulated Response Functions

In this section, we will summarize various simulated responses that office workers may exhibit.

- **"Deterministic Office Worker"**: We include three types of deterministic response, with the option for the user to specify a mixed office of all three.

  - In the linear response, we define simple office worker who decreases their energy consumption linearly below a baseline with respect to points given. Therefore, if $\vec{b}_t$ is the baseline energy consumption at time t and $\vec{p}_{RL,t}$ are the points given, the energy demand $\vec{d}_t$ at time t is $\vec{d}_t = \vec{b}_t - \vec{p}_{RL,t}$, clipped at d_{min} and d_{max} as defined in Section 2.3.3.6.
  - In the sinusoidal response, we define an office worker who responds well to points towards the middle of the distribution and not well to prices at the. Therefore, the energy demand $\vec{d}$ at time t is $\vec{d}_t = \vec{b}_t - \sin \vec{p}_{RL,t}$, clipped at d_{min} and d_{max}.
  - In the threshold exponential response, we define an office worker who does not respond to points until they are high, at which point they respond exponentially. Therefore, the energy demand $\vec{d}$ is

    $$\vec{d}_t = \vec{b}_t - [\exp \vec{p}_{\mathrm{RL},t} * (\vec{p}_{\mathrm{RL,\ t}} > 5)] \tag{2.14}$$

    Clipped at d_{min} and d_{max}.

- **"Curtail And Shift Office Worker"**: Office workers need to consume electricity to do their work, and may not be able to curtail their load below a minimum threshold, e.g. the minimum power needed to run a PC. They may have the ability to shift their load over a definite time interval, e.g. choosing to charge their laptops ahead

of time or at a later time. We model a response function that exhibits both of these behaviors. We can model the aggregate load of a person ($\vec{b}_t$) as a combination of fixed inflexible demand ($\vec{b}_t^{\text{fixed}}$), curtailable demand ($\vec{b}_t^{\text{curtail}}$), and shiftable demand ($\vec{b}_t^{\text{shift}}$), i.e., $\vec{b}_t = \vec{b}_t^{\text{fixed}} + \vec{b}_t^{\text{curtail}} + \vec{b}_t^{\text{shift}}$. All of the curtailable demand is curtailed for the T_{curtail} hours (set to 3 hours in practice) with the highest points, and for every hour t the shiftable demand is shifted to the hour within $[t - T_{\text{shift}}, t + T_{\text{shift}}]$ with the lowest energy price. For an example of "curtail and shift" in the wild, please see the endnotes[XXV].

2.3.3.6 Reward

Specification of the reward function is notoriously difficult, as it is generally hand-tailored and must reduce a rich and often multi-dimensional environmental response into a single metric. Although we include many possible rewards in the code, we outline the two rewards that we feel most accurately describe the environment. As we already demonstrated in prior work the ability to reduce overall energy consumption (Spangher et al., 2019a), we endeavor to direct this agent away from reducing consumption and towards optimally shifting energy consumption to favorable times of day.

- **Scaled Cost Distance**: This reward is defined as the difference between the day's total cost of energy and the ideal cost of energy. The ideal cost of energy is obtained using a simple convex optimization. If $\vec{d}$ are the actual demand of energy computed for the day, $\vec{p}_{\text{util}}$ is the vector of the grid prices for the day, E is the total amount of energy, and d_{min}, d_{max} are 5% and 95% values of energy observed over the past year, then the ideal demands are calculated by optimizing the objective:

$$\vec{d}_t^{\,*} = \arg\min_{\vec{d}} \ \vec{d}_t^{\intercal} \vec{p}_{\text{util, t}} \tag{2.15}$$

  Subject to the constraints $\vec{1}^{\intercal}\vec{d}_t = E_t$ and $d_{min} < \vec{d} < d_{max}$. Then, the reward becomes:

$$r(d) = \frac{\vec{d}^{\,*\intercal}\vec{p}_{\text{util}} - \vec{d}^{\,\intercal}\vec{p}_{\text{util}}}{\vec{d}^{\,*\intercal}\vec{p}_{\text{util}}} \tag{2.16}$$

  i.e. taking the difference and scaling by the total ideal cost to normalize the outcome.

- **Log Cost Regularized**: Although the concept of the ideal cost is intuitive, the simplicity of the convex optimizer means that the output energy is often an unrealistic, quasi step function. Therefore, we propose an alternate reward of log cost regularized. Following the notation from above, the reward is

$$r(d) = -\vec{d}^{\,\intercal}\vec{p}_{\text{util}} - \lambda \left[\sum_{t=1}^{H} \vec{d} < 10 * (.5 * b_{max}) \right] \tag{2.17}$$

  where b_{max} refers to the max value from the baseline. In practice, we set λ to some high value like 100. The purpose of the regularizer is to penalize the agent for driving down energy across the domain, and instead encourage it to shift energy.

2.3.3.7 Illustration of Features

We will now demo the environment's functioning. All comparisons are done with a vanilla SAC RL agent that learns throughout 10000 steps (where one step is equal to one day), with a TOU pricing regime fixed at a single day. The agent's points are scaled between -1 and 1.

We present the effect of using the Log Distance Regularized and the Scaled Cost Distance. Please see Figure 2.6 for side by side comparison of the reward types. The energy output of the simulated office workers is drawn in light blue, and corresponds to the primary axes. The grid prices are drawn in red, and refers to TOU pricing. It corresponds to the secondary axes. The agent's actions are drawn in dark blue, is scaled between -1 and 1 to improve readability of the plots, and correspond to the secondary axes. In this figure, you can see that not only is the agent capable of learning an action sequence that accomplishes a lower cost than if the simulated office workers were to respond directly to the untransformed grid prices, but also differs in how the learning is guided. The log cost regularized reward accomplishes smoother prices that result in the agent deferring most of the energy for the end of the day, whereas the scaled cost distance reward allows for more energy earlier in the day, guiding the simulated office worker to increase energy gradually throughout the day.

We present the effect of using different simulated office workers on the output of energy demand. Please see Figure 2.7, in the Appendix, for a comparison of two types of simulated office workers. The energy output of the simulated office workers is drawn in light blue, and corresponds to the primary axes. The grid prices are drawn in red and corresponds to the secondary axes. The agent's actions are drawn in dark blue, is scaled between -1 and 1, and correspond to the secondary axes. In the exponential response, we see an example of how the office worker's energy demand responds to points – that is, perhaps, too coarsely for a learner to make much difference. Meanwhile, the Curtail and Shift response demonstrates a much richer response, which enables a learner to learn the situation and perform better than the control.

2.3.3.8 Simulating DR in Your Building

The environment we provide contains many ways to customize your own building. You may choose the number of occupants, their response types, baseline energies, grid price regimes, and frequency with which grid price regimes change. You may also choose from a host of options when it comes to customizing the agent and its state space.

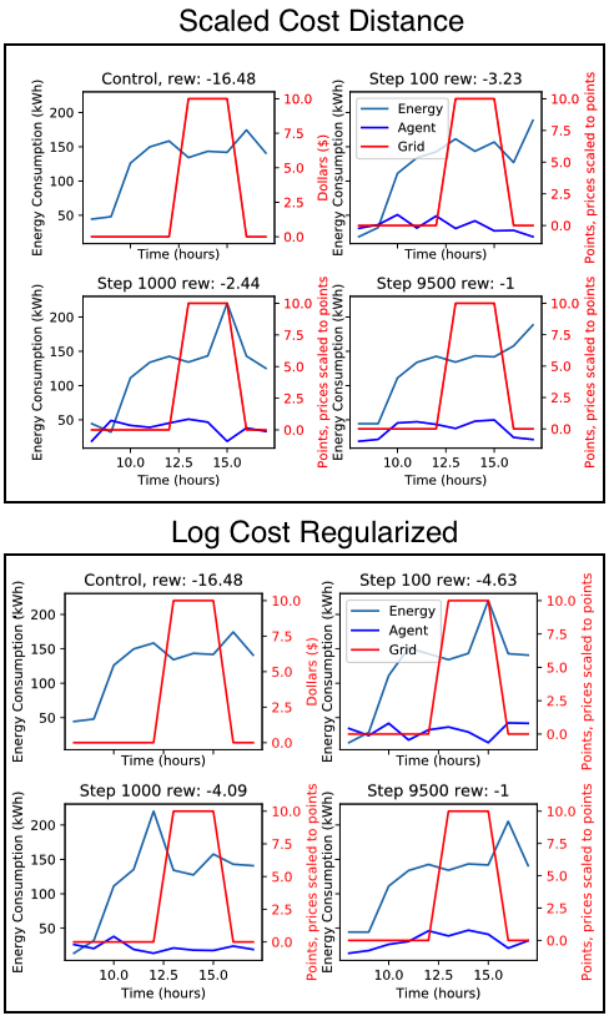


Figure 2.6: A comparison of the Log Cost Regularized and the Scaled Cost Distance rewards. The energy output of the simulated office workers is drawn in light blue, and corresponds to the primary axes. The grid prices are drawn in red, and refers to TOU pricing. It corresponds to the secondary axes. The agent's actions are drawn in dark blue, is scaled between -1 and 1 to improve readability of the plots, and correspond to the secondary axes.The control (top left) simply sets $\vec{p}_{\text{RL}} := \vec{p}_{\text{util}}$. (Spangher et al., [n. d.])

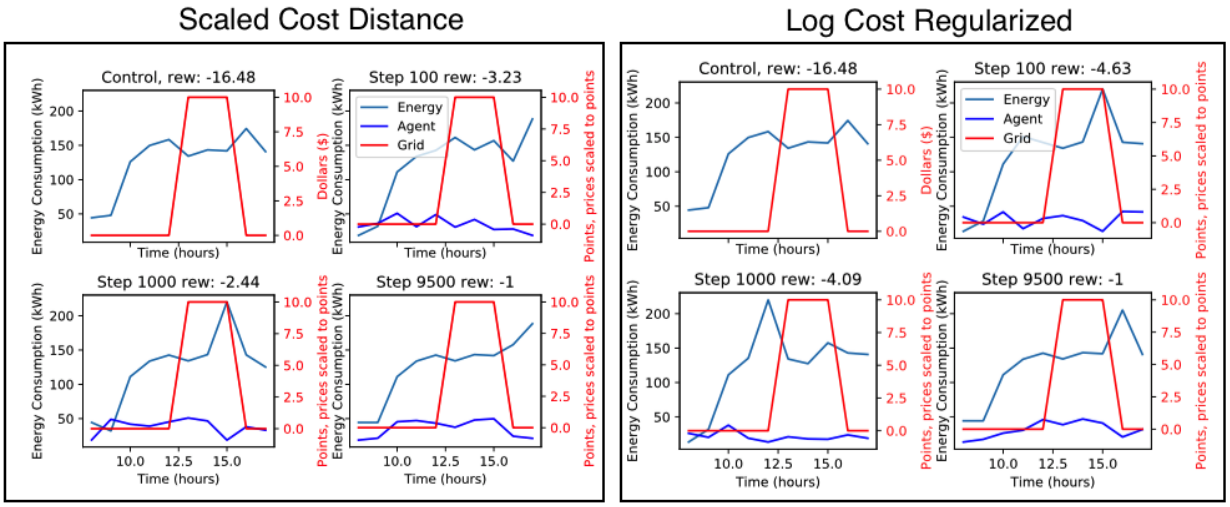


Figure 2.7: A comparison of the "Exponential Deterministic Office Worker" to the "Curtail and Shift Office Worker". The energy output of the simulated office workers is drawn in light blue, and corresponds to the primary axes. The grid prices are drawn in red, and refers to TOU pricing. It corresponds to the secondary axes. The agent's actions are drawn in dark blue, is scaled between -1 and 1 to improve readability of the plots, and correspond to the secondary axes. The control (top left) simply sets $\vec{p}_{\text{RL}} := \vec{p}_{\text{util}}$. (Spangher et al., 2020d)

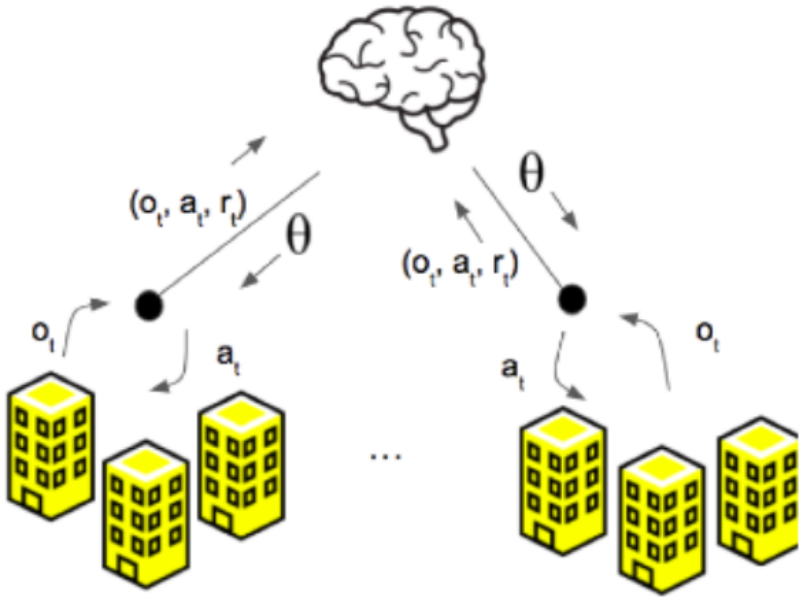


Figure 2.8: A description of the multi-microgrid environment. (Gunn et al., [n. d.])

2.4 Within Microgrid Clusters and Beyond

We now broaden our scope to applications above the level of just one microgrid or one building, and consider a similar transcative structure to larger entities.

2.4.1 RL Environment 3: Multi-Agent MicrogridLearn

Consider a setting of 100 microgrids. One RL agent sets the policy parameters θ of all 100 microgrid controllers, which transacts locally within each microgrid. Each microgrid consists of 7 prosumer office buildings. Every prosumer has a battery, solar panel array, and baseline energy consumption; each wants to minimize their energy cost. Prosumers see both grid-set hourly energy buy and sell prices and local microgrid controller-set hourly energy buy and sell prices. Prosumers choose to transact with either the grid or the RL aggregator at each hour. Prosumers also decide when to discharge their battery according to both their demand and the energy prices. The microgrid controller accepts all transactions the prosumers request of it. It does not produce or store energy, but sells energy it has bought from prosumers producing energy in a timestep to prosumers demanding energy in the same timestep. The aggregator balances the net load by purchasing from or selling to the energy utility under which they sit, usually at a loss. As the manager of the RL-aggregator, you see the grid's buy and sell prices, and wish to learn an automatic pricing strategy such that you consistently turn a profit. See Figure 2.8 for a graphical depiction of the environment.

We create this problem setup by extending MicrogridLearn to a multi-agent environment.

Here, each agent sets prices for a microgrid defined by the same hyperparameters as the original environment. Please see Figure 2.8 for a visualization of what this may looks like. In this figure, the brain is the multi-agent RL agent, the black dot is a distributed microgrid controller, which constitutes one agent in a multi-agent framework. Each collection of yellow buildings is a *separate* and possibly different microgrid that the brain controls.

2.5 A Brief Note on Leveraging Hierarchy in the Grid for Multiple levels of RL Control

As we demonstrate to show, similar price-setting architectures may be applied at multiple levels of the grid. At the simplest level, all a price-setter would have to know would be the prices set by those above it and an accurate reward calculation of the aggregate entities below it. The mechanics of RL would help different price-setters at different levels adapt to the dynamics of their position in the grid.

This observation relies on the fact that, for our purposes, the grid may be simplified into a real-life tree with no functionality lost from the simplification[24]. Each node receives information of actions taken first by a linear line of nodes above it takes an action that affects the subtree below it, and then observes the effect of its action. Peer nodes or nodes not directly in the linear causal chain from root to node would set prices for different subsections of the utility's base, and so would not effect the node in question.

We thus imagine that a minimal extension of our work is to put multiple RL controllers in hierarchy with each other. We leave this currently as a thought experiment (Spangher, 2021), but are excited to share it and possibly pursue it in the future.

[24]The physical grid is not an exact tree, given redundant lines and some amount of bidirectionality with solar. However, for the purposes of transactional directions only, it simplifies into a tree with no lost information (Von Meier, 2006).

Chapter 3

Implementing Reinforcement Learning

3.1 Introduction

3.1.1 Overview

RL is not a magic bullet. We seek to identify six different problems in Simulation-to-Reality (Sim-to-Real) RL and apply them in the environments we have previously designed. Although each will each be applied within a single environment, they are all extensible to other environments.

We will provide several different enhancements on RL to address the problems that we define. Thus, this section may be seen as a survey of the add-ons of RL that were promising at the time of this publication. We also seek to identify which of the methods seemed especially promising to us.

We borrow Sutton's RL cell to put rough approximations of the problems we surface in Sim-to-Real RL onto the RL agent. Please see Figure 3.1 for a visualization. We hope that this visual serves as a helpful overview of the topics that we will cover.

3.1.2 Simulating Sim-to-Real Test Cases

It may be non-traditional to structure a dissertation around Sim-to-Real RL when there is no "real" experiment that we have tested our agents in. While we had prepared for one[25], and so have some insight in the tools needed beyond pure RL (please see the API and proposed experiment design), we still cannot fully claim to test our modifications in the real world.

We persistently make the argument that one may create two levels of complexity in simulation. When testing Sim-to-Real techniques in simulation, we argue that one may limit their training to the lower level of complexity and test in the higher. When testing Sim-to-Real techniques in the real world, we argue that one can then use both levels of complexity in simulation to prepare the agent for the noise that it may face in the world. In using simulation complexity as a proxy for Sim-to-Real complexity, we hope to make the case for certain methods that we champion.

[25]Unfortunately, SARS-CoV-2 (COVID-19) prevented us from running the experiment in a test office.

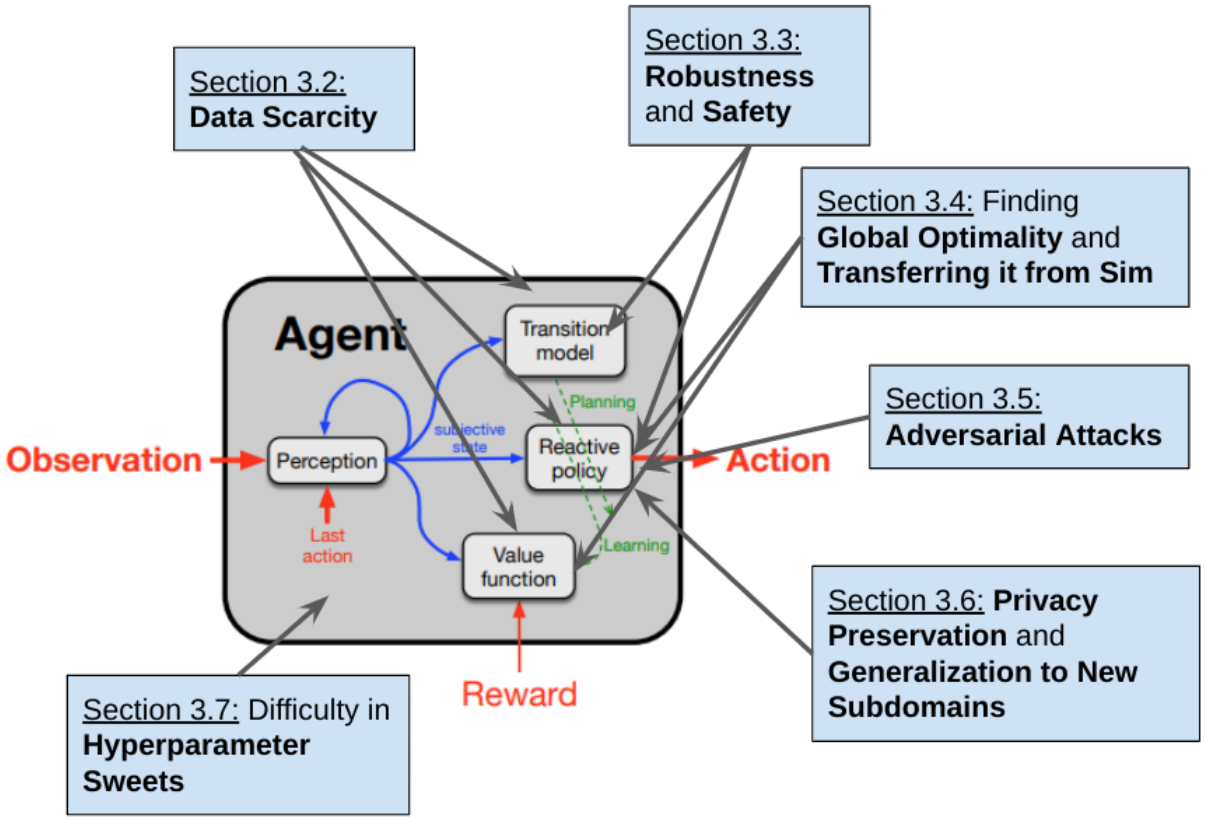


Figure 3.1: Overview of the problems we identify in sim-to-real RL. Base image from (Sutton, 2022).

3.2 Data (In)Efficiency

3.2.1 Introduction to the Problem

It is well known that RL requires large amounts of data to run, which hamstrings its ability to be performant in the real world. We demonstrate the problem in our own work and then recommend our preferred strategies after several we chose.

In the OfficeLearn environment, we were able to achieve performance that beat both a control and standard, TOU baseline. Indeed, we observe from Fig. 3.2 that it is possible to eventually achieve a pricing strategy that saves $4500 per year compared to buildings with no incentive structure (Flat pricing) and $3250 per year compared to TOU. However, 10,000 steps of data is necessary to achieve performance comparable to TOU: approximately 40 years of training data. We will define the "initial deployment cost" as the higher energy cost that the RL price controller incurs relative to the TOU pricing strategy during the initial time period when the RL prices perform worse than the TOU price. The initial deployment cost of the SAC agent is approximately $175,000, suggesting that at least 93 years of deployment is necessary to recoup the initial deployment cost. The return over 5 years would be $37,500 less than using TOU. This clearly unrealistic data requirement shows that an off-the-shelf SAC agent is in fact worse than the TOU price for all practical purposes. It suggests potential value in modifying the vanilla RL architectures in order to have a custom pricing agent for our setting.

3.2.2 Offline-Online RL and DAggr

3.2.2.1 Offline-Online Reinforcement Learning

With the online "vanilla" SAC optimization procedure, several decades worth of real-world training data would have to be collected to fully train an hourly price-setting controller (Spangher et al., 2020b) in our Social Game. We seek to leverage a detailed simulation with behaviorally reasonable dynamics encoded in a model that can train on both simulated and experimental environments to accelerate this process. SAC is an off-policy algorithm, which means that it can be trained on state transitions that did not originate from its policy. This allows SAC to be used to train networks on offline datasets of previously collected samples. Thus we propose pretraining SAC on an offline dataset of state transitions collected from our Social Game simulations, in order to learn a warm-started neural network initialization that can generalize to the experimental environment with few real world steps, decreasing data cost. We will refer to this procedure as "Offline-Online SAC", as this SAC is first trained on offline data before transitioning to online data.

3.2.3 Dataset Aggregation (DAggr)

Instead of the strict transition from offline to online training in Offline-Online SAC, we explore interleaving offline and online training through a DAggr inspired weighting scheme. For our "offline" component in this variant, we explore the possibility of using a planning model to accelerate training. This model is a neural network trained to predict the responses of people to a proposed price, essentially a trained simulation of the rewards an agent would

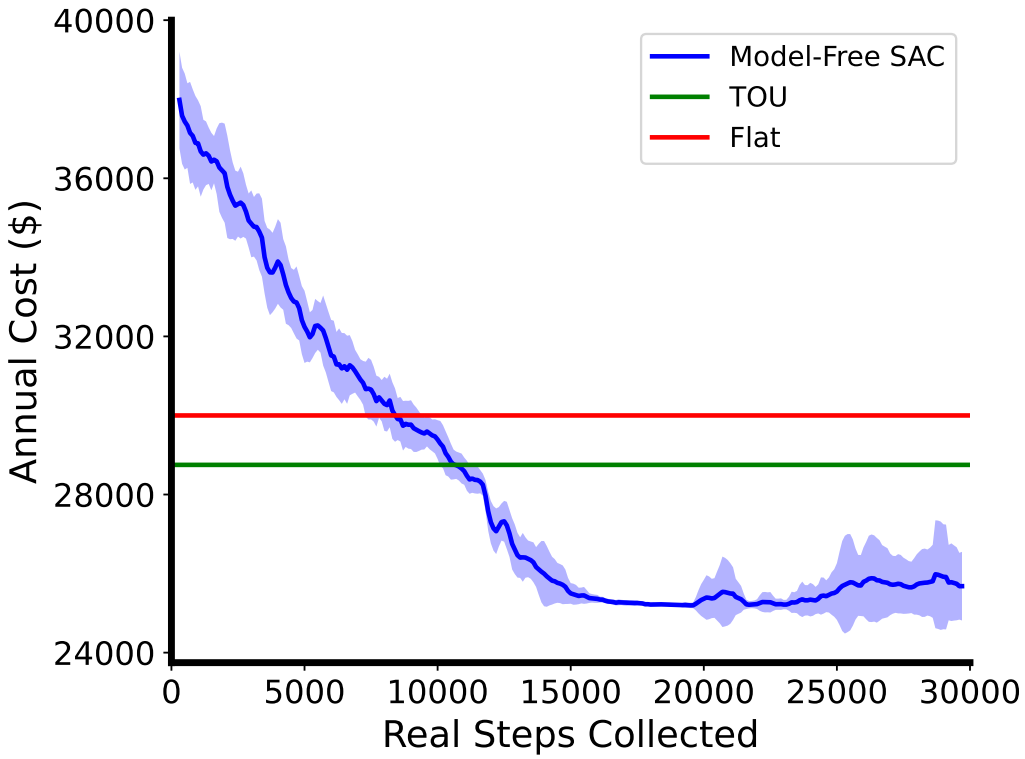


Figure 3.2: Vanilla RL price controller cost as training progresses, illustrating one vivid example of the "data hunger" noted in RL (Jang et al., 2021c).

receive given a state and an action in the real world. We use this planning model as our offline component here instead of the Social Game simulations from Section 2.3.3, because we believe training on data from two completely different distributions at the same time would not yield a model that learns efficiently for the real world task. We thus try to make our offline data source as close as possible to the online data. However, with a limited number of samples it is impossible to train a planning model $\mathcal{F}(s, a) : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ that exactly predicts the reward from the real environment, so we are still faced with the issue of having a source of training data that may not be aligned to the distribution of test data.

DAggr (Ross et al., 2011) is a meta-algorithm that helps solve the problem of distribution shift between training and test data in imitation learning. In the original paper, DAggr was used to solve the problem of training on one distribution (states reached by humans) and testing on another (states reached by the RL agent). Inspired by this, we attempt to adapt DAggr to bridge the gap between two distributions of training data: samples from the target environment, and samples from our planning model.

In order to mix data from the planning model and target environment, we employ a weighting strategy inspired by DAggr. We alternate training in the planning model and target environment, exponentially decaying the ratio of planning model steps as training continues. Our rationale is that our RL price controller should glean as much information as possible from the planning model first, since sampling from the planning model has negligible

cost compared to sampling steps from the target environment. Once the model has learned enough from the planning model, steps from the target environment are slowly introduced into the training dataset, ultimately producing a price controller that performs well on the target environment with fewer steps. In this way, we dynamically weight the two data sources for SAC for more efficient learning. For our experiments, we set the initial ratio of planning steps to target environment steps as $M = 10$, and exponential decay parameter $\beta = 0.99$. The algorithm for our data mixing procedure can be seen in Alg. 1. We refer to this training procedure as "DAggr SAC" since it interleaves online real world and offline planning model data to form an aggregated dataset for SAC to optimize the price controller. We will also refer to "Offline-DAggr SAC", which consists of employing DAggr SAC during Offline-0nline SAC's online portion. Though DAggr SAC does have an upfront data cost to train the planning model, our results show that this algorithm does ultimately decrease data cost by leveraging knowledge from the planning model.

Algorithm 1 Planning model $\mathcal{F}$ and target environment data mixing procedure

Initialize $D \longleftarrow 0$
Initialize π_{θ_1} to any policy in $\prod$
for $i = 1$ **to** N **do**
  Sample T-step trajectories using π_i
  for $j = 1$ **to** $\lfloor M_i \rfloor$ **do**
    Get dataset $D_{ij} = (s, \pi_{\theta_i}(s), r_{\mathcal{F}}(s, \pi_{\theta_i}(s)))$ of visited states and actions taken by π_{θ_i}, and rewards given by the planning model.
  end for
  Get dataset $D_{i0} = (s, \pi_{\theta_i}(s), r(s))$ of visited states and actions taken by π_i, and rewards given by the target environment.
  Aggregate datasets: $D \longleftarrow D \bigcup D_{i0}, D_{i1}...D_{iM}$
  Train policy $\pi_{\theta_{i+1}}$ on D
  Let $M_{i+1} = M_i * \beta$
end for

To test our hypothesis that Offline-Online SAC will enable faster adaptation to unfamiliar environments like a real-world Social Game, we pretrained SAC on several simpler models of simulated person response. We then evaluated how quickly SAC (starting from the pretrained weight initialization) can learn in an OfficeLearn environment with more complex models of simulated person response. We use "Curtail and Shift" office workers in place of real office workers. We argue that the transition from "Deterministic Function" workers to "Curtail and Shift" workers represents a similar step up in complexity as the transition from "Curtail and Shift" to workers in the real world. Training environments are instantiated by randomly sampling "Deterministic Function" response types and randomly sampling multipliers for how many "points" simulated humans received for reducing energy usage. The former sampling distribution is binomial with equal parts, and the latter is uniform between 0 and 10. Though the training environments used to train SAC had randomized parameters, the validation environments (with "Curtail and Shift" response types) were kept constant to ensure fairness. To ensure an accurate representation of each network's capabilities, we averaged the results from 5 different test trials and report the mean and standard error for each test. SAC is

trained with an ADAM optimizer with learning rate 3e-4, 0.9 β_1, and 0.999 β_2, where β_1 refers to the first moment and β_2 refers to the second. The offline dataset used to pretrain SAC was generated with 256,000 steps from each "Deterministic Function" type environment, for a total of 768,000 state transitions, evenly distributed among the three "Deterministic Function" response models, with a variety of randomized parameters. Our intention was to provide a wide, varied, and rich dataset of simplistic responses that would allow for our Offline-Online SAC model to learn the dynamics of a Social Game without overfitting to a specific model of human response to prices. Offline-Online SAC was pretrained on this dataset for approximately 15 epochs with an ADAM optimizer with the same parameters as above.

For the planning model used in DAggr SAC, we train a 4 layer neural network with 32 hidden units in each layer to predict people's energy usage given the hourly prices of energy for each day. The model is trained on 1000 randomly sampled state transitions from OfficeLearn with the 'Curtail and Shift' response type. The network was trained for 10,000 epochs with the ADAM optimizer with learning rate 0.001 and L2 regularization weight 0.001. Over the 10,000 epochs, the model with the lowest loss on a holdout validation set of 256 randomly sampled transitions was used for the rest of the experiment.

3.2.3.1 Results

We will now describe the results obtained from our pretraining and data aggregation approaches.

Offline-Online SAC Fig. 3.3 compares the performance of Offline-Online SAC, DAggr SAC, and Online SAC against our TOU and Flat Pricing baselines. In order to compare data costs, we define data cost here as the number of days' worth of data needed to train the price controller to beat the TOU baseline since it is the baseline with lowest energy cost.

First, note that the costs for TOU and the RL controllers shown in the figure are inflated by \$10,000 to account for the annual cost of running the Social Game (\$400 every two weeks for a 250 day business year.) So, if the figure at one point shows that the RL controllers cost \$30,000 per year, \$20,000 of it is the actual cost of the energy and \$10,000 is for Social Game incentives and logistics. This inflation does not occur for the Flat Pricing baseline since it would not make sense to run the Social Game for a flat price signal. Also note that each step in the simulated Social Game represents one day.

We observe that, for the first 4000 steps, Online SAC fails to beat the TOU and Flat Pricing baselines; while it makes significant progress toward learning a good policy, the learning speed is not enough to justify the cost of implementing it as a Social Game even with over a decade's worth of simulated training data. Our Offline-Online SAC, however, appears to have already learned a slightly better policy than TOU during its pretraining, with an effective data cost of 0 sampled steps. In contrast, Online SAC has a data cost of 8000 days (32 years). In addition, the model converges to a price controller that appears to provide over \$7000 in energy savings per year, with just 1000 days worth of simulated training data. The annual savings that Offline-Online SAC can provide clearly justify its implementation, even with the additional cost of running the Social Game. The success of the Offline-Online SAC model also suggests that, given a dataset of simplistic simulated models of human

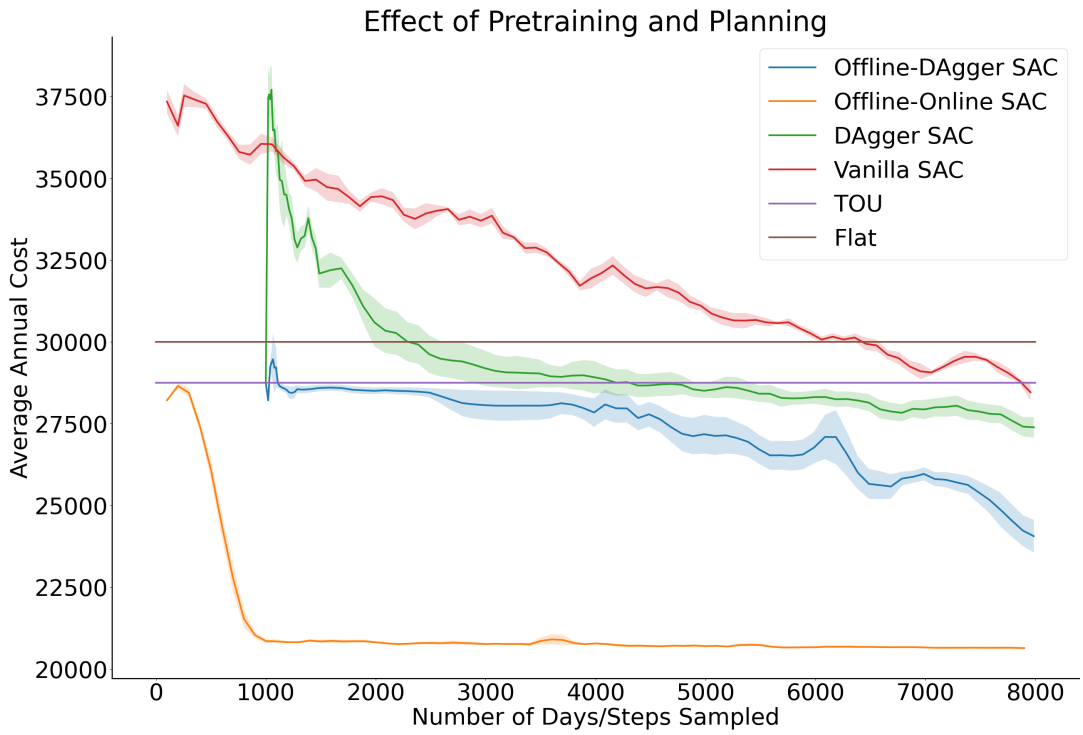


Figure 3.3: Offline-Online SAC and DAggr SAC Results. (Jang et al., 2021d)

Performance of Offline-Online SAC and DAggr SAC adapting to "Curtail and Shift" in comparison to Online SAC. We show the mean of 5 trials, with the standard error of the mean shaded.

behaviour, our price-setting model can learn helpful aspects of the price-setting problem that can warm-start it in a more complex environment. Our pretraining scheme appears to prepare the controller for steps up in complexity similar to what we might encounter transitioning from the simulation to the real world.

DAggr SAC The effect of DAggr SAC is less clear cut. We plot the cost of price controllers with models trained by the planning model 1000 steps to the right, to account for the up-front data cost of 1000 steps that must be collected to train the planning model in the first place. We assume during this planning model training period that TOU pricing is used as it is the cheapest baseline. As can be seen in Fig. 3.3, pretraining helps immensely in training the Online price controller, reducing the data cost by a factor of 2 compared to the Online controller without planning. On the other hand, the planning environment seems to slow down the training of Offline-DAggr SAC, performing only marginally better than TOU for the first 4000 days while Offline-Online SAC without the planning environment significantly diverges from TOU after just a few hundred steps.

Conclusion In conclusion, we found that vanilla offline training helped the performance of the algorithm by a significant amount. We believe that DAggr was not as performant, as it drew the agent away from local optima. Thus, we caution the reader to carefully consider distribution shift in their work.

3.2.4 Surprise Minimizing Reinforcement Learning

3.2.4.1 Background

Some stability of control is desirable, for both human subjects and the agent. Incentivizing the agent to minimize the surprise it experiences is equivalent to incentivizing it to minimize peoples' change in energy usage across days. This corresponds to adjusting people's habits in a stable system rather than forcing them to confront and attempt to understand an unstable one. Additionally, people behave predictably on aggregate, and thus choosing to minimize surprise may in fact make it easier for the agent to learn. In general, we define surprise minimization to be the reduction of novel events experienced by the RL agent. In this context, this equates to the RL controller experiencing stable energy demands by the office workers.

Surprise Minimizing Reinforcement Learning (SMiRL) is an algorithm that aims to reduce the entropy of visited states. SMiRL is useful when the environment provides sufficient unexpected and novel events for learning where the challenge for the agent is to maintain a steady equilibrium state (Berseth et al., 2019).

SMiRL maintains a distribution $P_{\pi_\theta(s)}$ about which states are likely under its current policy. The agent then modifies its policy π_θ so that it encounters states s with high $P_\theta(s)$, as well as to seek out states that will change the model $P_\theta(s)$ so that future states are more likely.

We make use of SMiRL as an auxiliary reward in addition to our usual reward to calculate a combined reward $r_{\text{combined}} = r_{\text{energy}} + \alpha r_{\text{SMiRL}}$. With a SMiRL weight α as a measure of how much the SMiRL reward influences the total reward. We will describe the explicit formulation of the SMiRL reward in Section 3.2.4.2.

3.2.4.2 Methods

Environment As a reminder to the reader, we are operating solely in the OfficeLearn environment. Please refer to Section 2.3.3 for precise definitions.

SMiRL Reward Formulation A SMiRL agent receives an auxiliary reward for experiencing familiar states based on an updating distribution of states it has experienced. This is exactly equivalent to learning a policy with the lowest entropy. Assuming we have a fully-observed Controlled Markov Process (CMP) with state s_t and action a_t at time t, and $P(s_0)$ as the initial state distribution, and transition probabilities $P(s_{t+1}|s_t, a_t)$, the agent learns a policy $\pi_\phi(a|s)$ parameterized by ϕ. As described earlier, we keep track of an estimated state marginal $\hat{P}_{\pi_{\theta_{t-1}}}(s_t)$ for the actual state marginal $P(s_{t+1}|s_t, a_t)$. As usual we denote

entropy of a state s_t by $\mathcal{H}(s_t)$. The entropy can then be calculated by the marginal as

$$\sum_{t=0}^{T} \mathcal{H}\left(\mathbf{s}_t\right) = -\sum_{t=0}^{T} \mathbb{E}_{\mathbf{s}_t \sim P(s_t, a_t)}\left[\log P\left(s_t, a_t\right)\right] \tag{3.1}$$

$$\leq -\sum_{t=0}^{T} \mathbb{E}_{\mathbf{s}_t \sim P(\mathbf{s}_t, a_t)}\left[\log \hat{P}_{\pi_{\theta_{t-1}}}\left(s_t, a_t\right)\right] \tag{3.2}$$

We bound Equation 3.1 by the entropy of an estimated marginal $\hat{P}_{\pi_{\theta_{t-1}}}$ in Equation 3.2. Minimizing the right side bound is then equivalent to maximizing an RL objective with reward $r_{\text{SMiRL}}(s_t) = \log p_{\theta_{t-1}}(s_t)$ We note that the optimal policy must also consider future changes to $p_{\theta_{t-1}}(s_t)$, as the distribution of visited states changes at each step. To account for the changing distribution of visited states, we use an augmented MDP that captures this notion (Berseth et al., 2019). We note that in our implementation of SMiRL, $p_{\theta_t}(s)$ is normally distributed. To construct the augmented MDP we include sufficient statistics for $p_{\theta_t}(s)$ in the state space such as the paramaeters of our normal distribution and the number of states seen so far.

SMiRL Implementation SMiRL is simply implemented in our existing OpenAI socialgame environment. We introduce SMiRL into our existing Social Game environment by initializing a buffer that tracks the agent's observed states and computes an estimated state marginal p_{θ_t}. As noted earlier, in the augmented MDP the state space also contains the number of observed states and this information is stored in the buffer as well. At each step in our simulation we add newly observed states to our buffer, and update p_{θ_t}. The agent then adjusts the its policy based on the combined reward.

We note that since $p_\theta(s)$ is modeled as an independent Gaussian for each dimension (hour) in the observation (consumption for a day), then the SMiRL reward is expressed as

$$r_{\text{SMiRL}}(s_t) = -\sum_i \left(\log \sigma_i + \frac{(s_i - \mu_i)^2}{2\sigma_i^2}\right) \tag{3.3}$$

where μ_i and σ_i are the sample mean and standard deviation from our state marginal and s_i is the i^{th} feature (i^{th} hour of day) of s (Berseth et al., 2019). With this formulation we can efficiently calculate the SMiRL reward from our buffer.

SMiRL as an Auxilary Reward We use SMiRL as an auxilary reward to provide faster learning and more stable outputs. We achieve this by calculating the SMiRL reward r_{SMiRL} as described in the previous section, and applying a SMiRL weight α to it and then using the sum with our usual energy reward $r_{\text{energy}} = \log(\vec{e}^{\intercal} p_{\text{util}})$. This gives us a combined reward of

$$r_{\text{combined}} = r_{\text{energy}} + \alpha r_{\text{SMiRL}} \tag{3.4}$$

and it is this reward that we use to train the RL agent. In our simulations, we found the optimal SMiRL weight α to be approximately $\alpha = 0.12$ after hyperparameter tuning. We will discuss the exact results of various SMiRL weights in Section 3.2.4.3.

3.2.4.3 Results

We now discuss our two aims: more effective learning and a more stable environment.

Observing Lower Sample Entropy As the primary objective of a surprise minimization technique is, in fact, to minimize surprise, it is natural to determine whether SMiRL does so.

We first quantify the degree to which the environment is more or less entropic when under the influence of an agent employing SMiRL. Since we assume each variable in our sample space is normally distributed in a given timeframe, we can compute its entropy as $\mathcal{H} = \frac{1}{2}\ln(2\pi e\sigma^2)$. For each step, we compute the sample variance of the last 100 steps and use this to compute a sample entropy, shown in Figure 3.4.

While both agents do reduce the sample entropy over time, the SMiRL + PPO agent does so earlier than the baseline PPO agent. The SMiRL + PPO implementation exhibits lower sample entropy for all iterations compared to baseline PPO, and hence energy usage over the course of the simulation is more consistent when using SMiRL.

Additionally, the SMiRL + PPO agent converges towards a policy which generates a more stable final environment than the baseline PPO agent. The SMiRL + PPO agent's behavior confirms our hypothesis that the stability of our environment is superior to our baseline PPO agent by incentivizing the RL agent to revisit familiar states. We also note that the both agents continue to explore changes to their policies after convergence, however the SMiRL + PPO maintains a lower sample entropy in late timesteps. Hence, the SMiRL + PPO agent explores in ways that maintain a more stable environment, while the PPO agent's exploration results in more unstable energy usage.

In this sense, the addition of the SMiRL reward can allow an agent to strike a balance between exploration and stability.

3.2.4.4 Discussion: Improved Learning with SMiRL (Compared to Baseline PPO)

While environment stability is important, it is essential that the agent still encourages efficient energy usage. We also wish to understand whether the SMiRL reward models our assumption of predictability in our system by improved learning speed.

Faster Learning and Consistent Outcomes We observe that our SMiRL + PPO implementation induces significantly faster learning and convergence to an equally optimal policy, with the same reward. As shown by Figure 3.4(a), both agents maintain similar rewards up until step $\sim 30k$, after which the SMiRL + PPO agent begins to achieve, on average, a higher reward. The SMiRL + PPO implementation converges in roughly half the time as the baseline PPO agent (step $\sim 50k$ v.s. step $\sim 110k$). The reward in the environment around step $\sim 60k$ whereas the PPO baseline does not converge to the maximum reward in the environment until step $\sim 110k$. Our results support the hypothesis that an auxiliary SMiRL reward improves learning speed.

Note that we compare the energy rewards (r_{energy}) here directly and not the combined reward which includes the SMiRL reward and weight. This demonstrates the influence of the SMiRL reward on the energy reward alone, which describes the overall effectiveness of

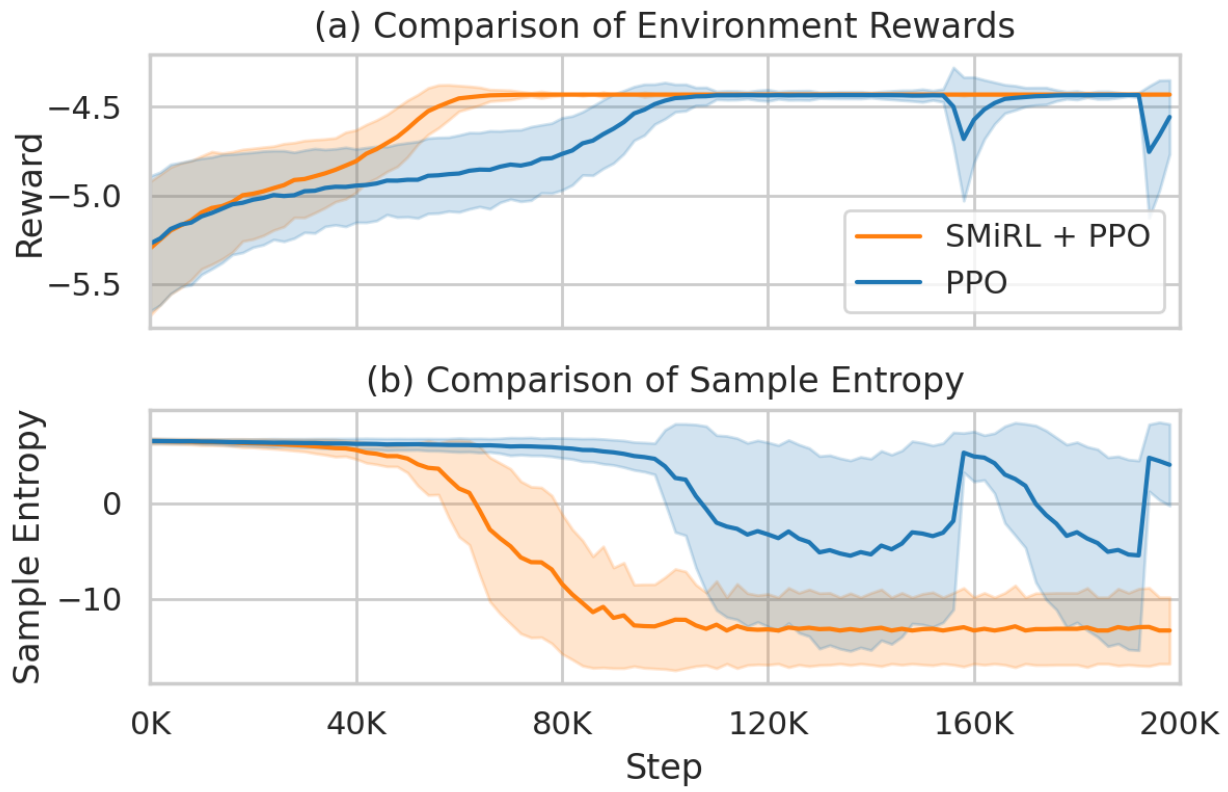


Figure 3.4: A comparison between the PPO + SMiRL agent and the baseline PPO agents' (a) rewards and (b) sample entropies over training steps. Shaded regions are one standard deviation of observations binned to every 100 steps (Arnold et al., 2021a).

our agent. Hence the inclusion of the SMiRL reward results in improved learning speed of our agent in the task it is given. Please see section 3.2.4.4 for a brief note on the seemingly adverse behavior in the SMiRL+PPO.

Faster convergence is demonstrated by Figure 3.5, which shows energy consumption throughout the day graphed at different training iterations (i.e., 10k, 40k, 80k, etc.) The grid price signal the agent receives is graphed as a dotted line in Figure 3.5. We observe that, for both the PPO and SMiRL + PPO agent, the agent ultimately learns a price signal that effectively shifts people's energy consumption away from hours when the grid price is higher. By step 40k, the SMiRL + PPO agent has already begun to greatly reduce consumption during peak, and people have shifted it towards just before that peak. By step 80k, the SMiRL + PPO agent has completely diminished any energy usage during the peak pricing, while the PPO agent only does so by step $\sim$ 120k (when its reward converges).

Optimal SMiRL Weights We note that while an appropriate SMiRL weight provides significant improvement in learning speed and sample entropy, an inadequate SMiRL weight can lead to poor learning that converges to a suboptimal result, and may not outperform a baseline PPO.

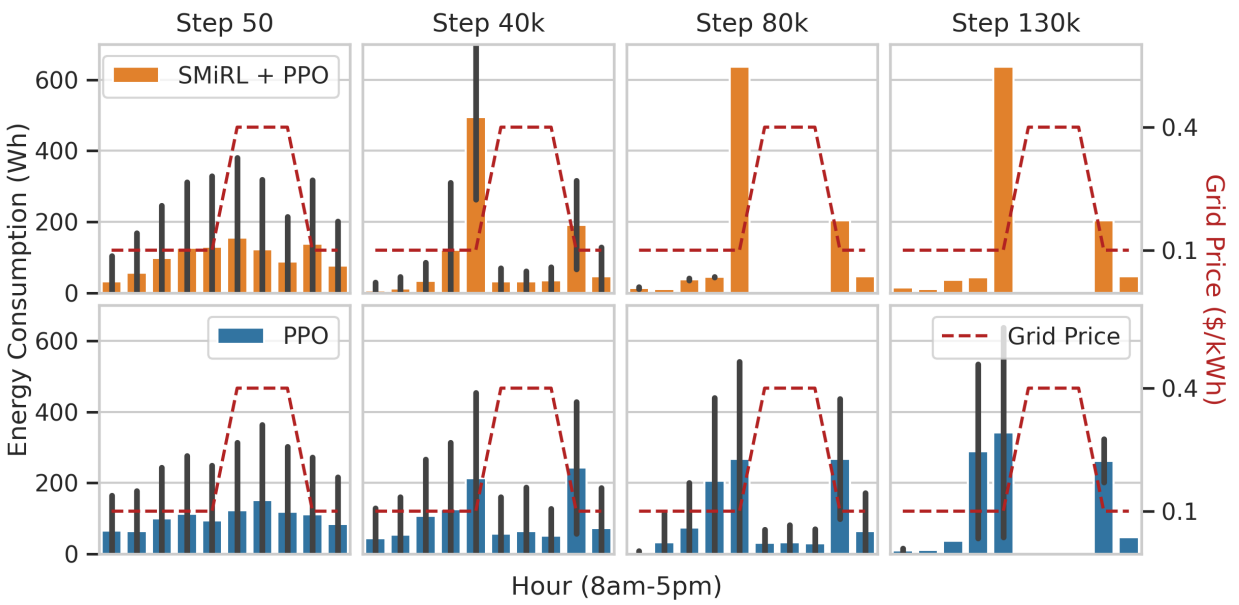


Figure 3.5: Energy consumption with the PPO and SMiRL + PPO agent at steps 10k, 40k, 80k, and 130k compared to the grid price (Arnold et al., 2021a).

Specifically, we found SMiRL weights of $\alpha = 0.25$ and higher performed worse than baseline PPO. In fact, we find that they converge to a suboptimal reward in the environment, hinting that too much surprise minimization might hinder exploration.

For much lower SMiRL weights such as $\alpha = 0.01$, we do not see any significant benefits when compared to baseline PPO; there possibly isn't enough weight on the SMiRL reward to have a meaningful impact.

Sample Entropy and Environment Reward Curves When comparing the sample entropy and reward curves in Figure 3.4 we see that beginning at steps $\sim 50k$, the SMiRL + PPO agent's observed sample entropy drops significantly below that of the baseline PPO agent. This correlates closely at $\sim 50k$ steps where the PPO + SMiRL implementation begins to experience significantly greater rewards than baseline PPO. Thus, a noisier environment, such as one which featured more house level wind turbines[26], may have been benefitted by SMiRL. Lastly, we observe that at $\sim 110k$ steps, both agents have observed a drop in sample entropy, and their environment rewards converge. This correlation between entropy and reward may support our hypothesis that aiming for a stable environment via surprise minimization can help the agent learn faster.

[26]The author has a tattoo of a wind turbine near his neck. If the reader sees this footnote, they are welcome to email and ask for a picture.

3.2.5 Pretraining Methods Under Consideration

For the sake of brevity, we will not detail every experiment that we made towards reducing dependence on in-environment training iterations. However, it is worth noting that you may find two techniques of interest, both of which use a planning model: (1) *extrinsic pre-training*, and (2) *intrinsic pre-training*. A *planning model* is a dynamics model of the environment; i.e. one that is capable of estimating transitions: $\mathcal{F}(s, a) = s_{t+1}$.

3.2.5.1 Extrinsic Pretraining

Background : The term *pretraining* is commonly used to refer to on-policy training in a simulation environment, and in *extrinsic* learning, an RL agent's reward explicitly encodes some of the programmer's desires. *Extrinsic pretraining* is thus the use of a planning model as a direct training environment for an agent, carrying over the environmental reward to guide the agent's search. Simply, this would look like training the agent in the simulation for M steps.

Method: In the following work, we will demonstrate the use of extrinsic pretraining in OfficeLearn. Constructing a good planning model for human behavior in the office – and specifically response to points – could provide a way for the algorithm to "imagine" the world's direct responses to its actions and explore more than it would otherwise. In our framework, the agent would step once in the office (i.e. deliver an action vector to the world and then observe the reward), and generate a vector of (state, action, next_state, done) that will contribute to a memory buffer specifically for office ("real world") observations. Next, stepping in the planning model space, the agent will deliver an action to the planning model, which predicts energy consumption. This tuple gets added to a separate planning memory buffer that the agent will train from. We experiment with:

- differing the batch size of buffer sampling
- whether the planning buffer has long term memory (never empties) or short term memory (empties after every step in the environment)
- a variety of planning model types.

We present four different planning models. Each model is trained on a simulated dataset of two months worth of energy.

- **NN**: Of various Automatic Machine Learning (AutoML) strategies, we present here the top performing Gaussian Optimization using GPy (GPyOpt) LSTM search (Knudde et al., 2017)
- **Ordinary Least Squares (OLS)**: We fit an OLS on the training dataset.
- **Oracle**: for a theoretical upper bound, we provide a model that calls the same function as the simulation would call. Thus, this model responds to the agent's actions in the same way the real world would.

- **Baseline**: For a theoretical lower bound, we test a model that only returns zero when queried. We hope that this model performs the worst.

Results: Comparison of Learning Curves. Here we compare the learning of agent with planning and those without: a full learning curve over 10k+ iterations for TOU and RTP regimes is shown in Figure 3.6, whereas performance over the first thirty days is shown in figure 3.7. Figure 3.6 confirms our hypothesis that the planning models indeed help the agent learn more quickly. On the left are the agents acting under a TOU price regime, and on the right are the agents acting in a RTP price regime. In both regimes, the agent's learning appears to be dramatically aided by the presence of a planning model. In Figure 3.7 the LSTM, labelled NN, generally outperforms the baseline model over the course of thirty days, and overperforms the OLS almost entirely. The performance of the LSTM is most pronounced when the batch size is lowest, implying that the lower the batch size, the more the agent's planning exploration is positively influenced by recent additions to buffer.

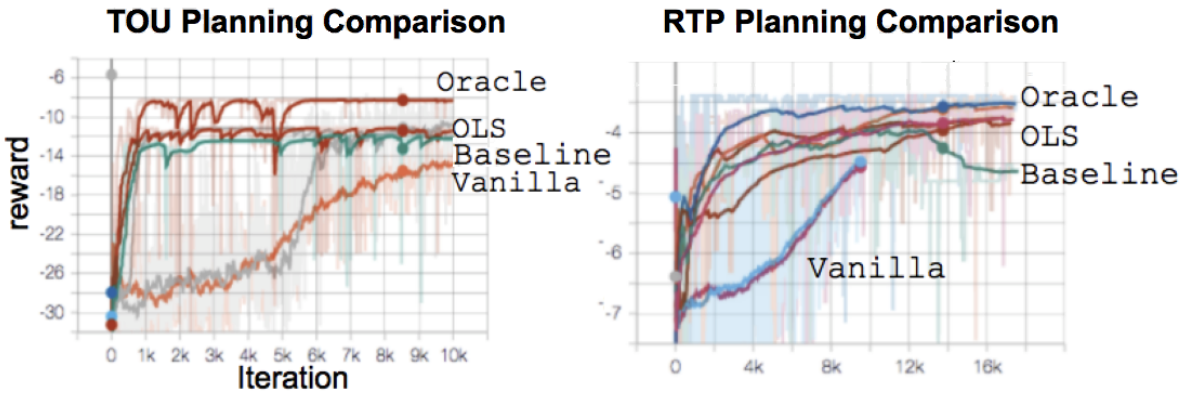


Figure 3.6: Comparison of the agents with and without planning (Spangher et al., 2020a).

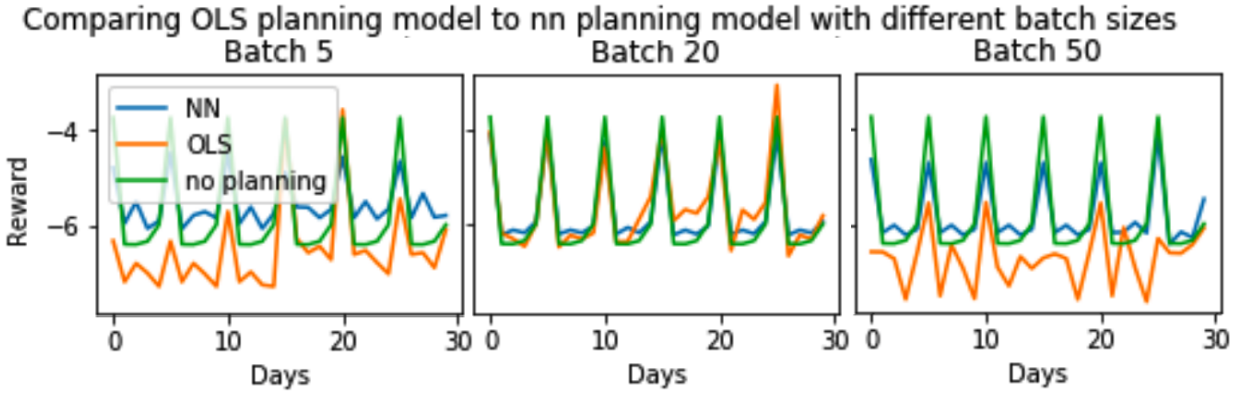


Figure 3.7: Effect of batch sizes in training when comparing the LSTM planning model to the OLS planning model, in an RTP pricing regime (Spangher et al., 2020a).

Results: Comparison of Short-Term vs. Long-Term Buffer Memory .

Following from that finding, we found that short term memory, i.e., learning produced when the memory buffer is emptied between every step in the environment, performed better than long term memory, shown in Figure 3.8. The result makes sense when considering that short term memory trains with examples that are more immediately relevant to the agent's next decision.

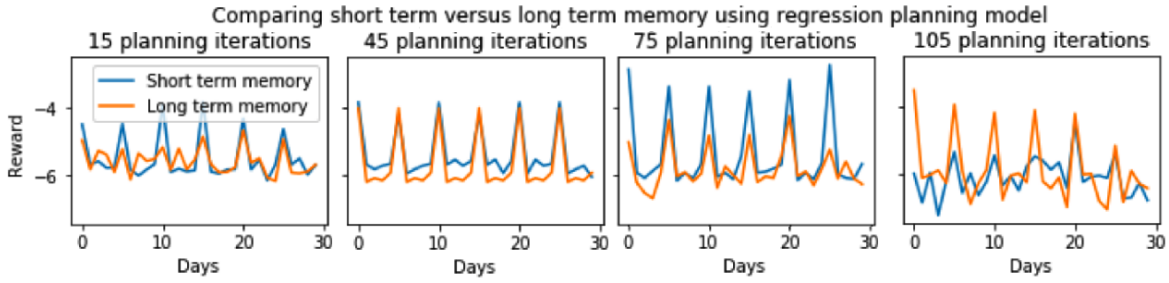


Figure 3.8: An exploration of the difference in memory buffer type. It is performed with the OLS because this required much lower compute than the LSTM predictions (Spangher et al., 2020a).

Discussion The results of this preliminary stop in our work are somewhat promising: we see that the machinery of the RL environment is working, and that planning models seem to have a generaly positive effect. While the OLS eventually helps the agent in lengthier training, it seems to hurt the RL agent's initial performance relative to the LSTM. From this, we assume that the planning model needs to be quite accurate in order to add to the simulation initially, but that when more iterations are possible, merely having a planning model present helps with training.

3.2.5.2 Intrinsic Pretraining

Background on Intrinsic Pretraining In contrast to extrinsic rewards, *intrinsic* rewards are ones which do not directly encode a programmer's desires. These include rewards such as ones that directly measure an agents' uncertainty in value prediction state by state. By pretraining with intrinsic rewards and then switching to extrinsic rewards in pretraining, researchers have shown that agents may be incentivized to explore states spaces and actions that they may not have explored if they were just optimizing for the same extrinsic reward. In a sense, cycling from intrinsic pretraining to extrinsic pretraining should lead to more well-rounded agents.

We believe that intrinsic motivation relies to heavily on the quality of the planning model. We also suspect that other methods guarantee some breadth of exploration in extrinsic motivation (i.e. SAC has a randomness parameter that rewards it for choosing an array of different actions) so encouraging exploration of uncertainty may already be built into state of the art methods. However, given that testing intrinsic motivation simply cycles out a different reward when a planning model is used, it is not difficult to implement.

We will now explain a brief series of experiments that we conducted to explore intrinsic motivation.

Experimental Methods To set up our experiments, we tested intrinsic motivation methods with three strategies: unsupervised exploration, unsupervised exploration, then RL, and RL with a so-called "curiosity" bonus[XXVI]. In order to capture uncertainty, we first created an planning model $mathcalF$ for our situation: an ensemble of twenty-four neural networks trained on data saved from a former run of the OfficeLearn simulation. The neural ensemble maps controller prices $\vec{p}_{\text{RL, t}}$ to baseline energy behavior $\vec{b}_t$, and so as such is not a traditional transition model, i.e. $f(s_t) = s_{t+1}$; part of our methods inquiry rests on a question too of whether this partial planning model is useful. We chose an ensemble of networks because of their distributional nature, and so we directly define six different intrinsic rewards that capture some aspect of uncertainty:

- **Mean of standard deviation of estimates**: The neural ensemble's twenty-four estimates can be summarized as a Gaussian distribution for hour of the day, i.e. $n(p_{\text{RL},1}) = N(\mu_{p_{\text{RL},1}}, \sigma_{p_{\text{RL},1}})$. We take the distribution of standard deviations and return the mean as the reward itself; i.e. $\overline{\sigma_{\vec{p}_{\text{RL}}}}$, the average spread in planning model estimate for that hour. We notice that this reward function behaves relatively well – the reward tends to converge after 400-500 steps.

- **Max of the std of estimates**: We return the max of the distribution of estimates; i.e. the direction along which the distributions produce the highest uncertainty. We notice that this reward function behaves poorly; it tends to get stuck on a single estimate.

- **L2 norm of the standard deviation of estimates**: We return the L2 norm of the standard deviation of estimates, i.e. $||\sum_{i=0}^{H} \sigma_i^2||_2$ and we see that this behaves very well, tending monotincally upwards in a concave manner.

- **Intrinsic + Extrinsic Rewards**: We return the mean of the standard deviation of estimates as an auxiliary reward to the extrinsic reward.

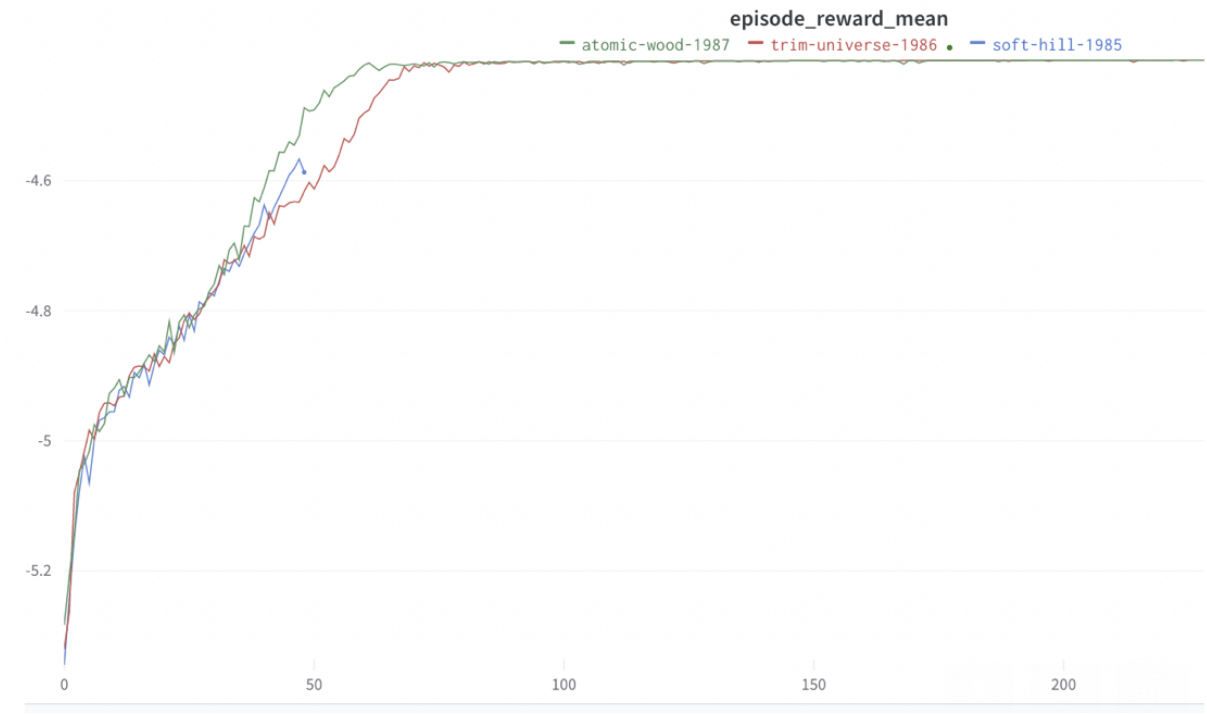


Figure 3.9: Baseline agent (red) compared to the L2 std norm (green) and Max std (blue). Un-published work.

- **Active Pre-Training (APT)**: The final method to add intrinsic motivation that we tested was APT. This consisted of the following strategy:

```
run model 10x
    take reward (1) 10 times
    adding unif(-.01, .01) for each hour's action
average these
```

Results: Best performing Intrinsic Methods vs. Baseline PPO : Please see Figure 3.9 for a reward plot for baseline PPO on our environment. From the results, it is apparent that most methods perform similarly to the baseline. One reason why this can be the case is that the exploration phase has not caused a diverse behavior to be learned. Another reason intrinsic motivation performs similarly to baseline is that the intrinsic motivation signal may be too weak for the agent to learn useful exploratory behavior.

3.3 Robustness and Safety

In many (but not all) real world cases, it may be very costly for an RL controller to make a mistake. For example, a self-driving car may crash, a power system controller may overload a transformer, or a Social Game controller may cause a worker to lose interest in the Game after the controller proposes some difficult prices.

Safety-critical cases pose a special challenge to RL as they eliminate the benefit of implicit understanding that an agent may bring to the situation after unfettered exploration. A first-cut solution may be that the programmer explicitly restricts actions or sequences of actions so that the agent does not reach dangerous parts of the state space. At best, this explicit action space pruning may force the agent into a local optimum as it converges to suboptimality in its reduced space. Worse, it may be difficult to actually determine ahead of time which sequence of actions may be dangerous, which may mean that the RL agent may still act unsafely despite certain action curtailment. Thus, it may be important to allow the agent some form of exploration despite safety concerns.

For safety-critical applications, we recommend pairing RL agent with a planning model of the environment. These may be used for the techniques previously noted: extrinsic and intrinsic pretraining, both of which help agents reap the benefits of exploration during training "for free" with respect to real-world training costs. However, simply using a heuristic in pretraining, such as forcing the model to step one time in the real world and 10 times in the planning model, does not directly guarantee that the actions emitted to the real world are safe. If safety is the goal, planning models may be used in more targeted ways.

We present a methodological innovation we call *guardrails* as our best recommendation of how to use a planning model to. Our *guardrails* technique: (1) weeds out unsafe actions, (2) decides when to explore or not, (3) allows the agent to learn from both a planning model and the real world. We are the first to propose the method, and we believe that it can be helpful in combining the insight of a planning model with optimal control. The main methodological contribution of the *guardrails* method is that it intelligently decides when to safely propagate an action into the world and helps the agent continue to "explore" when it is not safe.

3.3.1 Planning *Guardrails* for Risk-Aware Reinforcement Learning

3.3.1.1 Background

Notions of incorporating risk into action selection have been considered in RL. For instance, (Prashanth et al., 2022) review work that accounts for a risk-sensitive objective function, such as a conditional value at risk. Similarly, the explicit management of risk arising from unknown consumer behavior in online learning is also fairly recent (Chen et al., 2020). In our work, we consider the risk of the RL agent's price leading to higher energy cost than the baseline energy cost that would result from the passing on of a readily available TOU price signal. Our use of guardrails provides a new mechanism for managing the risk of posting poor prices in the real world that does not involve directly modifying the RL objective to be risk sensitive, which has the benefit of allowing us to continue to leverage established RL algorithms such as SAC.

Planning models are one way in which risk consideration is implemented. Methodologically, planning models used as surrogate environments for agents to train in has existed as a concept for a while (Sutton, 1991a), but creative ways in which they can be used are just now beginning to emerge. For example, planning models have gotten attention in complex domains like autonomous driving (Hoel et al., 2019) and robotic manipulation (Wang et al., 2019a) by providing alternate environments for agents to learn in. Only recently have fields emerged like "curiosity" driven learning, which make use of a planning model for uncertainty quantification (Frank et al., 2014). For example, (Lakshminarayanan et al., 2017) employ a neural network ensemble to quantify uncertainty, and (Kidambi et al., 2020) use a planning model that uses this uncertainty quantification method to minimize the risk of training an RL agent where the model is uncertain. To the best of our knowledge, the *guardrails* approach of using an ensemble planning model for detecting a risky action and rejecting its use in the *real world* is new. By employing guardrails to switch between real and planning steps, we can: (1) significantly reduce overall real-world data collection, (2) minimize the cost of the real-world data that is collected, and (3) achieve greater final reward.

3.3.1.2 Methods

We base the following work in the OfficeLearn environment.

Planning Model Creation We assume the availability of limited offline data, possibly from applying demand response at a secondary building. We train a planning model that transforms a price signal to a predicted energy consumption of the N workers using this offline data. In the context of the RL problem, the planning model can be used to estimate the reward associated with an action. We implement our planning model as an ensemble of neural networks, because neural networks are expressive at representing nonlinear functions (Goodfellow et al., 2016) but known to have high variance. We therefore consider an ensemble of J neural networks. The planning model represented by the j-th neural network is $\mathbb{E}_j^{\text{NN}}(p_{\text{RL},t})$, which again predicts the energy consumption of the U workers. We will use $\mathbb{E}_{j,i}^{\text{NN}}(p_{\text{RL},t})$ to denote the hourly energy consumption vector of worker i. Then, the energy cost for the j-th NN's energy consumption prediction is:

$$c_j^{\text{NN}}(p_{\text{RL},t}) := \sum_{i=1}^{N} p_t^\intercal \mathbb{E}_{j,i}^{\text{NN}}(p_{\text{RL},t}) \tag{3.5}$$

The consumption response of this ensemble planning model is the energy consumption d_i of each worker i across the N neural networks, i.e.:

$$d_i = (1/N) \sum_{i=1}^{N} \mathbb{E}_{j,i}^{\text{NN}}(p_{\text{RL},t}) \tag{3.6}$$

The expected energy cost associated with this mean consumption is:

$$\mathbb{E}^{\text{NN}}[c(p_{\text{RL},t})] := (1/N) \sum_{j=1}^{J} \sum_{i=1}^{U} p_t^\intercal \mathbb{E}_{j,i}^{\text{NN}}(p_{\text{RL},t}) \tag{3.7}$$

We will refer to Equation 3.7 henceforth as mean predicted cost. This can easily be converted to an estimate of the reward since we defined r_t as:

$$r_t := -\log\left[c(p_{\mathrm{RL},t})\right] \tag{3.8}$$

Guardrails Method Since some prices $p_{\mathrm{RL},t}$ could be more expensive to actually try in the real world than the readily available TOU price, can we use our planning model to estimate which prices are at "risk" to be worse than the TOU price and filter them out before they are sent to the real world? Here we will describe how to implement "guardrails" that use the planning model to identify prices that could result in high energy costs and query the planning model with these prices instead of the real world. Let $\mathbb{W}(a_t)$ be an indicator function expressing whether action a_t is accepted in the real world, and $P_{\mathbb{W}}(p_{\mathrm{RL},t}) \in [0,1]$ denote the probability of posting a price $p_{\mathrm{RL},t}$ to the real world to obtain a real energy cost and $1 - P_{\mathbb{W}}(p_{\mathrm{RL},t})$ be the probability of posting this price to the planning model to obtain predicted energy cost. Figure 3.10 is a schematic of our approach. Model-free RL corresponds to $P_{\mathbb{W}}(p_{\mathrm{RL},t}) \equiv 1$. The idea of guardrails is to define $P_{\mathbb{W}}(p_{\mathrm{RL},t})$ such that prices that result in high energy costs are unlikely to be posted in the real world. We define guardrails relative to the baseline energy cost associated with sending the TOU price signal to the workers. We denote this TOU cost by c^{TOU}. Let $\mathbb{1}[\cdot]$ represent an indicator function that evaluates to 1 if its argument is true and to 0 otherwise. The guardrail definition we will highlight here is the Quantile approach. At a high level, we compute N estimates of the expected cost of our price signal with a total of N neural networks in our planning model ensemble, and only send the price signal to the real world if a certain number of the N estimates agree the estimated cost is low enough.

$$\text{Quantile}_\alpha : P_{\mathbb{W}}(p_{\mathrm{RL},t}) = \mathbb{1}[n(p_{\mathrm{RL},t}) < \alpha N]$$

Here, α belongs in the interval $[0,1]$ and $n(p_{\mathrm{RL},t})$ is the number of neural networks in the ensemble with predicted cost $c_j^{\mathrm{NN}}(p_{\mathrm{RL},t})$ greater than the TOU cost c^{TOU}. To illustrate, suppose that α is 0.5. Then a price p_t is passed on to the real world only if 50% of the ensemble's N NNs predict a cost for this price that is less than c^{TOU}. We refer to this strategy as "Quantile" because it checks if the α^{th} quantile of the predicted costs is less than c^{TOU}. As α increases it is less likely that a price will be played in the real world.

Other rules we considered are defined in Supplementary Material in Section A.2.1. The efficacy of a guardrail depends on at least two factors: (1) the quality of its assessment that a price would have higher cost than the TOU price signal, and (2) if there is useful information in the planning model feedback to guide policy learning when such risky prices are posted.

Please see algorithm 2 detailing the algorithm. Also, please see figure 3.10 for a schematic of how the process would work.

Summary These components together represent a risk-aware deep RL approach for price setting. When used in conjunction with SAC, we refer to the resulting methodology as Risk-Aware Soft Actor Critic (RA-SAC). To sum up the flow of information (illustrated in Figure 3.10), the SAC RL agent first proposes a price vector, which is sent to the planning model to test for risk. Our "guardrail" method takes in the output of the planning model

Algorithm 2 Risk-Aware Soft Actor Critic (RA-SAC)

Input:
π_ψ: $s_t \to a_t$ Policy with parameters ψ.
$\mathcal{F}$: $(s_t, a_t) \to E_t$ Planning model, here a NN ensemble, predicting env response E_t,
G: $\mathcal{F}(a_t) \to P_{\mathbb{W}}(a_t)$ Guardrail rule giving probably of action being sent to World.
W: $(s_t, a_t) \to E_t$ "Real world"; i.e. the true response function to E_t.
$\mathbb{W}$: Indicator if a_t is accepted in the world.
Output:
π_ψ: Fully trained policy.
D: Dataset of $(s_t,\ a_t,\ s_{t+1},\ r_t)$ tuples
for $t \in 1, \ldots, T$ **do**
  Observe state s_t and select action $a_t \sim \pi_\psi(\cdot|s_t)$
  Predict $\mathbb{E}^{\text{NN}}(a_t)$ according to each NN $\in \mathcal{F}$, i.e., $(\mathbb{E}_1^{\text{NN}}(a_t),\ \mathbb{E}_2^{\text{NN}}(a_t),\ ...,\ \mathbb{E}_N^{\text{NN}}(a_t))$
  Resolve $P_{\mathbb{W}}(a_t)$ according to G
  Sample $\mathbb{W} \sim \mathcal{B}\left[P_{\mathbb{W}}(a_t)\right]$
  if $\mathbb{W} = 1$ **then** ▷ Evaluate action a_t in the real world
    Compute $E_t \leftarrow W(a_t)$. Derive $(s_{t+1},\ r_t)$ from E_t.
    Store $(s_t,\ a_t,\ s_{t+1},\ r_t)\ \in D$
  else if $\mathbb{W} = 0$ **then** ▷ Evaluate action $p_{\text{RL, t}}$ in the planning model
    Compute $E_t \leftarrow \mathcal{F}(a_t)$. Derive $(s_{t+1},\ r_t)$ from E_t.
    Store $(s_t,\ a_t,\ s_{t+1},\ r_t) \in D$
  end if
  if it is time to update **then**:
    Perform a SAC training update on ψ as per (Haarnoja et al., 2018) with D
  end if
end for

and determines whether the price vector is low or high risk. If it is low risk, we post the price vector to the "real" environment, where we collect a reward by observing the energy consumption of office workers in response to the price vector. If it is high risk, the reward is instead calculated from the planning model, and the price vector is never posted to the real environment. For pseudocode summarizing our approach, please see Algorithm 2.

3.3.1.3 Results

We compare RA-SAC with model-free and model-based benchmarks in simulation. The setup of our experiments is described in Section 3.3.2. Results and insights are discussed in sections 3.3.2 and 3.3.2, respectively.

3.3.2 Experimental Setup

In order to evaluate our approaches to guiding RL price controller training, we simulate a social game composed of 500 office workers. These simulated workers will compete for points in the social game by shifting their energy consumption behavior. We assume a homogeneous

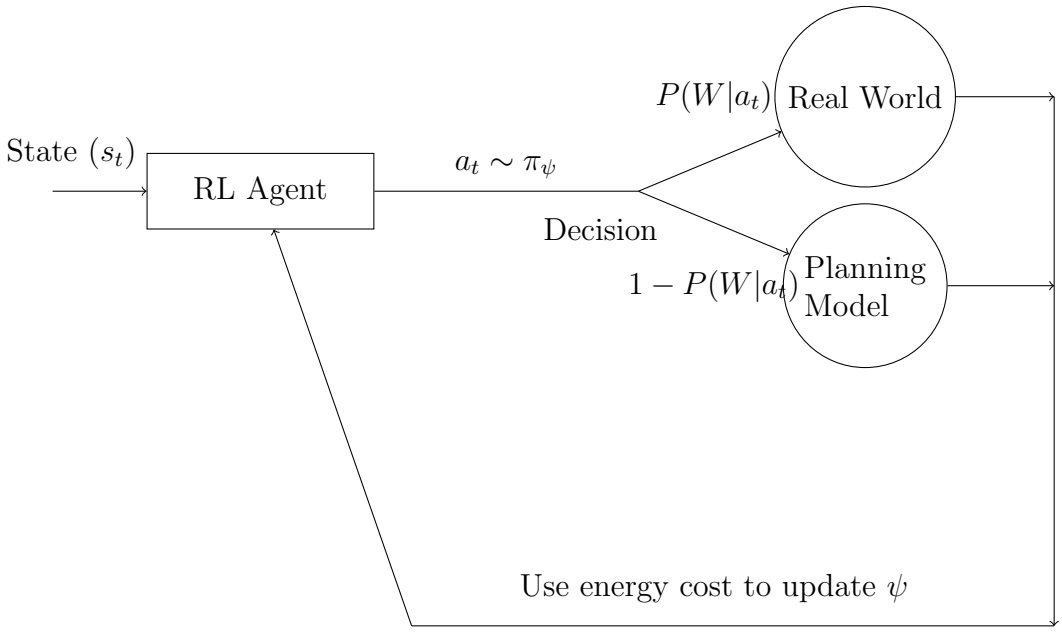


Figure 3.10: A ball-and-stick schematic detailing the flow of information and decisions in the RL with guardrails approach.

environment where each worker exhibits a Curtail and Shift response as discussed in Section 2.3.3. We assume a 250 day work year, a 5 day work week, and a 10 hour work day.

We setup the three components of RA-SAC as follows:

(i)**Model-free RL.** We use an out of the box implementation of SAC from RLLib with standard hyperparameter tuning. We chose the state space to include only the TOU grid prices [XXVII]. We replaced the cost function $c(s_t, p_{\mathrm{RL},t})$ by

$$\log\left[\sum_i p^{\intercal}_{\mathrm{util},t} d_i\right] + \lambda * \mathbb{1}\left[\sum_i d_i > \frac{\sum_i b_i}{2}\right]. \tag{3.9}$$

We use the logarithm of the energy cost as RL tends to learn with more stability when the reward signal is closer to zero. Experiments without this logarithmic term did not perform as well. In equation 3.9, the term $\lambda * \mathbb{1}\left[\sum_i d_i > \frac{\sum_i b_i}{2}\right]$ acts as a regularizer to encourage the controller not to reduce energy uniformly across the board[XXVIII]. In practice, we set λ to 10.

(ii) **Planning Model.** We use an ensemble of 20 fully connected networks, each with 5 hidden layers and 64 units in each hidden layer, batch normalized, with ReLU activations. Each network was randomly initialized, and the output of the ensemble is just the average of the outputs of all the networks. We train the planning model using an offline data set generated by sampling energy prices from a log normal distribution and then feeding them to the Curtail and Shift model to obtain energy consumption response data.

(iii) **Guardrails.** We considered the guardrails discussed in Section 3.3.1.2.

We use static grid prices from the PG&E Spring quarter for each simulation. We assume the energy consumption of our building is not large enough to impact the outside grid's energy prices.

We assume that the RL agents are trained using the building manager's personal computer. This is reasonable since it takes an entire day to collect one training sample of data, and data is not collected during the night. The computational time of training the agent is negligible compared to data collection time. No experiment described in this paper took longer than 1.5 hours on an Nvidia GTX 3060 and Intel i7-11700k, of which much of the overhead was CPU-bound environment simulation.

Value of TOU Pricing The cost of running the social game infrastructure as described above for an office of 500 workers is approximately $40 per day (see 3.3.2). Our simulated "Curtail and Shift" people use about $120 per day on energy without any social game (and its incentives), and $75 per day with TOU pricing. Since using the TOU pricing requires the incentive to be in place, the total cost of TOU pricing is $115 USD. Thus, there is an effective $5 per day savings with TOU, which yields a return of $1250 over 1 year and $2500 over two years. We will compare our RL pricing strategy against this baseline. This finding suggests that the building demand response Social Game adds value even without an RL agent by incentivizing workers to adjust their consumption based on the TOU price signal. Moreover, there is no deployment cost related to data since the TOU price signal is readily available. Nevertheless, the savings of $5 , which is a 4% reduction over the daily energy cost under a flat price, is likely insufficient incentive for the building manager to administer the social game. This motivates the exploration of an RL pricing agent.

Value of RL Based Pricing We investigate the design of an RL pricing agent that would increase the value of the social game, and if such an agent would provide sufficient incentive for a building manager to adopt a social game based building demand response program.

Model-Free SAC We begin by assessing the energy cost savings of an off-the-shelf and state-of-the-art SAC RL agent is deployed, henceforth model-free SAC.

We use the SAC implementation from Reinforcement Learning Library (RLLib) (Liang et al., 2018) with RLLib's default neural network architectures, batch size of 256, a learning rate of 0.0001, and starting learning immediately. These parameters were used for every experiment we report for the SAC price controller. Since we use a stochastic price controller, we run each of our experiments 5 times and report the standard deviation as error bands.

Model-Based SAC Since learning from the real world is costly, we evaluate a SAC agent, henceforth model-based SAC, that posts prices to and learns from only the neural network ensemble planning model (see Section 3.3.2). This model is trained using an offline dataset of 500 days' worth of data. The model-based SAC agent is evaluated on the real world environment for 256 days. This model-based SAC was only able to achieve an average cost of $114 per day (including the $40 per day in incentives). This is approximately the same performance as the $115 cost of TOU pricing.

Model-based SAC can be seen as substituting costly real world data with "free" data from the planning model. This helps reduce deployment costs substantially compared to model-free SAC. However, this fully supervised SAC adds essentially zero value relative to a TOU pricing signal, which is discouraging. Practically, we are back in square one, that is,

when using either a TOU pricing signal (see section 3.3.2) or a model-based SAC agent, there isn't sufficient value for the building manager to administer a demand response social game.

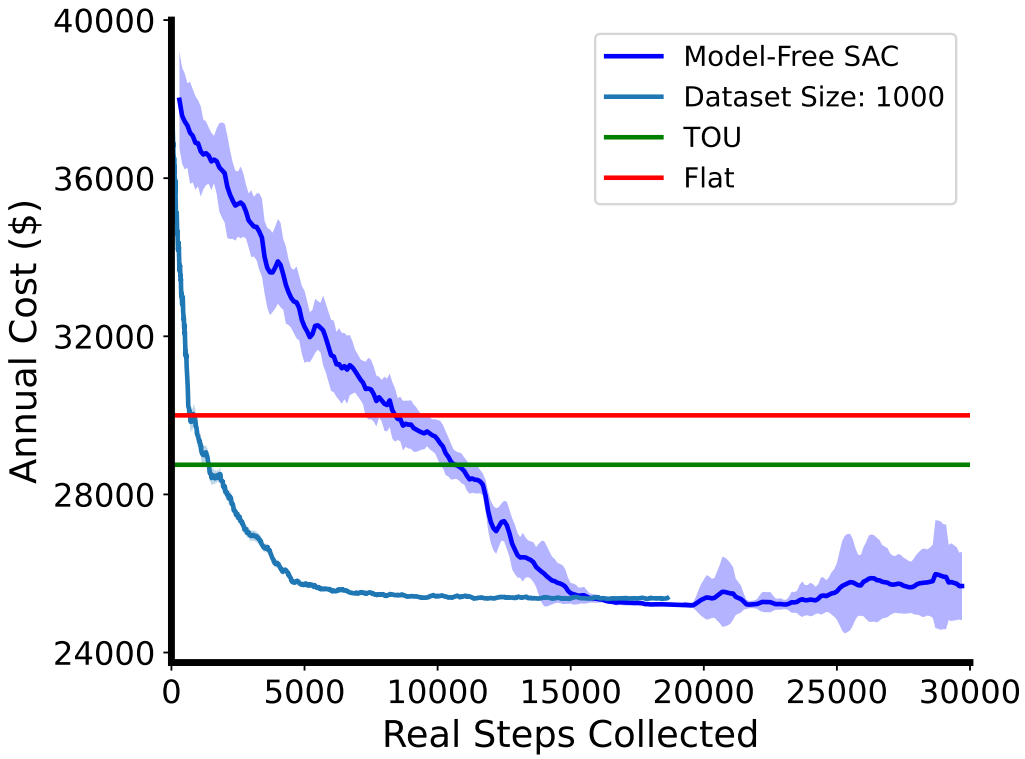


Figure 3.11: RL price controller with neural network ensemble guardrails (with a "hard" trigger) as training progresses (Jang et al., 2022a).

Risk-Aware Soft Actor Critic (RA-SAC) We begin by discussing RA-SAC with a $\text{Quantile}_{\alpha=50\%}$ guardrail based on an ensemble of neural networks (see Section 3.3.1.2). This is arguably the simplest specification of RA-SAC. Figure 3.11 shows that this specification of RA-SAC with a planning model trained using just 1000 steps of data is enough to provide substantial savings over model-free SAC. In particular, it beats TOU in 1200 days compared to model-free SAC, which takes 10,000. This is an 8x speedup in learning, which translates to \$127,000 in savings in initial deployment cost (\$48,000) compared to model-free SAC (\$175,000). Put differently, it would take 93 days for a fully trained RA-SAC to recoup deployment costs instead of about 18 years taken by model-free SAC. The total energy cost after 5 years would still be \$14,000 more than TOU, but this is a substantial improvement over the model-free SAC which is worse by \$37,500.

Figure 3.12.A compares the performance of RA-SAC with different guardrail strategies using neural network ensemble planning models. Most guardrails, defined in the endotes[XXIX], appear to have similar performance, except for the quantile based guardrails of section 3.3.1.2. This is likely because the other guardrails are similar in that they let all actions with predicted cost (the mean of the ensemble's predictions) lower than c^{tou} into the real world and gradually

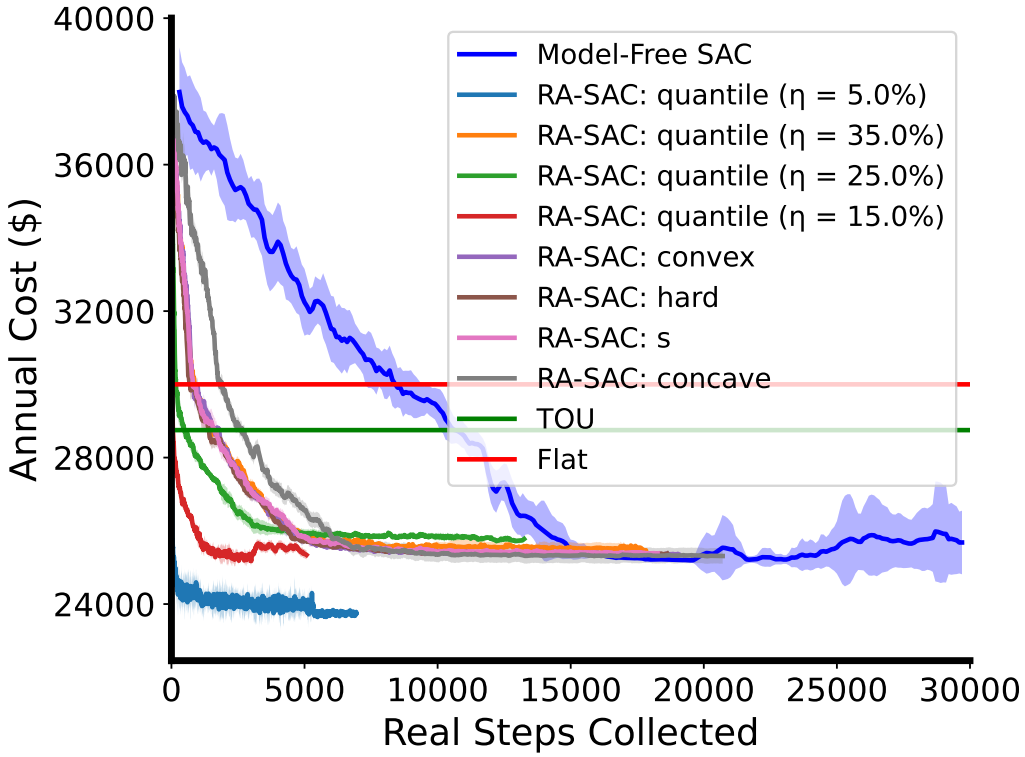


Figure 3.12: The costs of RA-SAC with neural network ensemble planning models trained on 1000 steps of training data, with different guardrail strategies (Jang et al., 2022a).

increase the probability of switching to the planning model as the predicted cost increases. On the other hand, quantile guardrails with lower α are much more conservative; all 20 neural networks in the ensemble predicting lower cost than c^{tou} is less likely than their mean doing so.

Notably, RA-SAC with a $\text{Quantile}_{\alpha=5\%}$ guardrail can substantially reduce the amount of early mistakes even over the RA-SAC with a hard guardrail (hard is equivalent to $\text{Quantile}_{\alpha=50\%}$). The initial deployment cost is cut to less than \$200 with a neural ensemble trained using 1000 samples, a reduction of 99.89% compared to the initial deployment cost of model-free SAC. Further, this version of RA-SAC appears to have identified a pricing strategy that is more effective than model-free SAC, costing \$24,000 per year instead of \$25,000. This represents an annual return of \$5,500 compared to Flat pricing and \$4250 compared to TOU. The building manager would only take 17 days to recoup initial deployment cost compared to 93 days when using a trained RA-SAC with hard guardrails. After 5 years, one would expect to see a return of \$45,500 **more** than TOU in energy savings. RA-SAC with a quantile guardrail provides significant value over Flat and TOU pricing signals quickly: a tempting incentive for building managers to implement social-game based demand response in their own buildings.

3.3.2.1 Conclusion

In sum, we have seen startlingly good performance gains based on how we propose to use the planning model. Thus, we recommend further study and use of this technique across the literature.

3.4 Finding Global Optimality and Transferring it from Simulation

3.4.1 Introduction to the Problem

Oftentimes, an RL agent may have found an near-global optimum in simulation that is very local in reality due to the variety of stochastic factors in the world.

We recommend methods on the environment-side to address this issue. The first option we propose is the use of meta-RL in domain randomization. We will demonstrate this inside OfficeLearn (please see Section 2.3.3.)

3.4.2 Meta RL with Domain Randomization

3.4.2.1 Background

Model Agnostic Meta-Learning (MAML) Model Agnostic Meta-Learning (MAML) ((Finn et al., 2017a)) is a meta-learning algorithm that seeks to learn a NN weight initialization θ that allows for fast adaptation to different types of tasks. We first define a task as a transition distribution $P(s_{t+1}|s_t, a_t)$ and a reward function r_i. MAML can be implemented over any model. In the RL variant, given a set of tasks $\mathcal{T}$ and a distribution of those tasks $P(\mathcal{T})$ MAML seeks to find an initialization θ that can achieve as high a reward as possible after K gradient update steps in any task. To maximize reward after K updates, MAML optimizes a policy network, π_ϕ for K gradient update steps starting from the learned initialization θ and computes the following meta loss:

$$L_{\eta_i}(\pi_\theta) = -\mathbb{E}_{s_t, a_t \sim \pi_\phi, P_{\eta_i}(s_{t+1}|s_t, a_t)}[\sum_{t=1}^{K} R_i(s_t, a_t)] \tag{3.10}$$

MAML then optimizes θ by gradient descent on the sum of the meta-losses over all tasks, computing parameters θ that produce good results after just K update steps. Training MAML on PPO in a task distribution can be described as consisting of an inner adaptation step per task and an outer, meta adaptation step between tasks. In the inner step, PPO is trained on a task η_i, randomly chosen according to $P(\mathcal{T})$; i.e. $\eta \sim \mathcal{T}$ for K PPO update steps, starting from the weight initialization MAML is optimizing. In the outer step, the weight initialization is updated according to the training trajectories of PPO that were collected in the inner adaptation steps. These outer and inner adaptation steps are alternated to compute the weight initialization that produces the highest reward after PPO is trained for K steps on a task from the task distribution η. We observe in this paper the performance of a trained MAML+PPO weight initialization on several test tasks that are outside of the training task distribution $\mathcal{T}$ but are still within our simulated Social Game setting.

We hypothesize that MAML applied to a PPO agent learning over different hyperparameter initializations of OfficeLearn will be able to warm-start the agent for better adaptation to new environments. We adapt MAML to the problem of optimizing a price-setting agent in OfficeLearn.

3.4.2.2 Methods

To test our hypothesis that MAML + PPO will enable faster adaptation to unfamiliar environments like a real-world Social Game, we trained MAML on several models of simulated person response. We then evaluated how quickly PPO, starting from the MAML weight initialization, can learn in an OfficeLearn environment with different models of simulated person response.

The training environments had randomized "Deterministic Function" response types and multipliers for how many "points" simulated humans received for reducing energy usage. Though the training environments used to train MAML had randomized parameters, the validation environments were kept constant to ensure fairness. To ensure an accurate representation of each network's capabilities, we averaged the results from 5 different test trials and report the mean and standard error for each test. MAML is trained with an ADAM optimizer with learning rate 0.0001, 0.9 β_1, and 0.999 β_2, where β_1 refers to the first moment and β_2 refers to the second. MAML sampled approximately eight PPO training trajectories in (parallel) training environments at a time between (sequential) meta-update steps. We trained PPO for 5 steps at a time in MAML's inner adaptation phase and trained MAML once per trajectory sampling. We trained MAML for up to 200 iterations. For evaluation, we trained PPO for 100 steps from the MAML-learned weight initialization and from a random initialization and compared the two. PPO (in both MAML+PPO and PPO) is trained with the clipped surrogate loss with a clipping parameter of 0.3 and a Stochastic Gradient Descent (SGD) optimizer with learning rate 0.01. The action and value estimators shared layers in the neural network for our implementation, and the value loss had a weighting of 0.5. We present three different experiments:

Adaptation to "Curtail and Shift" Response We train MAML on a distribution of environments with three simple models of human response to price, i.e. the three "Deterministic Functions", and see if it can adapt to a more complex model that we believe may be more representative of real-world behavior, i.e. "Curtail and Shift" function. Here, testing was done over 5 different test trials "Curtail and Shift" with different hyperparameter instantiations. Please see Figure 3.13 for a depiction of the experiments.

Adaptation to Threshold Exponential Response In this experiment, we train MAML on a distribution of environments with two different models of human response to price instead of three: "Deterministic Function" with linear or sinusoidal responses. Each variant environment, along with other parameters, is created by sampling environment hyperparameters at random within training task parameter distributions. We test this trained model's ability to adapt to an environment with a different model of human response: "Deterministic Function" Person with threshold exponential response, with all 5 validation runs of MAML+PPO and PPO.

MAML Training Behavior We observe the performance of MAML + PPO in the *Adaptation to Curtail and Shift Response* task after different numbers of MAML training iterations. When training MAML, we saved checkpoints of the meta-optimized weight

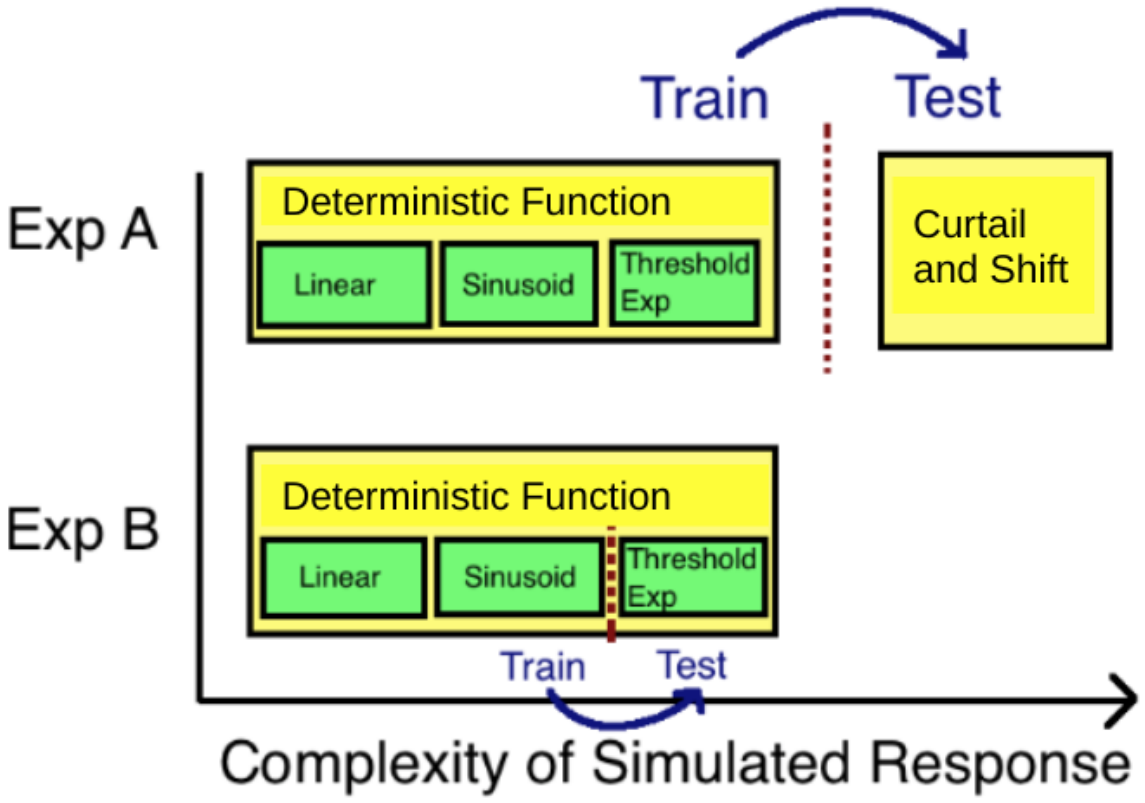


Figure 3.13: Setup of testing frameworks to simulate steps up in complexity.

initialization at 50, 100, 150, and 200 MAML iterations and observed the performance of PPO starting from each of those warm starts.

3.4.2.3 Results

We will now describe the results obtained from using MAML.

MAML Results (Compared to Baseline) As can be seen in Fig. 3.14.A and B, MAML + PPO significantly outperforms baseline PPO from the start to the finish of training. In both adaptation experiments, MAML+PPO converges to a solution with higher reward, and thus lower simulated energy cost than PPO. Note that PPO appears to have plateaued with constant reward. MAML decreases total energy cost by about 40% in both environments compared to PPO by the end of training on average. It takes fewer than two PPO steps worth of training data for MAML+PPO to outperform PPO after training for 100 steps, which suggests that the use of MAML with PPO to warm start learning from our simulation was successful. Using MAML allowed us to find a weight initialization that generalized to more complex tasks, thus decreasing the amount of data necessary to train the controller.

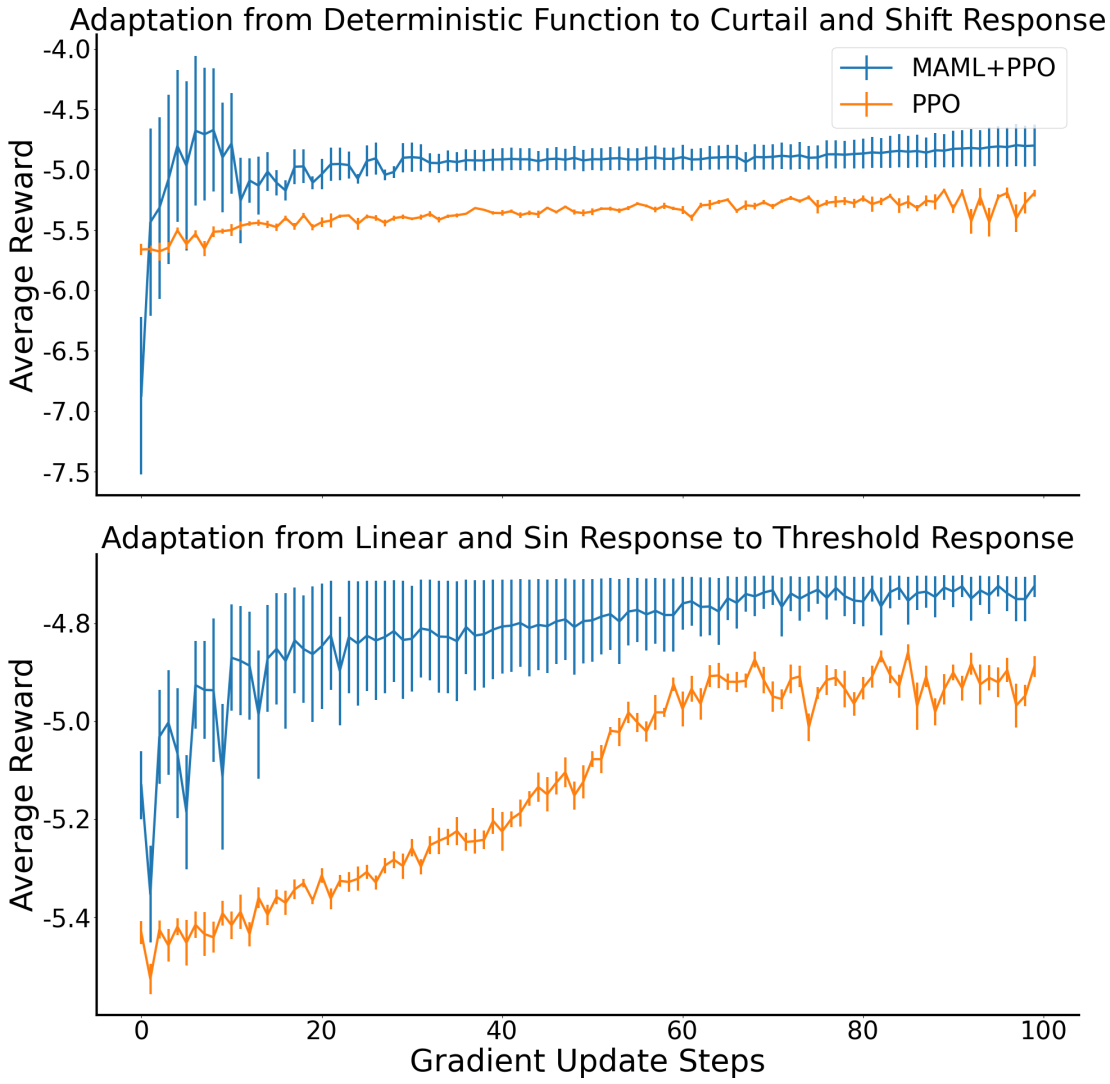


Figure 3.14: MAML+PPO Results (Jang et al., 2021a).

(Above) Performance of MAML+PPO in training on "Deterministic Function", and adapting to a more complex model, "Curtail and Shift" in comparison to PPO. (Below) Performance of MAML+PPO in a lateral shift in complexity, i.e. adapting to a different model of human behavior, "Deterministic Threshold Exponential" in comparison to PPO.

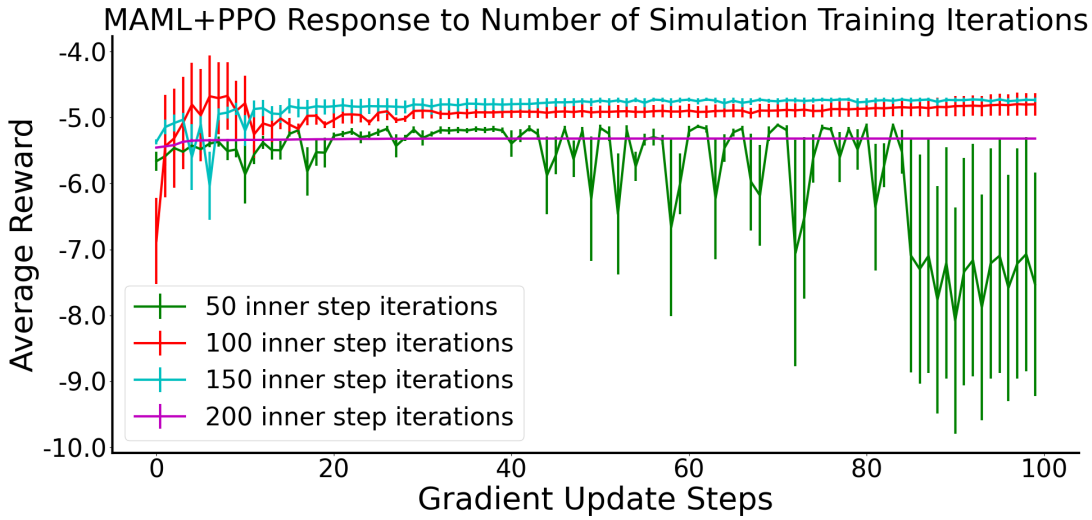


Figure 3.15: MAML Reward vs Update Steps (Jang et al., 2021a).

Performance of MAML+PPO on "Curtail and Shift" task after different number of iterations of MAML in simulation.

As we are interested in both the reward achieved by each algorithm and how quickly the reward is achieved, we chose to plot the training graphs of the validation runs of each checkpoint in Fig. 3.15 instead of simply plotting the total reward of each validation run. We observe that performance appears to increase with number of MAML training steps up until 200 iterations, where MAML appears to have overfit to a local minimum. Training MAML for too few steps, for example only 50 iterations, appears to increase instability in the PPO training on the MAML weight initialization.

3.4.2.4 Discussion and Conclusion

Meta-learning and K-shot learning are fields of machine learning focused on applying knowledge obtained from one domain to another. This makes techniques like MAML especially helpful in taking advantage of simulations, where data is plentiful, to increase performance in the real world, where data is costly. Whereas other RL tasks like robotic grasping can use highly accurate physics simulations to train a model and successfully employ the same model to real world equivalent tasks, the human response to changes in electricity price is not as well studied as physics. The models of human behavior we employ in simulations are simplistic and likely inaccurate with regard to real human behaviour. However, the novelty of our experiment is in showing how MAML might be applied to transfer knowledge acquired from training in simulation to accelerating adaptation in the real world, by demonstrating that MAML allows fast adaptation to different models of human responses to energy price. From our results, we are hopeful that MAML+PPO can be applied in our real world price controller. By training MAML+PPO in an environment different from the test environment, we were

able to affirm that the technique can be used to adapt training in simulation to a different task in simulation. The jump in complexity between the "Deterministic Function" response the model was trained on and the "Curtail and Shift" response shows that MAML+PPO can be used to adapt to tasks that are outside of its training task distribution and are more complex. We believe that this ability for MAML to close gaps in complexity has a high probability of enabling a MAML optimized model to work as a real world RL price controller.

3.4.3 Environment Search

We conclude our discussion of Meta RL under domain randomization with a proposition: if simply randomizing the parameters behind automatic environment selection provided a gain in performance, can intelligently selecting the parameters for successive environments do even better?

We believe the question warrants further merit, and we believe that we are the first to propose using environment auto-curricula techniques to prepare the agent for sim-to-real transfer. We will discuss this idea further in Future Work.

3.5 Adversarial Attacks

Energy grids are known to be lucrative targets for cyberattacks (e.g., (Kshetri and Voas, 2017)). Our work investigates the robustness of an AI-based microgrid controller to malicious actors. We present a novel attack that enables a few compromised microgrid controllers to adversely affect the behavior of connected controllers by *poisoning the data* on which it is trained. Our adversarial work expands on a recent explosion of interest in adversarial attacks (Madry et al., 2018; Rakhsha et al., 2021; Goodfellow et al., 2014). We pair our attack with a gradient-based defense that eliminates the threat of our attack.

More concretely, we examine MicrogridLearn: a setting in which a network of microgrid controllers collect supply and demand data that are continually aggregated by a central agent. The agent uses RL to optimize its profits. In our attack, a few microgrid controllers are compromised by a malicious adversary. The adversary applies a perturbation to the collected data, severely impacting the provider and *the entire network* of controllers. The provider is made to operate at a loss, and all prosumers are made to pay higher energy costs, use their batteries less, and violate more transformer power constraints.

Our work is set against a backdrop of developments in energy grid control that hold both promise and peril: RL-based controllers allow for sophisticated control in unprecedented granularity. Yet, we must be careful to minimize risk enabled by the opaque nature of deep learning. Our attack stands out in its subtlety and its scope. Other forms of large-scale interference such as blackouts and line disruptions are, by definition, easily detectable and local. Yet our attack causes harm by interfering with the agent's learning, and may not be detected until significant financial damage has been incurred. Furthermore, by interfering with the central agent's learning, our methods can damage systems that are physically disconnected from the energy grid under attack.

Outline In Section 3.5.1 we briefly contextualize our work within RL and energy controls. In Section 3.5.2 we describe the threat model, attack, and defense. In Section 3.5.2.4 we introduce our experimental setup, which is used in Section 3.5.2.5 to evaluate the efficacy of our attack and defense in an energy grid environment. Finally, in Section 5.2.6 we discuss limitations and future work.

3.5.1 Background: Adversarial Attacks on RL for Prosumer Energy Pricing

The literature on adversarial attacks for RL in demand response focuses on *responding* to prices (Wan et al., 2021) rather than *setting* them. To our knowledge, there are no works on adversarial attacks on dynamic price setting for demand response.

3.5.2 A Novel Adversarial Attack and a Defense Against It

3.5.2.1 Threat Model

In our setting, N controllers continuously collect data to be aggregated by a centralized agent. Learning takes place over multiple *iterations*; in each iteration, each controller collects a

trajectory $\tau \sim= (s_i, a_i, r_i)_i$ collected according to the agent policy π_θ. The agent's policy π_θ is described by a neural network. Nodes are required to feed observations through π_θ so as to collect policy-specified actions (pricing schemes), so we assume that the network parameters θ and architecture are shared with the controllers.

The attacker's power is determined by a fraction of *corrupted controllers* $\varepsilon \in (0, 1)$, and a *perturbation bound* $\rho > 0$, as follows: An attacker controls $\varepsilon \cdot N$ of controllers. The attacker *perturbs* the trajectories collected by each compromised controller, causing it to report back a trajectory $\widetilde{\tau}$ instead of the collected trajectory τ. Crucially, these perturbations are of small norm, that is,

$$\|\widetilde{\tau} - \tau\|_\infty \leq \rho$$

for some *perturbation bound* $\rho > 0$. Note that our attacker adheres to the suggested policy π_θ, but lies about the result to the agent.

Remark 1 *In our setting, the attacker may only perturb the actions of each trajectory. Observations and rewards remain unperturbed, because such perturbations would be expensive or easily noticed. This is in contrast to previous work in RL poisoning in which only rewards are poisoned (Rakhsha et al., 2021).*

3.5.2.2 The Attack

At a high level, our attack aims to perturb each trajectory to reverse the direction of the estimated gradient $\nabla_\theta f(\theta)$. Let θ be the parameters of the agent's policy, τ_ι be the unperturbed set of compromised trajectories (the trajectories collected by compromised controllers), $\tilde{\tau}_\iota$ be the set of perturbed adversarial trajectories (reported back to the agent), and $\tau_{\mathbb{H}}$ be the set of honest trajectories (unaffected by the adversary). Our adversary minimizes the correlation of the gradient post-perturbation with the honest one by solving the following constrained optimization problem:

$$\begin{aligned} \min_{\tilde{\tau}_\iota} \quad & \langle \nabla_\theta f_\theta \left(\tilde{\tau}_\iota\right), \nabla_\theta \left(f_\theta(\tau_\iota) + f_\theta(\tau_{\mathbb{H}})\right) \rangle \\ \text{such that} \quad & ||\tilde{\tau}_\iota - \tau_\iota||_\infty \leq \rho. \end{aligned} \tag{3.11}$$

Since the compromised controllers report $\tilde{\tau}_\iota$ to the agent instead of τ_ι, the agent will take gradient steps according to $\nabla_\theta \left(f_\theta(\tilde{\tau}_\iota) + f_\theta(\tau_{\mathbb{H}})\right)$. Therefore, choosing $\tilde{\tau}_\iota$ to minimize Equation (3.11) should maximally mislead the gradient towards a sub-optimal policy. Equation (3.11) is optimized by the adversary using the Fast Gradient Sign Method (FGSM) (Goodfellow et al., 2014). Interestingly, we find that our adversaries can obtain nearly identical results by solving Equation (3.11) without the $\tau_{\mathbb{H}}$ term, meaning that the adversary does not require any information about the honest (uncompromised) controllers.

3.5.2.3 The Defense

We propose a defense to protect an RL agent from the attack described in Section 3.5.2.2. Our defense works by identifying and removing the trajectories which have the largest influence on the gradient from the training data. Intuitively, this defense works because the honest trajectories are not expected to have out-sized gradients. Note that the poisoned trajectories

are not easily identifiable without calculating the gradient through the policy; while the adversarial perturbations significantly influence the gradient estimate, the perturbations themselves are small. More formally, if the RL agent suspects that some fraction $\hat{\epsilon}$ of the microgrids are adversarially controlled, then, when estimating the gradient $\nabla_\theta f(\theta)$, it ignores the $\hat{\epsilon}$-fraction of trajectories τ with largest $||\nabla_\theta f_\theta(\tau)||_2$.

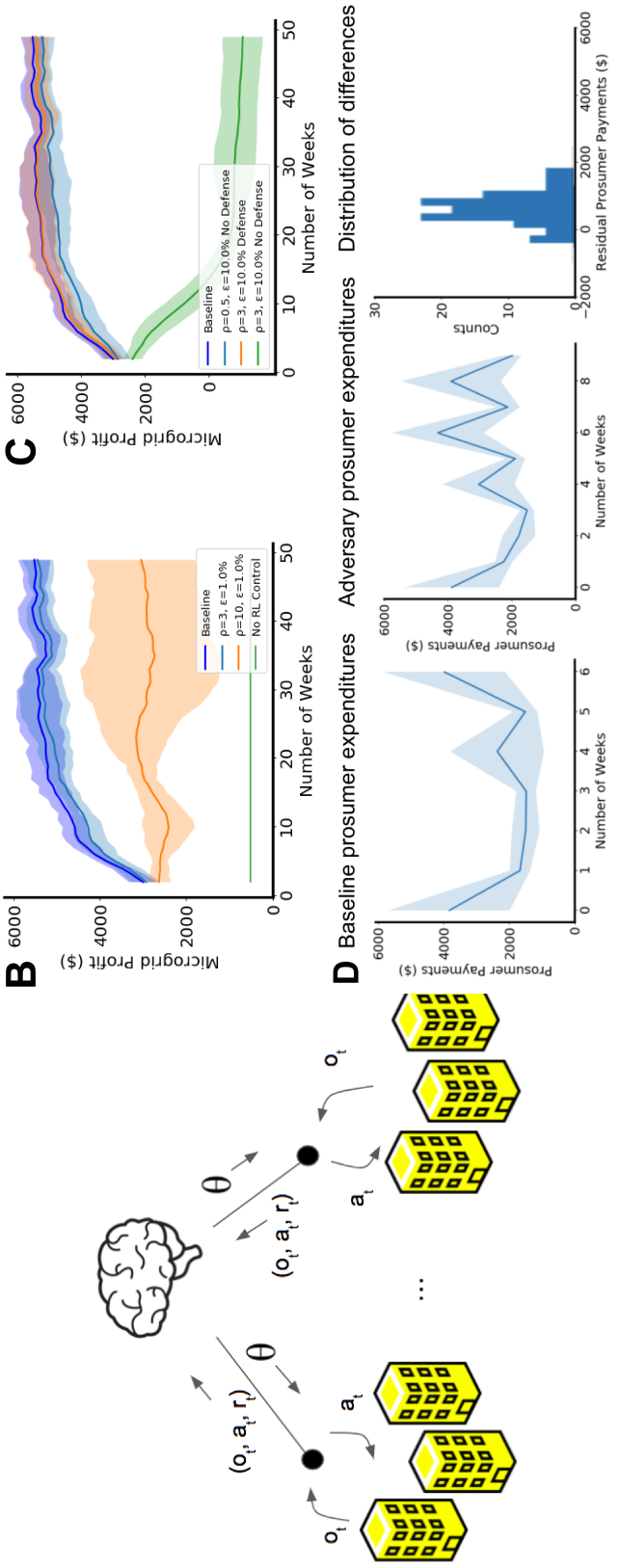


Figure 3.16: **A.** A description of the microgrid environment, reproduced from above. In this figure, the brain is the RL agent, the black dot is the microgrid controller, and the adversary attacks the a_t that is sent back to the RL agent. **B.** Effect of the adversary on the agent's learning. Note that $\epsilon = 1\%$ corresponds to only one adversarial microgrid. **C.** Effect of our defense in the presence of an adversary. **D.** Characterization of prosumer costs in the baseline and adversarial scenarios. The prosumer consistently pays more in energy when the adversary interferes (Gunn et al., [n. d.]).

3.5.2.4 Experimental Setup

We use a multi instance of MicrogridLearn to demonstrate this attack.

Adversarial Microgrid Poisoning "In the Wild" We briefly present a potential real-world example of our adversary in action.

Suppose that Eastern Gas and Electric (EG&E) is piloting a dynamic, local pricing program. To do this, EG&E instantiates an RL agent to train across a sample of building clusters (i.e. microgrids grouped locally). Unfortunately, there is an attacker who wishes to disrupt the functioning of EG&E, and they intercept the outflow of data from one of the local microgrid controllers. In one attack strategy, the attacker wishes to minimize the extent to which the outgoing prices are perturbed so as to escape detection. In another attack strategy, the attacker considers high perturbations in order to maximally disrupt profitability.

3.5.2.5 Results

Next, we present experimental results demonstrating the gradient-reversing adversary's harmful potential, as well as the efficacy of the filtering defense.

All of our experiments used the MicrogridLearn environment (Agwan et al., 2021b) consisting of 100 microgrids of 7 buildings each. The RL agent is an Actor-Critic agent which updates every week over the course of one year.

The Attack Figure 3.16.B shows our attacker can significantly hinder the RL agent's learning by co-opting a single microgrid controller. The maximal difference between successive actions taken by the true policy is around 6, so the strongest attack in the single-trajectory setting requires a relatively high perturbation budget $\rho = 10$. However in Figure 3.16.C, our attack utilizes a smaller perturbation budget of $\rho = 3$ with ten ($\varepsilon = 10\%$) compromised controllers to achieve significant damage.

The Defense We find that our defense recovers the original performance of the RL agent, even under generous ϵ and ρ. See Figure 3.16.C.

3.5.3 Characterizations of Environmental Response

We investigated several ways in which the environment responded to adversarial attack beyond the sheer profit: individual prosumer energy costs (i.e. the sum of the building's energy expenditures in an environment with the adversary and without), battery utilization (i.e. the number of times batteries were charged and discharged, and the total capacities) and transformer power constraint violations. According to all measures, the environment performed worse with an adversary, even those that were not directly targeted: the prosumers paid on average *more* for the energy, the battery was used *less* when the microgrid controller was adversarially perturbed, and transformer power constraints were violated more. We present the prosumer prices in Figure 3.18 and omit the rest due to space constraints.

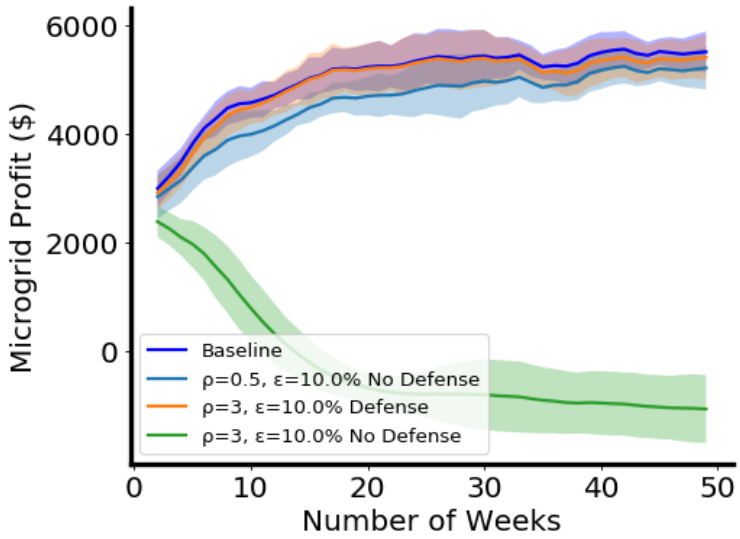


Figure 3.17: Effect of our defense in the presence of an adversary (Gunn et al., [n. d.]).

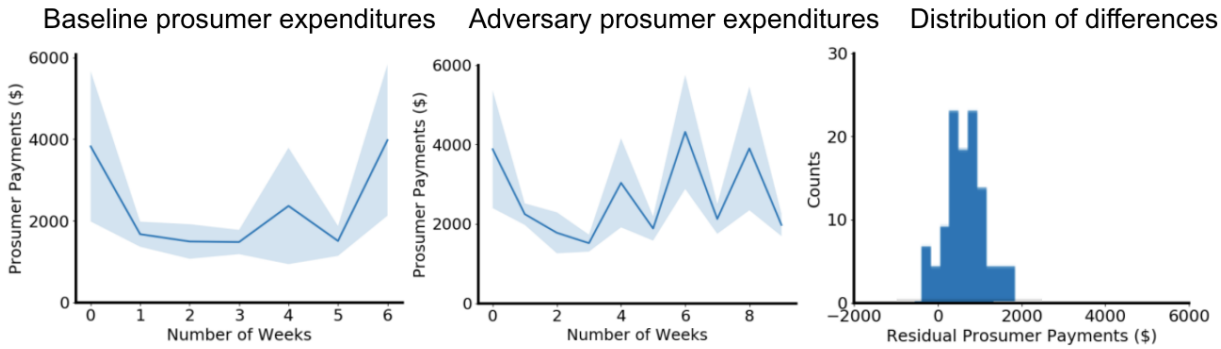


Figure 3.18: Characterization of prosumer costs in the baseline and adversarial scenarios. The prosumer consistently pays more in energy when the adversary interferes (Gunn et al., [n. d.]).

3.6 Privacy Preservation and Generalization to New Subdomains

3.6.1 Introduction to the Problem

As RL methods are applied to real-world problems and novel environments, challenges arise that are not present in classical RL environments. Two challenges and desiderata we will now focus on are: *privacy preservation* and *a simultaneous objective*, often competing, of generalizations to new tasks.

3.6.1.1 Privacy Preservation

The **preservation of privacy** within local *tasks*[27] of the environment is an important challenge in RL. Real applications at scale may require privacy guarantees which are not provided by modern multi-agent RL algorithms, as they may need to be trained on privileged or corporate data (Lowe et al., 2017; Sunehag et al., 2017; Rashid et al., 2018). Indeed, any app that personalizes an RL agent to individual users must take care to protect their privacy by not storing all their data in a central server.

We present a hypothetical setting in which one may imagine privacy demands of our application of interest: energy. One may imagine a hypothetical company, CovertAI, that is concluding a multi-month endeavor to train their new 80 quintillion parameter language model, CPT-4. Now imagine that a malicious hacker is listening to the energy consumption of CovertAI's datacenter, and is able to infer when more of their computing cores are active. From this, the hacker can tell when CovertAI is performing model training, and can thus conclude the right moment to intefere with training. This may come in the form of sabotaging power lines at the right moment to erase learning gains or launching programs on certain cores that would scramble their datasets (i.e, please see the previous section on adversarial attacks.) This could result in CovertAI's model training to be incorrect or incomplete.

Separately, we can imagine attacks on residential neighborhoods. In this setting, hackers might observe the energy consumption of domestic buildings to figure out when people are not home. They could time a theft for maximal effectiveness; or they they could disaggregate energy signals to learn the appliances the homeowner has or glean sensitive health information if medical devices produce noticeable patterns in energy consumption.

Classical RL is fundamentally ill-equipped to deal with the problem of privacy preservation, as it aggregates data into a central, hackable server. Not only would keeping data of buildings' energy consumption at one central location present a major privacy concern if this central machine is compromised, but message passing of the raw data could present an additional source of vulnerability. Although each microgrid might have access to the data of a few buildings at a time, the scale of damage would be much larger if data was stored in a central server across multiple microgrids.

[27]For the purpose of this writing, we define "tasks" to be a local group, goal, or objective in an environment that may contain many heterogenuous tasks.

3.6.1.2 Generalization to New Tasks

One may imagine that as privacy of unique tasks is increased, the ability to learn from prior tasks is necessarily limited: the most privacy preserving setup may be one in which an agent "starts fresh" each task. Thus, we pair the real world demand of privacy with another demand of learning another task, hoping to use the simultaneous use of prior subdomains to provide a jump on learning in new subdomains.

Real world applications are always likely to feature heterogeneous tasks; every user, robot, energy system will have different traits that cannot be accounted for by "one size fits all" algorithms. As previous work in privacy-preserving RL (Qi et al., 2021; Wang et al., 2020b; Ren et al., 2019; Anwar and Raychowdhury, 2021) does not extend to personalized models, the competing goals of privacy and personalization must be accomplished at the other's expense.

We relate these two characteristics inextricably in the solution, thus we present them together in a single section. We will now proceed to discuss our solution.

3.6.2 Privacy-Preserving Personal Federated Hypernetworks

3.6.2.1 Background

One approach toward privacy preservation by decentralizing data servers within *supervised* learning is **federated learning** (Shokri and Shmatikov, 2015). Federated learning algorithms train a global model from gradient updates sent by individual clients training on their own data, which is never sent to the central server. An extension of federated learning technique is Personal Federated Hypernetwork (PFH) (Shamsian et al., 2021), which allows for behavior tailored to individual heterogeneous tasks by splitting the model into a global common component (i.e. the **hypernetwork**), and a local individual component (i.e. a **local network** generated by the hypernetwork), which is tailored to each client. This task specialization allows for the learning of common features together in a global component network while allowing for learning client-specific knowledge in the local agents.

The combination of the two fields, federated multi-agent reinforcement learning, has focused mainly on learning global models, not personalized models for heterogeneous tasks Qi et al. (2021); Wang et al. (2020b); Ren et al. (2019); Anwar and Raychowdhury (2021); Kwon et al. (2020); Zhang et al. (2021b); Xu et al. (2021). Decentralized multi-agent reinforcement learning does learn personalized models Zhang et al. (2021a, 2018), but it may be difficult to scale up a decentralized system such that each agent can benefit from the experiences of all the others without large communication costs. The superlinear combinatoric communication increase is not as much of an issue for federated learning, as communication only needs to occur between clients and a server rather than clients and all their peers. Although decentralized systems have their benefits, we focus mainly on federated systems in this work.

3.6.2.2 Caveat on "Privacy Preservation"

We note that "privacy-preservation" might be to an extent an overstatement, as works have shown that the transmission of gradients can allow one to recreate private data (Xie et al., 2019; Hitaj et al., 2017; Melis et al., 2018). Thus, while we note that our work guarantees

privacy to the extent of other works within the field of federated learning (Wang et al., 2020c; Li et al., 2019; Acar et al., 2021; Zhang et al., 2020), one should apply the term privacy-preservation with the same caveats to our work as to the rest of the field.

3.6.2.3 Disambiguation between Multi-task and Multi-Agent

We wish briefly to disambiguate between **multi-task** and **multi-agent** for the reader's convenience. We use them in the conventional sense: multi-task relates to multiple, related settings (in our case slightly different MDP's in each different microgrid) whereas multi-agent refers to multiple different policies.

3.6.2.4 Related Work

We position our literature within an ecosystem of work related to transactive pricing in microgrids. A price-setting RL agent was first shown to help an energy aggregator improve demand response and generate a profit (Agwan et al., 2021a). Since then, a number of works have explored the issue (Shojaeighadikolaei et al., 2021; Wen et al., 2020; Han et al., 2021; Rolnick et al., 2022), with some work exploring different configurations of RL.

We wish to provide an example of federated learning. "Distributed Selective Stochastic Gradient Descent", i.e. DSSGD (Shokri and Shmatikov, 2015), is an interesting example which deserves further exploration from the interested reader. In DSSGD, each local model exchanges select parameters and gradient updates with the central server. In contrast, FedAvg (McMahan et al., 2017) averages all local model gradient updates and syncs all local model parameters. There exists much application of federated learning in RL, but very little attempts to personalize RL and no attempts to do so using PFH Nadiger et al. (2019); Wang et al. (2020a); Qi et al. (2021).

Existing multi-agent environments are often solved through multi-agent RL algorithms like MADDPG (Lowe et al., 2017), VDN (Sunehag et al., 2017), and Q-Mix (Rashid et al., 2018), but these aggregate data from all the agents onto one central machine during training, and take advantage of joint action-values from all agents. Other works use federated hypernetworks for multi-task setups, but specifically not those in RL.

3.6.2.5 Our Contribution

We present a novel application of PFH to RL in a realistic power systems setting that requires both privacy and heterogeneity in agents to accommodate diverse, sensitive environments.

Methodologically, our paper is novel in its presentation of an adaptation of a state of the art privacy-preserving algorithm to RL. To our knowledge, we are the first to explicitly apply personalized federated learning to multi-task, multi-agent RL when centralized learning and joint action-values are unavailable. Application-wise, our paper is also novel in its improvement in energy demand response across heterogeneous microgrids. We hope our work highlights an important microgrid environment to the RL community, helps establish the use of PFH within RL, and allows for RL to address problems where learning speed and privacy are fundamental.

3.6.2.6 Environment Setup

To increase the amount of data available, we consider multiple RL agents, each managing their own (slightly different) microgrid through energy prices and collecting data in parallel. We modify the MicrogridLearn environment into the Multi-MicrogridLearn environment, which we define as a multi-task, multi-agent setup in which the management of each microgrid, through prices, constitutes the task of a single agent. We characterize our problem as multi-agent because we have multiple RL agents optimizing a shared reward (total profit), and multi-task because optimization of profit in each of the different microgrids presents tasks that are related but also independent due to differences in size, number of batteries in each building, etc. We hypothesize that we can accelerate training by incorporating data from multiple microgrids with different characteristics. Learning to set prices using data from multiple microgrids (i.e. the source tasks) also opens the door to few-shot learning in new microgrids (i.e. the target tasks), wherein we learn to generate near-optimal prices for a microgrid very quickly. It is our hope to contribute to privacy protection by aggregating learning, *not* data, to one central source.

3.6.2.7 RL setup

Although our technique is RL architecture agnostic, we will briefly explain the RL agent that we use. We use PPO (Schulman et al., 2017) to train all of our RL agents to solve the MDP introduced in 2.4.1 because PPO is reliable and highly performant. Note that both algorithms introduced in 3.6.2.8 and 3.6.2.8 are agnostic to the architecture of the local policies, so one could use any gradient-based model.

3.6.2.8 Technical Setup of Personal Federated Hypernetworks

Federated Learning In order to learn a shared model between multiple microgrids in a privacy preserving manner, we turn to federated learning. McMahan et al. (2017) presented what is now the most popular federated learning scheme: Federated Averaging (FedAvg). FedAvg is simple to implement. Let the parameters for the policy for microgrid i at timestep t be $\theta_{i,t}$. All $\theta_{i,0}$ are initialized with the same weights, so $\theta_{1,0} = \theta_{2,0} = ... = \theta_{n,0}$, etc. Then, each policy trains on its own microgrid for k local steps, producing a new $\theta'_{i,k}$ for each microgrid, which has adapted to be better at price-setting in microgrid i than the original $\theta_{i,0}$. All $\theta'_{i,k}$ are transmitted back to a central server, where they compute the shared model for the next iteration by averaging all the $\theta'_{i,k}$.

$$\theta_{1,k} = \theta_{2,k} = ... = \theta_{n,k} = \frac{1}{n}\sum_{i=1}^{n}\theta'_{i,k} \tag{3.12}$$

Then the local models train on their own, send trained models back to a central server, and repeat. Sending model information only preserves privacy because only the parameters $\theta_{i,t}$' and not any data are communicated with the central server. Note that in our setup, every client participates in the weight exchange process, not just a sampled subset of the clients[XXX]. While FedAvg is a simple algorithm that performs well in supervised learning, it learns a global policy for all the price-setting agents. In our case, a global model is not ideal as microgrids may have different energy consumption/supply behaviors.

Hypernetworks for Personalized Federated Learning (PFH) To learn a shared model that is still able to personalize to individual microgrids, we turn to hypernetworks for personalized federated learning. (Shamsian et al., 2021). Personalized federated learning algorithm has found great success in supervised learning, beating FedAvg and personalized federated learning approaches based on meta-learning (Fallah et al., 2020), Moreau Envelopes (T Dinh et al., 2020), and Personalization Layers (Arivazhagan et al., 2019). However, personalized federated learning has never been used before for RL.

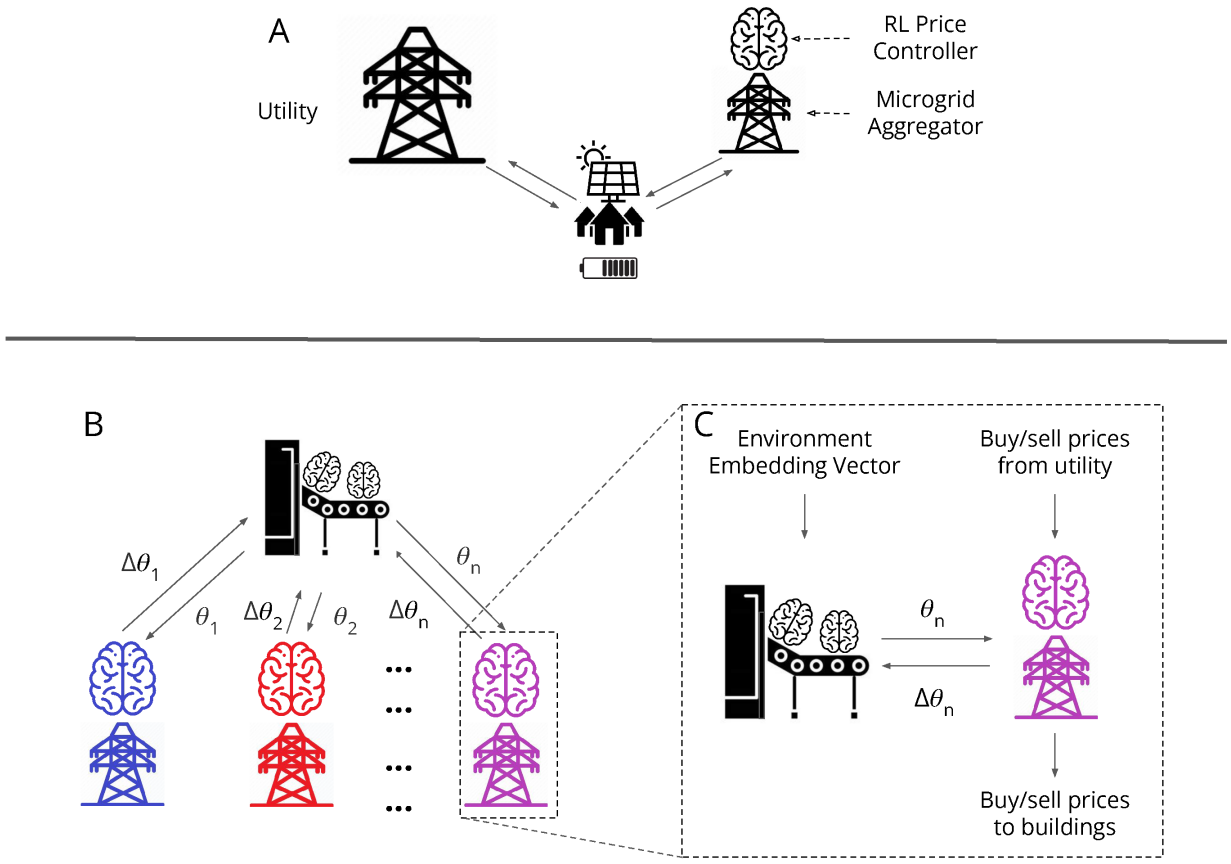


Figure 3.19: **Microgrids and PFH: A.** We imagine a prosumer that can, at each hour of the day, choose to sell energy surplus or purchase unmet energy demand from the larger utility or to the microgrid aggregator. The microgrid aggregator's energy buy/sell prices are determined by an RL controller. **B.** A Hypernetwork for Personalized Federated Learning (PFH) receives gradient updates from RL controllers and sends back weights. **C.** The hypernetwork takes as input an environment embedding vector and outputs weights for an RL controller. The RL agent takes as input buy/sell prices from the utility and outputs buy/sell prices to the buildings in the microgrid the agent manages. The RL agent sends back a gradient update to the hypernetwork, which uses the update to compute the gradient update for the hypernetwork's own weights (Jang et al., 2022b).

Now we will describe PFH more formally. Please refer to Fig. 3.19 for a visual of the algorithm, and Algorithm 3 for pseudocode. Consider again $\theta_{i,t} \in \mathbb{R}^m$ as an m dimensional

Algorithm 3 Personalized Federated Hypernetworks in RL

Input: Task set $\mathcal{T}$ and hypernetwork ξ_ϕ. For each task $\eta \in \mathcal{T}$, an RL policy $\pi_\theta(\eta)$ and hypernetwork-specified parameters $\xi_\phi(\eta)$.
Hyperparameters: Number of training rounds N, number of local training steps per hypernetwork update K, learning rate α.
for $i = 1, \ldots, N$ **do**
  for task $\eta \in \mathcal{T}$ **do**
    Get parameters $\tilde{\theta}(\eta) := \theta(\eta) := \xi_\phi(\eta)$.
    for $k = 1, \ldots, K$ **do**
      Collect rollouts $\tau_{\eta,i}$ from η using policy $\pi_{\theta_{\eta,i}}(\eta)$ with parameters $\hat{\theta}(\eta)$.
      Update $\hat{\theta}(\eta)$ using PPO with rollouts $\tau_{\eta,i}$.
    end for
  end for
  Initialize $\bar{\phi}_{\text{update}} := 0$.
  for task $\eta \in \mathcal{T}$ **do**
    $\Delta\theta(\eta) := \hat{\theta}(\eta) - \theta(\eta)$
    $\bar{\phi}_{\text{update}} := \bar{\phi}_{\text{update}} + \frac{\nabla_\phi \theta(\eta)^T \Delta\theta(eta)}{|\mathcal{T}|}$
  end for
  Update hypernetwork parameters $\phi = \phi - \alpha \cdot \bar{\phi}_{\text{update}}$.
end for

vector denoting the parameters of the policy for microgrid i at timestep t. A *hypernetwork* is a neural network that outputs the parameters of another neural network. We will have one global $\xi_{\phi_t} \in \mathbb{R}^l \to \mathbb{R}^m$ parameterized by ϕ_t. ξ_{ϕ_t} takes as input an environment embedding vector $v_i \in \mathbb{R}^l$, which is learned for each environment along with the hypernetwork. We initialize $\theta_{i0} = \xi_{\phi_0}(v_i)\ \forall i \in [1, ..., n]$. Then each[28] local agent trains for k steps, producing new parameters $\theta'_{i,k}$. Then, $\Delta\theta_{i,k} = \theta'_{i,k} - \theta_{i,0}$, is sent back to the central server, where it is used to update the hypernetwork:

$$\phi_k = \phi_0 - \frac{1}{n} \sum_{i=1}^{n} \alpha \nabla_{\phi_0} \theta_{i,0}^T \Delta\theta_{i,k} \tag{3.13}$$

Since the hypernetwork outputs NNs conditioned on the environment, it is able to create RL agents that are personalized to the needs of each microgrid. We also still preserve privacy by only communicating parameters with the central server instead of data.

Diversity and Optimal Use of PFH One factor that could affect the relative performance of PFH is the heterogeneity of the scenario. A homogeneous scenario (imagine a cookie-cutter residential neighborhood) could be suitable for federated learning methods due to similarity in behavior. In contrast, an extremely heterogeneous scenario (imagine mixed-use city blocks with night-life, shopping, and residential real estate) could have wildly different energy

[28]Note our setup is slightly different from the original PFH (Shamsian et al., 2021); every client participates in each round, not just a sampled subset of clients. We made this small variation to better understand whether scaling the algorithm to larger numbers of microgrids would be useful.

demands, which may be better learned by individual local networks without any mechanism to share learning. We hypothesize PFH will perform competitively in some average of these two extremes. If local environments are diverse yet share similar underlying mechanisms, PFH will be able to fit to local conditions while sharing information on common trends.

3.6.2.9 Experimental Setup

Simulating Diverse Microgrids Because each microgrid is defined by a distribution of photovoltaics and battery sizes, we propose a simple way to tweak the amount of diversity in a system. We sample photovoltaic and battery sizes from normal distributions, changing the variance σ^2 as the diversity parameter, and round outcomes to the nearest integer[29] We sample from $\mathcal{N}(\mu = 100, \sigma = 10)$ for low diversity cases, $\mathcal{N}(100, 30)$ for medium diversity and $\mathcal{N}(100, 50)$ for high based on the spread of two standard deviations[30]. We note here that we have chosen the low, medium, and high cases such that 95% of samples (i.e. 2 standard deviations around the mean) in the high case hit realistic bounds in the environment; i.e., 0 (an obvious lower bound) and 200[31].

Baselines We compare PFH against FedAvg and two other baselines. First, we observe what happens with no RL control at all; the microgrid aggregator outputs prices that are exactly the same as the utility's. We assume buildings choose to meet half their energy demand/surplus with the utility and half with the aggregator. Our second baseline is the approach used in Agwan et al. (2021a): training all the local RL controllers with only their own data: no central model or inter-microgrid communication. These two baselines, no RL and local control, are designed to highlight the added value of RL[XXXI] to the task of price-setting for energy demand response in microgrids, and the added value of having some central model that aggregates learning across multiple microgrids, respectively.

For specification on how we selected hyperparameters, please see Section 3.7.1.

Multi-Task Transfer An interesting feature of our hypernetwork-based setup is the potential for multi-task learning and few-shot transfer learning. The optimization problem of setting prices for each microgrid can be viewed as an individual task, $\eta \in \mathcal{T}$. Since the hypernetwork should learn some common strategies for each task, we tested whether it can generalize to unseen tasks with little training. To test this hypothesis, we simply take a hypernetwork that has trained for to manage a microgrid cluster with 20 microgrids of medium diversity and train the hypernetwork to manage a *new* microgrid cluster of 20 microgrids with the same level of diversity. By pretraining our hypernetwork on 20 varied source tasks, we hope to encode enough knowledge applicable to the new target tasks to make few-shot transfer learning possible. We will refer to such a pretrained hypernetwork as a Few-Shot PFH.

[29]As we are sampling from a "hyper" distribution to instantiate houses, the means of the distribution are not as important as the variances in instantiating diversity.

[30]All distributions are truncated at 0.

[31]200 is a realistic upper bound in both solar panels and batteries: 200 solar panels would require an area of 60 x 70 ft, which bounds the square footage of many commercial roofs, and 200 batteries would be a realistic upper bound of entities not engaging in commercial grid services.

Table 3.1: Cumulative profits above base utility pricing after 10,000 days, in hundred thousands.

Scenario	PFH	FedAvg	Local Baseline
Simple, 5 agents	39.23	**45.75**	43.72
Simple, 10 agents	42.11	41.65	**43.18**
Simple, 20 agents	40.85	34.52	**42.82**
Medium, 5 agents	**47.50**	40.95	40.12
Medium, 10 agents	**46.89**	43.82	41.73
Medium, 20 agents	**48.22**	39.38	39.66
Complex, 5 agents	34.77	32.66	**35.60**
Complex, 10 agents	43.01	42.08	**45.24**
Complex, 20 agents	**44.39**	38.70	40.78

3.6.2.10 Results and Discussion

PFH Accelerates Learning in Medium Diversity microgrid clusters Fig. 3.20 shows average daily profit gained by each microgrid in a microgrid cluster with 5, 10, and 20 microgrids, with varying amounts of diversity. The middle column of Fig. 3.20 shows PFH is more efficient and profitable for a microgrid cluster than a microgrid cluster under a FedAvg or local control scheme. As shown in Table 3.1, PFH results in up to $8,500,000 of additional cumulative profit after 10,000 days over the local control baseline in a microgrid cluster with 20 microgrids. However, this advantage does not carry over to cases of small or large diversity. For less diverse scenarios, PFH was comparable or less profitable than FedAvg or local control. For more diverse scenarios, local control was generally more profitable. The number of microgrids in microgrid clusters also did not seem to have much effect on learning speed here.

FedAvg Recovers Local Performance at Best Curiously, our results indicated that FedAvg presented did not improve the management of a microgrid cluster over a collection of local agents. We had expected FedAvg to perform better in the homogeneous case, and to scale with the number of agents, but neither effect appears in our results. Although FedAvg may perform well in supervised learning (McMahan et al., 2017), it may not extend well to RL. We explain FedAvg's poor performance as follows: unlike supervised learning, RL requires exploration. One could imagine each microgrid's local agent explores in different direction to another, causing gradient updates that are not well conditioned when averaged together. Meanwhile, the hypernetwork is able to learn how to build RL policies with different exploration behaviors from the aggregated gradient because it outputs agents personalized to each task.

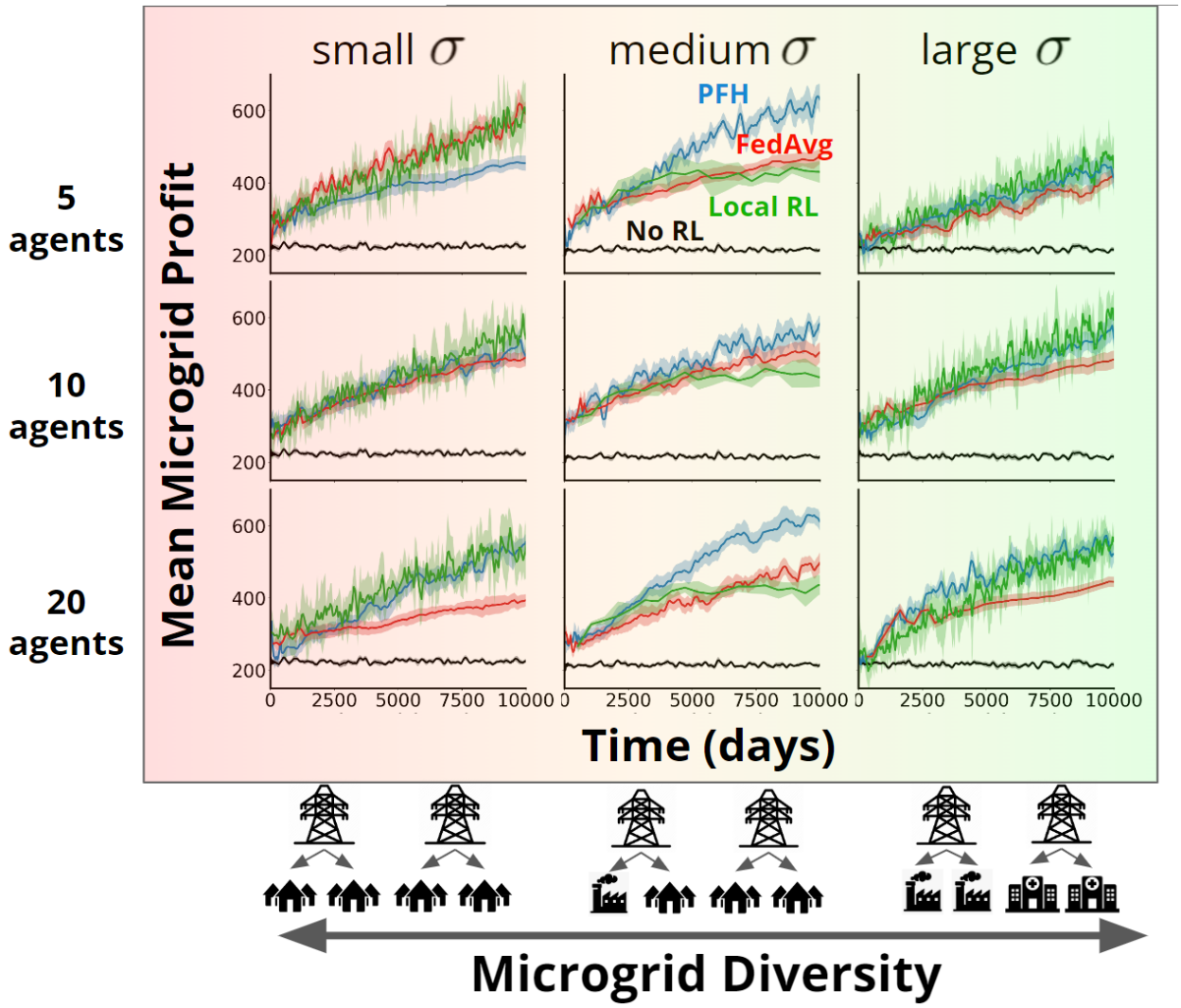


Figure 3.20: **RL Agent Performance:** The performance of the RL price-setting agent as a function of the number and diversity of the microgrids in the microgrid cluster . Performance is measured by looking at the average daily profit gained by each microgrid (Jang et al., 2022b).

PFH Enables Few-Shot Learning Fig.3.21.A shows the hypernetwork adapted to a new set of microgrid management tasks extremely quickly. On average, within ≈ 1.5 months (42 days), each new microgrid achieved ≈ $380 in daily profit, which is about the daily profit of the local agents baseline after 13 years (5000 days) of training. The original, randomly initialized PFH required 3000 days to achieve similar performance. Thus, Few-Shot PFH achieved a 119x speedup over local agents and 71.4x over a randomly initialized PFH over the first 1.5 months. Within 7 months (210 days), Few-Shot PFH achieved a daily profit of $565: 44% higher profit than the local agents ever achieve. A randomly initialized PFH required ≈ 22 years (8000 days) to achieve similar performance: a 38x speedup in the first 7 months of training. Cumulatively, having a pretrained PFH on 20 microgrids saves ≈ $1,500,000

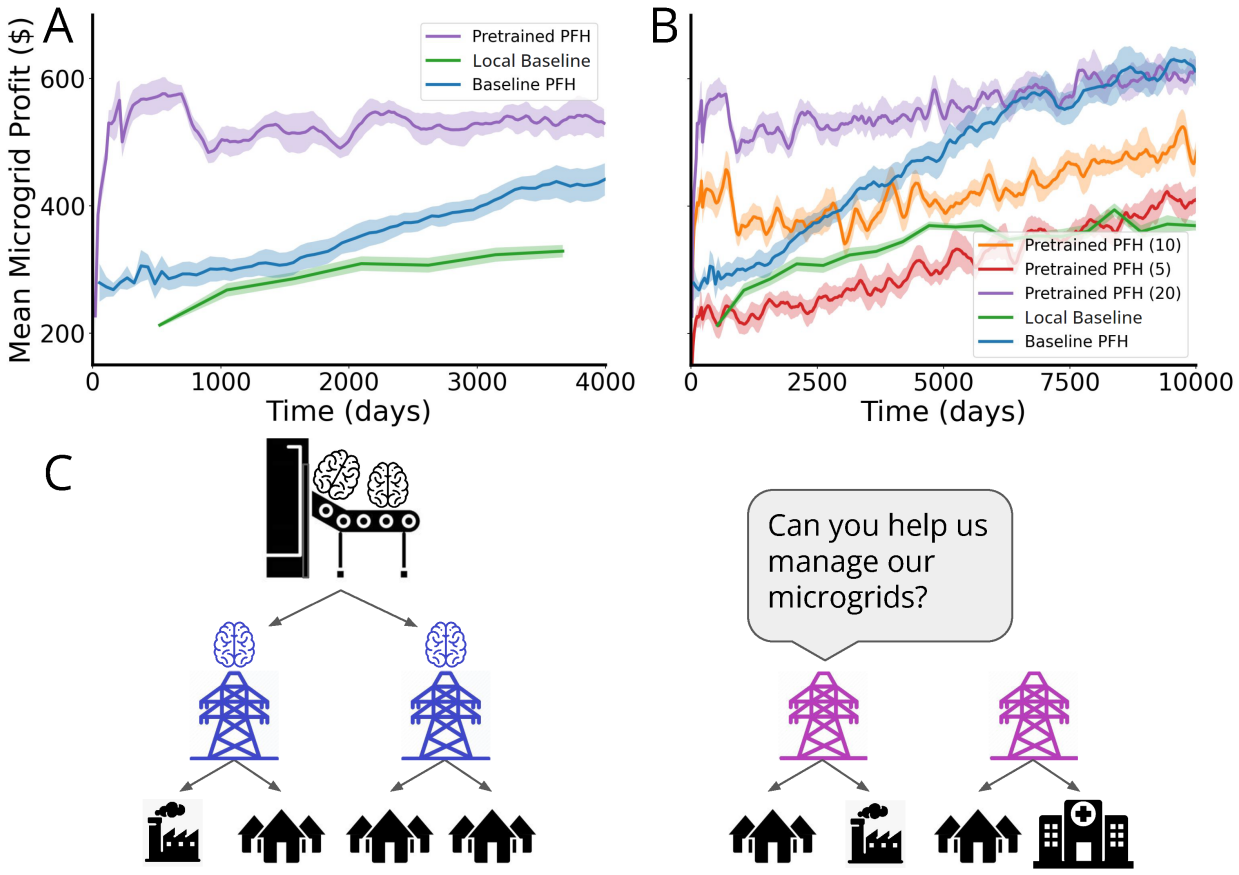


Figure 3.21: **PFH Enables Few Shot Learning: A.** Mean microgrid profit of PFH pretrained on 20 microgrids learning to manage 20 new microgrids ("Pretrained PFH"), compared to randomly initialized PFH ("Baseline PFH") and the local agents baseline ("Local Baseline"), over training days on the new microgrids. **B.** Mean microgrid profit of PFH pretrained on 5, 10, and 20 microgrids on a new set of microgrids, over a longer time than A. **C.** A plausible scenario in which PFH may need to quickly adapt to new microgrids (Jang et al., 2022b).

over the course of training on the new microgrid management tasks compared to a randomly initialized PFH.

Few-Shot Learning Capability Scales with microgrid cluster Size When we tried the same experiment with hypernetworks that were trained for 10,000 days on 5 microgrid management tasks and 10 tasks in Fig. 3.21.B and tested on 5 and 10, respectively, we saw significantly smaller boosts in the mean reward over groups of new tasks with fewer training tasks. The smaller scale of benefit was expected given a multi-task learning strategy with fewer source tasks and data. Indeed, when trained on 5 tasks, there was hardly any initial training speedup. Starting from 10 tasks, we observed a large initial boost (although not as large as with 20.) Rather strikingly, Few-Shot PFH pretrained on 5 and 10 tasks converged to lower reward curves than even the baseline PFH (i.e., a randomly initialized hypernetwork.)

With 20 tasks, we saw both a large initial boost in training speed and no adverse impact on long term training. We hypothesize the fewer microgrid source tasks provided, the more information is stored in the environment embedding, which makes the hypernetwork brittle to new environments. Thus in the 5 and 10 case, the net has not learned enough shared dynamics in the other parameters to generalize to new settings. In the case of 20 and above, we expect that enough shared dynamics are learned that the net can generalize. The range of training speed benefits we observed suggested the potential in some configurations for a Few-Shot PFH to quickly adapt to new tasks depends on how many tasks it was initially trained on.

3.6.3 Societal Impact

What are potential negative societal effects of our work? Overall, negative effects to prosumers are limited, as the focus of our work is in protecting consumer information. Furthermore, prior work demonstrated that the presence of an aggregator consistently reduced energy costs for consumers.

However, a persistent danger of Artificial Intelligence (AI) is that it is often deployed through centralized profit-seekers. Our work is no different in this regard. Although our specific innovation protects prosumers, it may improve the economic viability of a profit-seeking entity whose scale may eventually enable it to further its own profit at the expense of prosumers.

Also, the act of setting prices in systems may raise fairness concerns. If initial training microgrids are biased towards wealthier residents, the PFH may initialize new policies with pricing that benefits consumption habits of wealthier clients but not poorer clients. A vivid illustration may be seen in the types of prosumers who are best poised to benefit from economic aggregation: prosumers with large solar panels and batteries are able to shield themselves from or profit off of high prices by consuming their own energy, and may fully charge their batteries when prices are low. Prosumers with smaller or no storage capabilities do not have this luxury, and thus are more vulnerable to the negative effects of price fluctuation.

3.7 Hyperparameter Sweep Difficulty

Training RL agents can be fairly tricky even when we happen to choose the right *hyperparameters*, i.e. the parameters that define structural aspects of the agent such as the number of layers in the policy, the learning rate, etc. Choosing these hyperparameters, or defining a heuristic to choose them, may be even more challenging. It is essential that these hyperparameters are chosen in simulation, as any attempt to sweep on hyperparameters in the real world would be extremely data costly. Thus, we will present a technique from classic statistics to help alleviate hyperparameter search complexity.

3.7.1 Regression Analysis of Hyperparameter Optimization

We suggest the use of inferential statistics to gain an intuition of hyperparameter fitness. Simple linear regression can be a powerful tool for this application. We regress the average reward during training on the hyperparameter combination:

$$\overline{y} \sim w_0 + w_1\alpha_1 + ... + w_n\alpha_n \tag{3.14}$$

Where each weight w_i is a linear coefficient on a corresponding hyperparameter α_i. Using standard, classic methods of linear regression fitting, one can eliminate hyperparameters from consideration that have low impact on the reward, thereby saving considerable compute by dropping these parameters.

We will demonstrate fitness in the data from the experiments detailed in the above section, PFH.

3.7.2 Regression Explorations of Hyperparameter Sweeps

Here we present, for the reader's interest, a regression fit on the hyperparameters that were swept over. In this regression, each observation is a single run of the sweep, the dependent variable in both is the reward mean of the learning trajectory, and the independent variables are those listed in the rows.

We believe that this regression contains some interesting information; specifically on the direction of coefficients (i.e. whether they are negative or positive) and on which parameters were significant in producing a positive reward. We note that the basic assumption of linear regression, that observations are sampled IID from a distribution, is not the case here; observations are loosely dependent on each other as the parameter configurations in each batch are determined by the performance of parameters in the previous batch. Thus, we caution that the positive results may be mused over as a curiosity only. We are more confident in the *negative* results of this regression, i.e. which variables are insignificant after controlling for the others, than the positive results, as this indicates parameters that the sweep chose not to focus on. We believe that further work in regressions of hyperparameter values may be an interesting research endeavor for understanding ML models as well as for ML applications like AutoML.

3.7.2.1 Full Regression Model (Table 3.2)

Of specific interest in the full regression are which hyperparameters did and did not effect the average reward. Many variables in the hypernetwork itself do not seem to matter: the hypernetwork's learning rate, number of layers, and L2 regularization did not matter. However, whether or not the hypernetwork selected for dropout *did* matter, and it hurt the performance, implying that hypernetwork fitting was more important than robustness. Some parameters of the PPO agents, such as the clip parameter or number of gradient updates, did not matter much either.

Table 3.2: Full regression model

Dep. Variable:	Avg Reward	**R-squared:**	0.362
Model:	OLS	**Adj. R-squared:**	0.348
Method:	Least Squares	**F-statistic:**	24.62
Date:	Thu, 19 May 2022	**Prob (F-statistic):**	1.08e-43
Time:	08:16:34	**Log-Likelihood:**	-3188.0
No. Observations:	533	**AIC:**	6402.
Df Residuals:	520	**BIC:**	6458.
Df Model:	12		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
Intercept	344.6188	29.679	11.612	0.000	286.314	402.924
sizes	-1.9453	0.777	-2.504	**0.013**	-3.472	-0.419
n_layers	5.2696	2.806	1.878	**0.061**	-0.243	10.782
learning_rate	-1835.5718	128.380	-14.298	**0.000**	-2087.779	-1583.364
ppo_clip_param	49.4402	38.009	1.301	0.194	-25.230	124.110
hnet_lr	77.4179	63.747	1.214	0.225	-47.815	202.651
hnet_num_local_steps	1.0579	0.150	7.043	**0.000**	0.763	1.353
ppo_num_sgd_iter	-0.3889	0.741	-0.525	0.600	-1.845	1.068
hnet_num_layers	-2.3439	2.571	-0.912	0.362	-7.395	2.708
batch_size	-0.6685	0.171	-3.910	**0.000**	-1.004	-0.333
hnet_embedding_dim	0.0077	0.028	0.271	0.786	-0.048	0.063
hnet_l2_reg	-10.4826	19.912	-0.526	0.599	-49.601	28.636
hnet_dropout	-68.5849	21.070	-3.255	**0.001**	-109.978	-27.192

Table 3.3: Regression parameters from hyperparameter search (Jang et al., 2022b).

3.7.2.2 Reduced Regression Model with Only Significant Coefficients Included (Table 3.5)

We believe that this test is more interesting for examining the direction of significant variables after controlling for the other variables. Here, it is interesting to note greater sizes of policy

Omnibus:	16.401	**Durbin-Watson:**	1.670
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	19.555
Skew:	0.336	**Prob(JB):**	5.67e-05
Kurtosis:	3.655	**Cond. No.**	9.08e+03

Table 3.4: Regression parameter metrics of fit (Jang et al., 2022b).

networks, ("sizes"), has a negative effect, while number of layers of policy networks have a positive effect, offering a mixed view on whether policy complexity is important. The learning rate is negatively important, implying that more stability in network is preferred. The number of local model update steps allowed is positively correlated to mean reward, implying that the more the local models are allowed to fit, the better. The combination of a negative effect of batch size and positive effect of learning rate implies that the local RL agents found it easier to take few steps (lower batch sizes) but update policies at a more conservative rate (lower learning rates), which makes sense in the RL context.

Table 3.5: Reduced regression model

Dep. Variable:	Avg Reward	**R-squared:**	0.357
Model:	OLS	**Adj. R-squared:**	0.349
Method:	Least Squares	**F-statistic:**	48.63
Prob (F-statistic):	1.79e-47	**Log-Likelihood:**	-3190.3
No. Observations:	533	**AIC:**	6395.
Df Residuals:	526	**BIC:**	6425.
Df Model:	6		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
Intercept	364.7212	15.515	23.507	0.000	334.242	395.201
sizes	-2.1970	0.627	-3.502	0.001	-3.429	-0.965
n_layers	5.1216	2.759	1.856	0.064	-0.299	10.542
learning_rate	-1829.2013	127.609	-14.334	0.000	-2079.887	-1578.516
hnet_num_local_steps	1.0482	0.149	7.055	0.000	0.756	1.340
batch_size	-0.6409	0.167	-3.841	0.000	-0.969	-0.313
hnet_dropout	-65.1701	20.923	-3.115	0.002	-106.273	-24.067

Omnibus:	17.230	**Durbin-Watson:**	1.669
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	19.988
Skew:	0.360	**Prob(JB):**	4.57e-05
Kurtosis:	3.618	**Cond. No.**	2.57e+03

Table 3.6: Reduced regression coefficients pruned for significance (Jang et al., 2022b).

3.7.3 Conclusion to Hyperparameter Regression

We are not confident that this method will outlast our focus on it. However, it is our hope to present it to remind the reader that (a) some cursory attention to hyperparameters in a fit is important (b) classical statistical techniques may still have value entering this new age of automated learning.

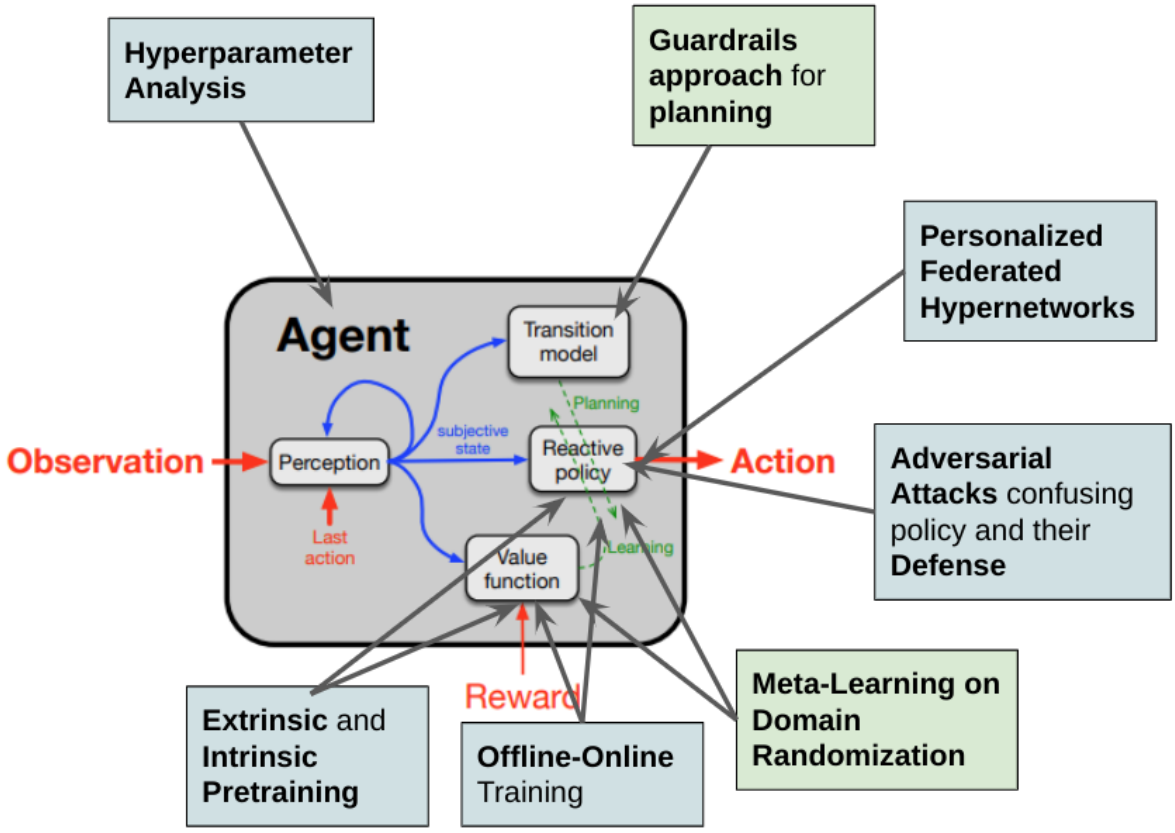


Figure 3.22: Overview of the methods presented as solutions to sim-to-real problems. Base image from (Sutton, 2022).

3.8 Comparison of Sim-to-Real RL Methods

Here we endeavor to compare across the methods we have presented here. Please see Figure 3.22.

We again reproduce the cell of the generally intelligent agent[32] presented by Sutton. Here, we map each solution approximately onto the area that it affects. In this case, we put arrows in to point towards both policy and value function when a method affects training, as both are included in Actor-Critic methods.

We color in green the methods that we believe are most promising to help RL in the real world. The reasons we select these two are the following:

Guardrails Method: We expect that any RL agent operating in a space in which real data is costly to collect will have a simulation or planning model associated with its development[33].

[32]We unfortunately pay little attention to the perception module in our work, leaving it only as identity.

[33]We have seen that planning models are common in development in our experience. developing real life RL agents: sometimes it is for as trivial a reason that a team of software engineers completes the RL agent but is waiting on a team of hardware engineers to hookup the system, so they make a realistic simulation of the system in its stead.

However, many methods that we surveyed were only as good as the accuracy of the planning model. The guardrails method we proposed, instead, was shockingly effective in its effect on learning, precisely because it *takes advantage* of the uncertainty in the model output in accommodating distributions. Thus, we recommend it for general use.

Meta-Learning in Domain Randomized Environments One way to anticipate some of the variation in the real world is to train in version of the world that have variable dynamics. We believe that the general approach that meta-learning provides paves the way for more interesting approaches in auto-curricula and guided environment design. We believe these directions hold tremendous promise in RL. We hope to contribute to these directions in the future.

Personal Federated Hypernetworks It is the opinion of the author alone, and probably not his colleagues, that PFH deserves an honorable mention. If PFH were to work well, the benefits are tangible and obvious. Of *course* we would want new subdomains to be instantiated with knowledge from previous. However, it is unclear the scale required from previous subdomains, and the efficacy of the controller before the point of PFH effectiveness is important to guarantee.

Chapter 4

Experiments

Finally, we wish to address the "real" part of our Sim-to-Real RL goals. We will first present the setup of a proposed experiment demonstrating the effectiveness of OfficeLearn, with the experimental setup and supporting software described. We will then present two real experiments ran in a similar setup prior to our investigation of RL in demand response.

Our goals in Chapter 4 are as follows.

First, although the experiment was postponed indefinitely due to mandated office shutdowns in Singapore's COVID-19 response, we laid groundwork for implementing RL in experiments. Hopefully a reading of our experimental setup can help germinate ideas on how to implement the reader's own RL-related experiments.

Second, we wish to present hard evidence from prior experiments that demonstrates that aspects of the RL-related setup are well tuned. The incentive allocation, for instance, turned out to be tuned finely enough to elicit differences in behavioral response. A strong decrease in energy consumption was noted when an experimental group was shown trees depicting their energy use, whereas no effect was noted when an experiment group was only shown barcharts. Had the incentive allocation been an incredibly strong motivator, we believe we would have seen significant responses from both over a control group that were not part of the experiment. Thus, we present these experiments to justify why choices were made in what would have been the later experiment administering RL.

4.1 Proposed Experiment: Price-Setting Reinforcement Learning Optimizing Office Behavioral Energy Demand Response

In 2019, we proposed to experimentally validate our OfficeLearn through an experiment. Here we will describe the implementation of the experiment, starting with contextualization in the literature, the experimental design, and social game implementation.

4.1.1 Background

Notable work has been performed in creating and administering "Social Games" – for our purposes, defined as competitions around energy use. Social Games tend to comprise of: (1) informing each player of their energy use in a friendly and easy to consume manner and (2) an accompanying competition in which higher energy savings relative to others is rewarded. Many studies have either created Social Games themselves ((Konstantakopoulos et al., 2017), (Ratliff et al., 2014), (Papaioannou et al., 2018), (Papaioannou and Stamoulis, 2018)) or studied existing Social Games ((Cowley et al., 2011), (Ayres et al., 2012)) to draw insight. To this end, the general finding has been that Social Games increase the extent to which an individual is motivated to and able to save energy in their daily functioning within their office or home.

4.1.2 Allocation of Incentives

We would like to give special importance to our treatment of incentives, I, as it has been tested in practice and will be an important part of running this experiment.

The expense I can be seen as a fixed amount prize that is awarded to high performing workers. It serves as a level of behavioral engineering to stimulate interest in the building manager's price signal. In other words, given a price signal, the presence of a prize-based incentive ensures that office workers modify their consumption to minimize the cost of energy in accordance to the price signal. An example of such a mechanism is a points-based energy competition developed and implemented in experiment in (Konstantakopoulos et al., 2019b) and (Konstantakopoulos et al., 2019a) with significant success. Each player is given a baseline amount of points daily depending on historical energy usage without the price signal, and the price of their energy that day in the game is subtracted from the baseline. At the end of each two week period, the building manager hosts a building-wide meeting in which the top third of players (by accumulated points) are entered into a lottery, and gifts such as Amazon gift cards are allocated randomly per the Vickrey-Clark-Groves (VCG) auction mechanism. The total value of the prizes (I) given out at the end of every two week round is fixed at around \$400. I is independent of the total number of people playing the game, which makes the incentive mechanism scale nicely for larger offices.

4.1.3 Proposed Experimental Timeline

We observe two experimental units for a period of five months, from August to December. We are interested in estimating the causal effects of two distinct reinforcement learning architectures, RL_j, for $j \in \{1, 2\}$, in addition to the causal effect of combining reinforcement learning with behavioral feedback from the planning model.

We estimate seasonality effects at period t (δ_t) and improvements in learning due to the accumulation of observations (Ω_t), by taking the average difference between performance across all conditions at time t and t_{-1}. We also estimate the effect of incorporating parameters from the social cognitive model by comparing observations *within* a single RL architecture, controlling for seasonality. Finally, we causally identify the effect of each RL_j, which is the difference between the score for RL_j versus $RL_{\text{-}j}$, and the difference in scores between RL_j and $RL_{\text{-}j}$, *conditional on incorporating the social cognitive model*, controlling for seasonality in each case.

Month	Group 1	Group 2	Control
July	— System ID —		
August	RL_2	RL_1	
September	RL_1	RL_2	

Table 4.1: Experimental Timeline in which we compare two different RL architectures and the effect of a planning model Spangher et al. (2020a).

We then use later experimental periods to train the model on smaller subsets of our experimental subjects: smaller groups and then at the individual scale.

Although we have leave RL_1 and RL_2 unspecified in this chapter, the reader may refer back to Section 3.8 to see what we recommend to be most promising RL architectures that may be tested in this forum.

4.2 Bridges between Proposed Experiment and the Real World: Software Methods

4.2.1 RL API

Implementing RL in the real world is a challenge that requires attention to scaffolding software well beyond the RL agent itself. Here we describe the development of an Application Programming interface to allow the Social Game agent to communicate with the administrator of the building Social Game.

4.2.1.1 What is an API?

An API is a software interface that connects two pieces of software to each other in intentionally simplified ways. The connection must be done such that the first piece of software can access the functionality of the second without needing to know what happens under the hood,

and vice-versa. The main type of API, a RESTful API, is a framework for caching and guaranteeing on-demand interoperability. In this case, we propose an API to connect the software used to manage a social game in Singapore with an RL controller that determines how many points to assign each participant in the social game.

Popular libraries for RL include RLLib and StableBaselines. StableBaselines does not present a formal API. While RLLib has an API, it is not RESTful, and limited to functionality that governs agent operation within a contained simulation environment. We require greater functionality than it can provide: it would have difficulty interacting with databases, switching between different agents, or performing other operations that are necessary for our server side. Thus, a custom API is needed.

Our interface facilitates plug-and-play of any kind of underlying RL model for the points controller, while maintaining a single data store. This allows researchers to freely iterate on their social game API containers with improved models without any risk of interrupting their data collection pipeline. Our API was created using Flask (a Python backend software facilitating API creation), backed by a disk-local SQLite database (which is based on Structured Query Language (SQL)), and containerized in Docker. Docker is a secure packaging software, facilitating creation of isolated containers in which to deploy software environments. One would query our API using Postman, curl, or any other HTTP request creation software.

Our API exposes asynchronous endpoints to add participants to the game, get energy pricing for each participant for the upcoming day, submit participants' energy usages for the day, and more. It provides a centralized, standardized way for social game researchers to interface with our RL models, upload social game history and retrain models with the most up-to-date data, and receive suggestions for points to assign participants the next day. It performs all point calculations automatically upon data entry, and can output a running tally of scores and a leaderboard at the conclusion of the game. It also contains comprehensive logging for debugging and observability. As per RESTful guidelines, PUT requests are ones that modify information on the server side (i.e. our side) whereas GET requests provide information for the client side.

Endpoint Descriptions

- **Submit Game Users [PUT] [/participants]**

  Function: Add participants to the social game; i.e., add a coded player to the internal list of players that is maintained as part of the social game. Later endpoints will iterate over each of these participants to save their energy data.

- **Get Hourly Energy Pricing [GET] [/energy/pricing]**

  Function: Load price signal for the upcoming day for each participant in the game, calculated using loaded model parameters. This is either the agent's predicted actions, or baseline square waves hard-coded for initial training periods.

- **Submit User Hourly Energy Consumption [POST] [/energy/consumption]**

  Function: Add participant energy usage data from the previous day to the SQLite database. The agent will draw from this database to train through an internal function.

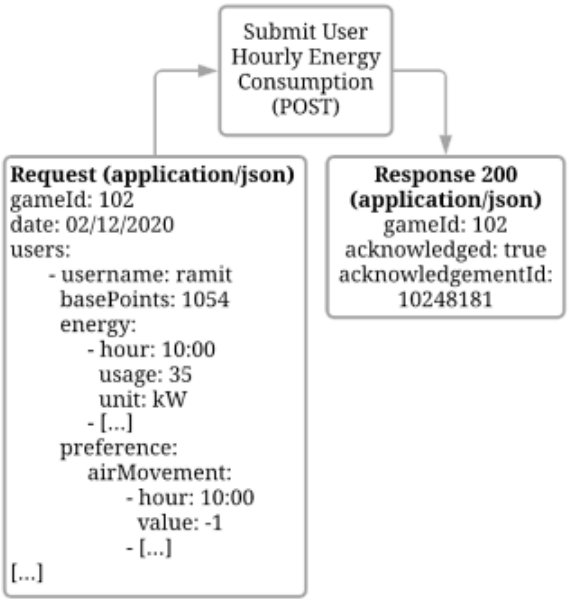


Figure 4.1: An illustration of a possible path of commands that the API endpoints expose. Un-published work by the author.

- **Get Points Earned and Base Points [GET] [/energy/points]**

  Function: Calculate how many points to award each participant based on each individual's price signal and energy usage. These values are calculated from (historical) base points and are not related directly to the RL agent.

- **Calculate Game winners [POST] [/game/winners]**

  Function: Maintain leaderboard throughout the social game. Position in the leaderboard is an aggregation of points throughout the game.

In sum, the API allows for the deployment of the RL agent and scaffolds surrounding aspects for its proper functioning. It allows the RL agent to interact with databases for more robust data storage, and creates a record of operations so that long term logging can occur. In this way, it is a novel and custom advancement from standard off-the-shelf RL APIs.

4.2.2 Experiment Implementation Indefinitely Postponed Due to COVID-19

The target office for our experiment was the Singapore Berkeley Building Energy in the Tropics (SinBerBEST) research office in Singapore. The office was composed of 40-50 graduate students, postdocs, faculty researchers, and administrative staff who used energy in structured ways. We began planning this experiment in Spring 2019.

Unfortunately, the experiment was relied on conventional office dynamics pre-COVID. Specifically, office workers needed to generally keep to a certain schedule and come into the office regularly. As the experiment was set to deploy in Spring 2020, it was first postponed as the COVID pandemic drastically changed office habits and then cancelled altogether as we could not predict when office workers would return to the office with some semblance of regularity. Indeed, the office has as of this writing not returned to normal routines.

Thus, we will now describe prior experiments we ran in the same office building as a proof of concept for how the price-setting experiment may run. Specifically, we wish to demonstrate that the prior experiment elicited enough signal from the study population that we saw an effect from the experimental treatment but not the baseline or control treatments.

4.3 Energy Reduction Experiment 1: Testing Visualizations Guiding Energy Awareness

Here we will describe an experiment that we ran to reduce energy consumption during work hours. As the experiment is too coarse a signal to mediate energy during certain parts of the day, we do not wish to imply that it is a demand response experiment. However, we do want to note that shortening the time window of hours in which energy is reduced may be a simpler way to implement demand response generally. It is our hope that in describing this experiment carefully, the reader may understand nuances in how to implement their own experiment in energy reduction.

4.3.1 Background on Energy Visualization

Data visualizations as they relate to energy have also been studied extensively ((Holmes, 2007), (Börner et al., 2012), (Murugesan et al., 2013)). The literature generally taxonomizes the space of visualizations into two main types that we consider here (Rodgers, 2011):

1. **"engineering-type"** visualizations composed of barcharts, line graphs, scatterplots – i.e. a formal, work-like presentation of the data in a way that would be used in scientific papers

2. **"ambient-type"** visualizations, in which a linear scale is communicated by some abstract, pleasant, artistic visuals[34].

Examples of an "ambient-type" visualization would include the so-called "power-aware" cord, a power cord that glows more brightly the more energy is being used (Gustafsson and Gyllenswärd, 2005), a "thrifty faucet" which shines a red or blue light depending on the temperature of the water (Togler et al., 2009), or an ambient battery communication system that communicates the battery's charge with a proportional intensity of light (Elbanhawy et al., 2016).

The consensus in the literature is centered around the idea that ambient visualizations are more effective than engineering type visualizations at consistently communicating to a user their energy usage ((Chetty et al., 2009; Piccolo et al., 2014; Kim et al., 2009; Quintal et al., 2010; Polson and Selin, 2012; Spangher, 2018)). The effect has been studied qualitatively. However, we are unaware of a study that attempts to quantify the behavioral differences that types of data visualizations engender. Therefore, we aim to reproduce a Social Game experiment and quantify the difference in energy consumption from participants exposed to different types of visualizations.

[34]There is a third type, the (3) **"natural type"** visualizations, in which images closely match the natural world and map energy onto environmental impact. However, it has been generally found that this type of visualization invokes guilt in the viewer and is thus less often used.

4.3.2 Experimental Methods

4.3.2.1 Pre-treatment Survey

We conducted a pre-treatment survey to identify trends in the social game population as it pertains to energy literacy and climate change opinions. The survey also collected demographic information from participants, including age, ethnicity, and job position. Our objective was to explore the extent to which the social game population was heterogeneous in their demographic, energy literacy, and beliefs about climate change. The survey was emailed through a Google Form to all participants six weeks before treatments were first administered. 18 responses to the survey were collected of the 28 individuals studied, representing a 64% response rate. In order to look for relationships between question responses, we used exploratory data visualizations and multiple regressions.

4.3.2.2 Social Game Experiment

The experiment was conducted in the Campus for Research Excellence and Tech Enterprise (CREATE) tower, a building in Singapore. Interaction with study participants took place through an online platform where participants could monitor their progress in saving energy relative to others, receive tips on how to change their energy saving behavior, and, crucially, to schedule times in which certain plugs at their desk would be on or off. Office workers were voluntarily enrolled in the Game at the start of August, 2018.

The Game was structured in the following manner: first, a normal distribution was fit to the historical energy data of each participant to make energy savings relative to their own baseline. Then, for five periods of two weeks each, participants competed against each other to reduce their energy consumption. Amongst the top five of every round, a random number generator picked the first, second, third and fourth place winners, all of whom received prizes of various amounts.

We were interested in exploring the effect of different types of data visualizations on the user's engagement with the system and energy savings. To this end, two types of visualizations were created: an engineering-type bar chart (please see Figure 4.2), and a ambient-type visualization (please see Figure 4.3). The ambient-type was intended to communicate a sense of linear scale without appearing too formal. Although we developed it in house, we relied heavily on the tree design presented by Froelich et al in their 2009 ACM CHI Conference on Human Factors in Computing Systems Proceedings (Froehlich et al., 2009). All participants were emailed a treatment once a week, on either a Wednesday or a Thursday.[35] The emails consisted of Treatment A, which included the barchart and the following language:

[35]There was no pattern in whether a Wednesday or Thursday was chosen; it was due to technical limitations on the time of our software engineer.

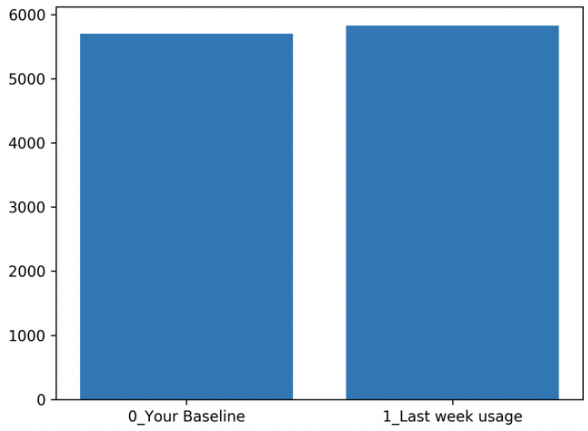


Figure 4.2: The "engineering type" visualization sent to Social Game participants as a treatment (Spangher et al., 2019a).

> Thank you for your participation thus far in the Social Game!
>
> Through participating in this game, your actions are being studied to learn more about how energy is used in office settings.
>
> Your ranking so far is:
>
> `__ of 27.`
>
> You progress with regards to last week and your baseline use is as shown:

Treatment B consisted of the same language and an image of the tree visualization, and finally a control email consisted only of the language.

The curious reader might wonder why we opted to include language in the control instead of simply not sending an email at all to a control group. Our experimental setup was modeled after the experimental setup noted in (Gerber et al., 2008). Here, the authors send a control postcard noting that the participants are being studied in order to deal with the *Hawthorne Effect*, a noted positive effect that occurs simply because the subject is aware of her being studied. Therefore, comparing the effects of a control whose email differs from a treated subject only by a visualization, we have in the difference of the responses the effect of the visualization itself.

The treatments are labelled as follows. Treatment A is an engineering-type visualization with the above language. Treatment B is an ambient-type visualization with the above language.

Participants were randomly selected to receive either Treatment A, Treatment B, or a control. Assignments switched three and six weeks following a step-wedge assignment protocol.

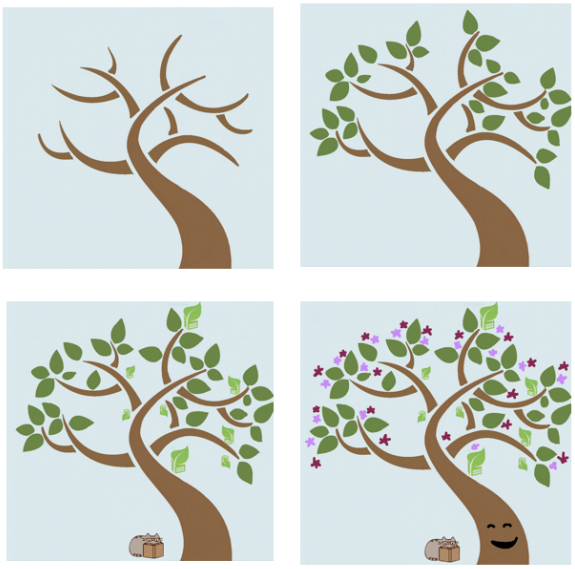


Figure 4.3: The "ambient type" visualization sent to Social Game participants as a treatment (Spangher et al., 2019b).

4.3.2.3 Post experiment survey

Following the conclusion of the Social Game, all participants were sent a questionnaire seeking feedback on the energy visualizations and the competition's impact on their energy consumption habits. Responses were compared with findings from the experimental data to analyze trends between the qualitative participant feedback and quantitative outputs from Social Game. Results from the post experiment survey were also used to frame the future direction of our experimentation.

4.3.2.4 Simulation of people's energy responses

Our ultimate goal is to make predictions about the effect of visualization type on total energy usage at the scale of an entire building, using the energy usage data we have collected from a small number of participants.

One way to use individual energy data to make predictions about energy use at the building scale is to construct a probabilistic model for individual energy use based on the individual energy data, and "simulate" the energy use of a building's worth of participants.

Here we introduce one such model, a Hidden Markov Model (HMM) for an individual's energy use per unit time (daily, for instance) (Baum and Petrie, 1966). An HMM is intended to be as simple as possible while capturing the effect that the visualization has on individual energy use.

Our model assumes that the energy use of an individual can be modeled at least roughly in the following way. In the absence of any kind of treatment, we assume that the individual's energy use per unit time follows some probability distribution, which is their *baseline distribution*. Some time after receiving treatment (that is, the energy visualization), we

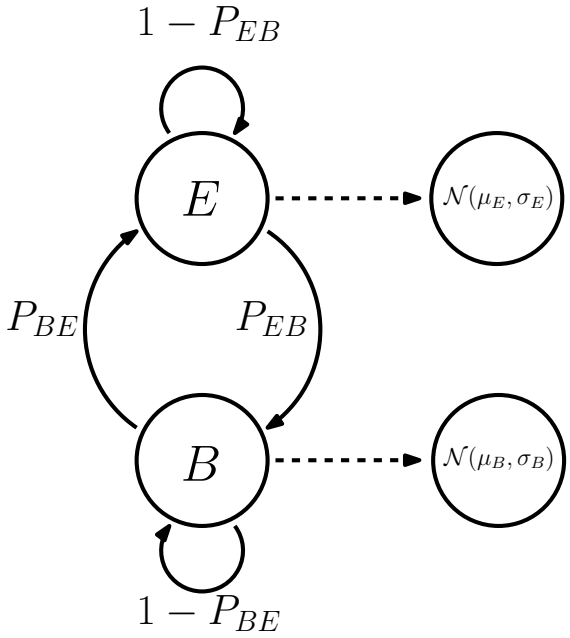


Figure 4.4: Diagram of the HMM model for individual energy use. The hidden states E and B represent whether the individual is following their energy-saving distribution (E state) or their baseline distribution (B state). The observations, all normally-distributed in this model, correspond to the energy use of the individual per unit time (Spangher et al., 2019a).

assume that the individual will modify their energy use in such a way that their energy use follows a new distribution, which is their *energy-saving* distribution. After spending some time in the energy-saving distribution, the individual eventually resumes their baseline distribution, and remains there indefinitely.

With this in mind, our model is an HMM with Gaussian observations, shown in Figure 4.4. The states correspond to whether the user is following their baseline distribution (state B) or their energy-saving distribution (state E). An execution of the HMM model represents a trace of the modeled individual's energy use over some time period. Since we have collected individual energy use over time for several participants, we can use this data to fit model parameters for each individual using an EM algorithm.

Under this model, how much energy does an individual save when they are influenced by the treatment? Without the treatment, the individual would remain in the baseline state the entire time, so the difference in energy use effected by the visualization comes entirely from the time spent in the energy-saving distribution. Assuming that the individual enters the E state once per treatment, and otherwise remains in the B state, the number of time units spent in the E state n_E is distributed geometrically, that is $n_E \sim \mathcal{G}(P_{EB})$. As such, the total energy used during the user's duration in the E state is a random sum of random variables, which we will denote as $\sum_{t=1}^{H} E_t$. By Wald's equation, the expected energy use during the energy-saving state is:

$$\mathbb{E}[\sum_{t=1}^{H} E_t] = \frac{1 - P_{EB}}{P_{EB}} \overline{E} \tag{4.1}$$

Had there been no treatment, the individual would have spent the same amount of time in the baseline distribution, using an amount of energy $\sum_{t=1}^{H} B_t$, whose expected value is:

$$E[\sum_{t=1}^{H} B_t] = \frac{1 - P_{EB}}{P_{EB}} \overline{B} \tag{4.2}$$

Therefore, the expected energy savings effected by the visualization for an individual is

$$\mathbb{E}\left[\sum_{t=1}^{H} [B_t - E_t]\right] = \frac{1 - P_{EB}}{P_{EB}} \left[\overline{B} - \overline{E}\right]. \tag{4.3}$$

For a more detailed look at the behavior of this model, in particular the effect of having models of multiple participants each with unique parameters, we turn to simulations, which we discuss in the next section.

4.3.3 Results

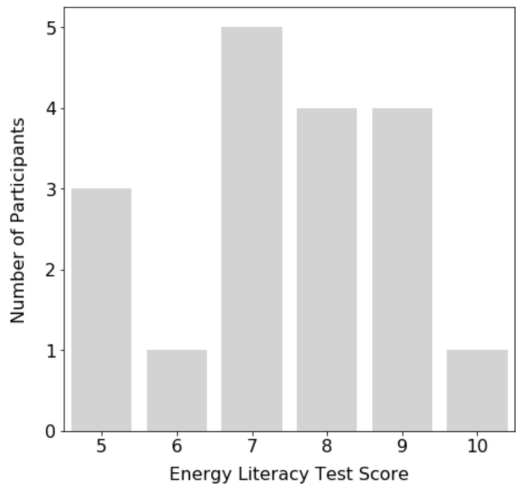
4.3.3.1 Pre-treatment survey

Distributions of scores on the energy literacy questions (Figure 4.5a) indicate a varied knowledge of climate related topics, although average participant scores were somewhat respectable at 68%. When pressed on whether Singapore is managing its energy correctly, participants responded across a uniform distribution centered at 6.4 points out of 10 (Figure 4.5b). This result indicates a spread in faith in the energy management of the Singaporean government.

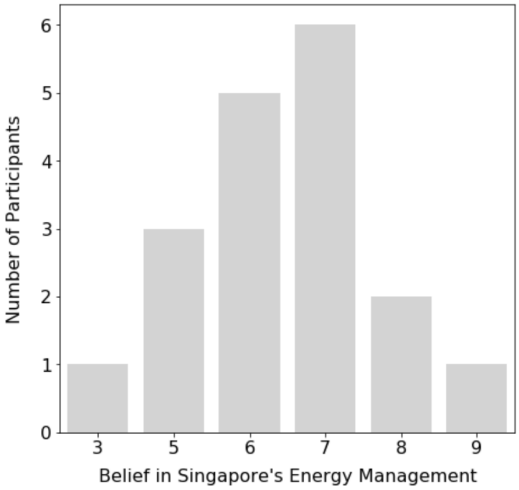
We completed a regression analysis on bivariate relationships in the data and present two takeaways. First, those who believed that the world should take collective action on climate change tended to also value scientists' portrayal of climate change as accurate (p=0.001) (Figure 4.6a). Second, those who believed Singapore is managing its energy correctly also tended to view it as a responsible actor in the world's climate goals (p=0.005) (Figure 4.6b).

Our results concur with other literature on Singaporeans' energy literacy and climate change opinions. Rosenthal and Ho note that most Singaporeans (84%) view climate change as somewhat or very serious (Rosenthal et al., 2013), which agrees with our observation that respondents tended to place an importance on the necessity for global climate action. Interestingly, a 2016 survey conducted by Singapore's National Climate Change Secretariat of Singapore (NCCS) found that most Singaporeans strongly believed that climate action should be driven by the government as opposed to individual actions (Jamil, 2017).[36] The NCCS results may explain the correlation in our survey between positive appraisals on Singapore's energy management and how responsible the Singaporean government is in the world's climate goals.

[36]Many attribute this trend to the so-called "nanny-state" syndrome, whereby years of state intervention have created a dependency on the government to address climate issues.

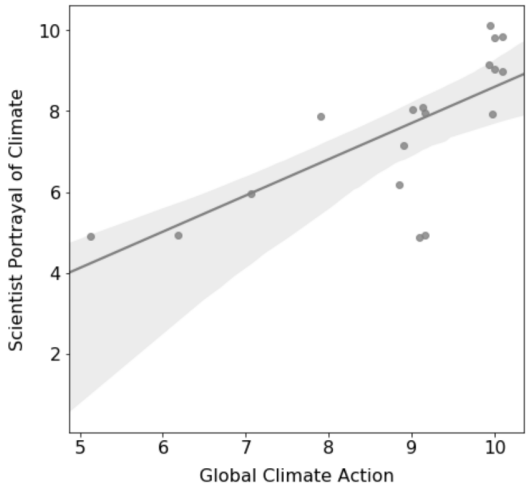


(a)

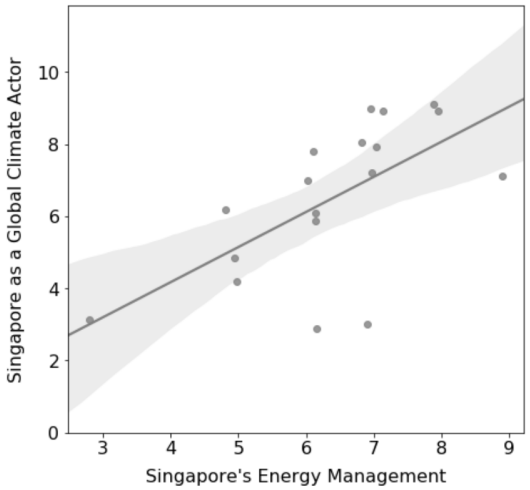


(b)

Figure 4.5: (a) Distribution of scores on energy literacy test survey questions amongst participants (out of 11 points) (b) Distribution in opinion on whether Singapore is managing its energy correctly amongst participants (10 indicates highest agreement) (Spangher et al., 2019a).



(a)



(b)

Figure 4.6: (a) Relationship between participant opinion on whether the world should take collective action on climate change and whether scientists' portrayal of climate change matches the phenomenon (10 indicates highest agreement) (b) Relationship between participant opinion on whether Singapore is managing its energy correctly and whether Singapore is responsible actor in the world's climate goals (10 indicates highest agreement) (Spangher et al., 2019a).

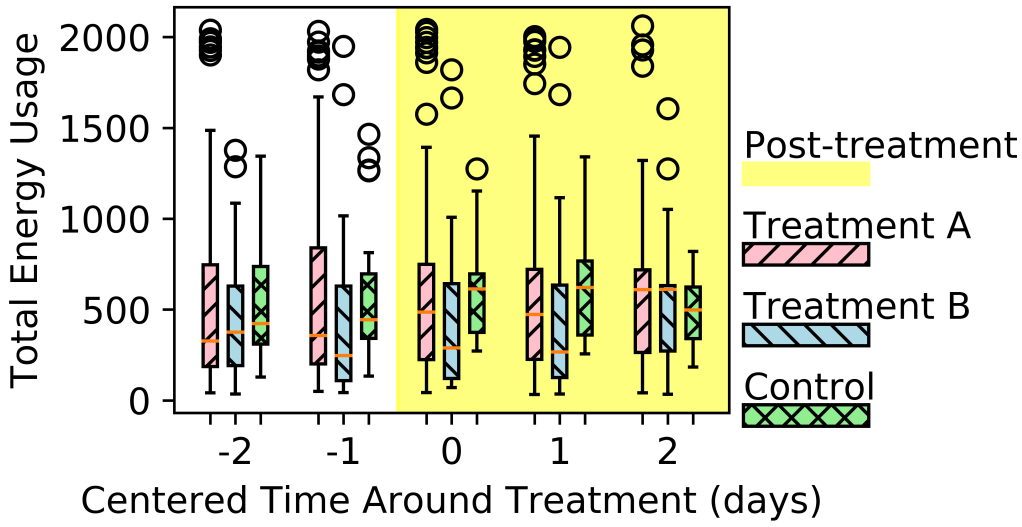


Figure 4.7: Summary of the main results from the experiment. Here we show averages of groups per day across the week, with the yellow section highlighting the part of the week which occurred after the treatment. The x-axis, "centered time around treatment" is the days translated such that the treatment day always occurs on day 0 (Spangher et al., 2019a).

4.3.3.2 Results from Regression Analysis

This section describes the results of an extensive exploratory analysis. First, we show a display of the data in figure 4.7. There are several items of note herein, which we will note and support through through regression analysis: (1) working days are generally indistinguishable from each other, (2) people tend to engage with the Game less, even though energy savings persist over the duration, (3) greater logins to the online platform correlated with greater energy savings (4) numerous predictor variables and interactions significantly predict daily aggregate energy use (E), which we present in a regression output shown in Figure 4.2 (Section 4.3.3.2).

Main regressions results of note:

- **Weekend Indicator (WI): Uniformity of weekday vs. weekend**: First, we observe that the working days of the week (i.e. Monday through Friday) are visually indistinguishable from each other with respect to E, yet significantly different than weekend usage. This was further confirmed in a regression analysis predicting E; a categorical variable treating each day of the week as an independent predictor found four of the five weekdays to have insignificant[37] predictive power with respect to the

[37]When we say a predictor was insignificant in predicting a response, we use conventional definitions of

response, after the variation from the first weekday was controlled for.

- **Treatment Number (TN): Effect of repeated treatments**: Second, we observe that participants login less after multiple treatments, but save energy consistently throughout the experiment. We tested multiple treatments by coding a TN: an integer value that counted the number of times an individual had been treated. TN returned as significantly negative when included in a regression predicting a users' Login Count (LC), which was an integer value counting how many times a user logs onto their online profile. The significant relationship implied that there was a significant decrease over time with how often users interacted with the Game. However, despite this, TN was insignificant when predicting E across time, which implies that the energy habits that the Social Game encouraged persisted throughout the treatment period.

- **Login Count: Effect of interaction with the platform**: Third, we observe a positive relationship between LC and E. A safeguard was put into place to ensure that LC was not unrealistically inflated; the count could increment only once every five minutes. Regressions showed that Login Count robustly predicted E when controlling for a variety of factors such as weekday vs. weekend and TN. As expected, the regression coefficient in LC was negative, implying that the more a user logged into their system, the more energy they tended to save.

- **Main treatment effect on energy savings**: Finally, we observe numerous significant predictors of E, which, again, is our main response of interest, measuring daily aggregate energy from each plug load of each desk. Given that E is a periodic time series, we segmented the resulting data into three days before treatment and three days after treatment. We include the previously noted variables, testing for a multiple regression of the following form:

$$\mathrm{E} \sim \mathrm{TN} + \mathrm{WI} + \mathrm{LC} * \mathrm{TT} + \mathrm{TI} : \mathrm{TT} + \text{out of office} \tag{4.4}$$

  Here, E is predicted by TN, WI, an interaction between LC and Treatment Time (TT): i.e., A, B, or control, and an interaction between TT and "Treatment Indicator" (TI). TI is an indicator variable that is 0 for three days before each treatment and three days after, which measures the average energy use that persists after a few days. Note that the equation follows standard `R` syntax, where $*$ is an interaction with terms considered independently, and : is an interaction with only bi-variate terms considered.

- **Interaction terms**: The first interaction, LC:TT, attempts to elucidate an effect between people's LC and their TT that is above the average effect of LC; i.e., does a different type of treatment engender different impact from interacting with the system? The second interaction, TI:TT attempts to provide the average treatment effect for different treatment types several days after treatment. The final term, Out of Office (OO), is an indicator that keeps track of whether or not a person was physically in the office, which we guessed from looking at distributions of energy demand.

5% significance in a frequentist OLS framework.

Dep. Var:	Energy	**R-sq:**	0.284
Model:	OLS	**Method:**	Least Sq
No. Obs:	1223	**Df Resid:**	1213
Df Model:	9		

	coef	**std err**	**P>|t|**
Intercept	599.27	34.03	0.00
TN	-4.43	5.54	0.42
WI	77.73	34.157	0.023
LC	-71.98	18.30	0.00
LC:TT[Ambient]	22.21	29.38	0.45
LC:TT[Control]	-64.17	56.52	0.26
TI:TT[Engineering]	34.93	27.75	0.21
TI:TT[Ambient]	-82.67	37.07	0.03
TI:TT[Control]	19.58	54.56	0.72
OO	-610.19	33.20	0.00

Table 4.2: Summary of the OLS regression results (Spangher et al., 2019a).

Results of this regression conform to our hypotheses. TN is insignificant, which implies that users' energy savings does not majorly diminish during the extent of the game. Whether or not a day is a weekend significantly predicts E, as we expect: during the weekends, desks use 77Wh less energy than during weekdays after controlling for the other factors. LC remains significantly negative in this regression; for an average increase of one login, we expect energy use to be approximately 71 Wh less per day. The TT interacted with LC, meanwhile, do not significantly predict total energy, implying that the treatments do not effect the relationship between interacting with the platform and reducing energy use. The TT interacted show a strongly significant negative effect of treatment B: for the three days following treatment B, we can expect on average 82 Wh less energy use after controlling for other variables. When people are out of the office, we expect 610 Wh of energy use less (this variable is highly collinear with, but not entirely explained by, WI.)

4.3.3.3 Post experiment survey

Through the post experimental survey, we endeavored to explore whether the quantitative results we measured were perceived by the participants.

First, we queried the participants on the negative effect of TN on LC (see section 4.3.3.3): people seemed to interface less with the game through repeated treatments. When asked whether interaction lessened over time, two thirds of the respondents answered "Yes", which confirms the quantitative findings.

Second, we queried the participants on our observation that the tree visualisation seemed more engaging. Of the tree, a respondent stated:

"The tree visualization provided insight and view on energy usage."

Meanwhile, of the bargraph, a respondent stated:

"It was clear but I did not feel it to have any distinctly engaging quality."

We received several comments similar to the ones shown above.

Finally, we received confirmation that the format of the treatment was sound, addressing a concern that the treatment emails may have often been left unread. When asked how often they opened the weekly email (the treatment), 83% of respondents said that they opened it "all the time" or "almost all the time". This was significantly different from responses to being asked how important the email was to their interactions with the Game: responses were uniformly distributed across a five point scale from "unimportant" to "important", confirming that while the emails may not have been impactful, at least that the method of treatment delivery was sound.

4.3.3.4 Results from the Simulation

We will now argue that a scale up in simulation is appropriate. Our test population reflects the wider Singaporean office worker population enough that behavioral dynamics from the test population can be seen as generalizable. This is important because individual energy savings only significantly improve sustainability when the savings are part of a collective shift in behavior.

We claim generalizability for a number of reasons. First, the Social Game was run in a general office with standard desk equipment[38] and a normal 40 hours per week working schedule (Monday-Friday). Second, the distribution of energy literacy's width (see section 4.3.3.1) indicates that the participants are not uniformly experts on the subject — as one might expect in a general office. Third, baseline energy consumption varied widely amongst participants, indicating a spread in energy consumption as well.

We ran our simulation with 1000 simulated actors across 100 days. In total, the energy savings were estimated to be 1.1 megawatt Hour (MWh), or approximately 1 kWh a person over these days. Each person was in the "energy saving state" for a different amount of time, so producing a estimate of savings per day would be a less meaningful task.

4.4 Energy Reduction Experiment 2: Persistence of energy reduction behaviors following an intervention

4.4.1 Background on Persistence of Energy Savings

An extensive study on energy reduction competitions found a mean electricity savings of 5% *during the program period* (Vine and Jones, 2015). Other studies have created energy

[38]Laptop or desktop computer, 1-2 monitors, desk lamp, etc.

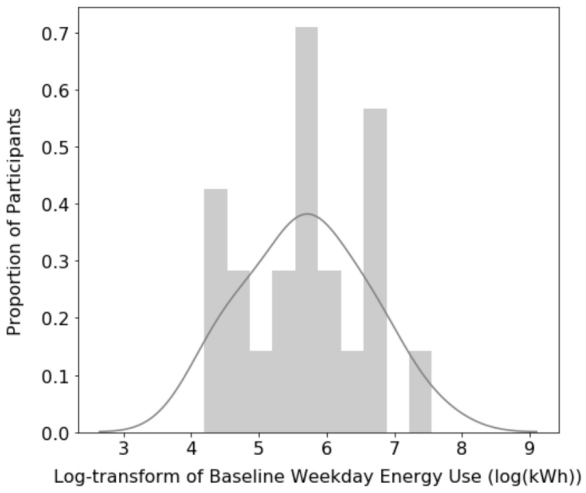


Figure 4.8: Distribution of participant weekday baseline energy consumption in log(kWh) (taw, 2020).

competitions and determined that gamified frameworks motivate participants to save energy (Ratliff et al., 2014; Dimitriou et al., 2018).

However, selected works have cautioned against the persistence of energy reduction schemes past program periods. (Prindle and Finlinson, 2011) observed that energy competitions are traditionally perceived as problematic because they require costly prizes, are zero-sum, lack sustainability, and lack substantial evidence on the persistence of effects. (McKenzie-Mohr and Schultz, 2014) further suggested that while extrinsic rewards can produce large changes in behavior, they may also induce various side effects when retracted including reversion to prior habits and "overjustification" effects, whereby intrinsic motivation to perform a task decreases. Thus, we postulate that currently, reliance on extrinsic motivators (i.e. prizes) prohibitively increases the cost of energy competitions, and intrinsic motivation to save energy (developed through social norms, personal satisfaction, and heightened awareness) varies based on individual characteristics once extrinsic rewards are removed. Many studies have accordingly concluded that further research is necessary to determine how energy competitions may address these challenges to support behavior change beyond the short term (Vine and Jones, 2015; Geelen et al., 2012).

Recent evidence suggests that short-run energy policies can shape behavior in the long term (thereby extending savings beyond the program period). While the evidence is collected in a unique context (population, type of intervention, presence of crisis), we seek to quantitatively understand if the measured effect generalizes to energy competitions. (Costa and Gerard, 2021) conducted an extensive study on the importance of hysteresis for the welfare evaluation of corrective policies. The paper examined residential electricity use in Brazil during the 2001 electricity crisis and determined that energy savings produced by a 9-month-long policy persisted for 12 years. Specifically, the energy-saving policy realized half of the short-run impact (–23%) in the long term (–12%). The authors noted that the mechanism of hysteresis was changes in utilization habits, and formulas presented in the paper could inform the design of persistence experiments in other contexts.

From this, we endeavor to answer two questions about the persistence of energy reduction behaviors that— provided substantiation— would begin to reconcile issues of cost and response heterogeneity in the energy competition environment.

First, do energy saving competitions— in lieu of energy policies— benefit at all from savings persistence after the program period? Here, we hope to uncover extended energy savings that increase competition cost-efficiency (extrinsic incentive dollars / kWh saved). Currently, persistence in behavior programs remains a great unknown and most competitions only assume savings to be counted while behavior change efforts are active. (Vine and Jones, 2016) methodically selected twenty energy competitions to review and found only anecdotal information on the persistence of energy savings. While competitions did not measure persistence, the authors wrote that certain programs did study the persistence of actions. For example, they cite that the winning fire department in El Paso's Energy Savings Challenge developed sustainable air-conditioning and lighting habits. Though qualitative, the habits are indicators for persistence given that they are difficult to break (Becker and Murphy, 1988). We are unaware of a study that attempts to quantify the persistence of energy savings in an office space following a competition. The literature accordingly calls for more contemporary and thorough research into savings persistence to assess the full value of energy saving interventions (Hoffman et al., 2015).

Second, do reductions in energy consumption following a competition persist differently for different people? Here, we attempt to provide greater clarity on the variation of intrinsic motivation to save energy among participants after extrinsic rewards are removed. However, energy forecasting of individual participants is difficult to model given the complex stochastic structure of consumption (Shchetinin, 2019). Therefore, we attempt to group participants on pre-competition energy signatures and evaluate the post-competition predictive power of the resulting clusters (i.e. whether certain clusters end up displaying greater savings persistence). We hope to isolate participants that are more likely to exhibit savings persistence post-competition and recommend behavior change efforts be tailored to them. The literature provides empirical justification for a clustering-based approach to energy forecasting. (Mikulik, 2018) performed a clustering analysis based on k-means to isolate daily energy demand and meteorological patterns in an office building. The study concluded it was possible to distinguish three daily energy demand patterns and established that clusters could be used in forecasting models.

Reliable persistence measurements and predictive clustering outcomes would promote cost-effectiveness and personalization in the energy competition environment. This paper attempts to build towards that end. We will next examine the results of our tests for persistence on the complete and cluster subdivided datasets. Finally, we conclude with an analysis of the results and future directions.

4.4.2 Methods

4.4.2.1 Clustering

Per our second research question, we endeavor to understand whether certain groups of competition participants exhibit greater savings persistence than others. This requires a clustering regime that groups participants with similar pre-competition energy use patterns.

Furthermore, the ideal regime will also produce groupings that exhibit similar post-competition energy use patterns (this data is not visible to the clustering algorithm). Here, we evaluate six regimes for clustering participants on pre-competition energy consumption. This section covers our methods for narrowing down a tool that evaluates the post-competition predictive power of the clustering regimes, selecting a clustering regime by means of that tool, and finally applying the selected clustering regime to the dataset. The outcome is three distinct groups of participants that we are then able to each test for energy saving persistence. Clustering describes a process in which a given set of data points are grouped in k clusters such that:

- points within each cluster are similar to each other (homogeneous)
- points from different clusters are dissimilar (heterogeneous)

Data points are typically in a high-dimensional space and similarity is defined using a distance measure such as Euclidean, cosine, or Jaccard distance. Here, we experiment with both probabilistic– so-called "soft" clustering– and its counterpart– "hard" clustering. Hard clustering assigns each data point to a single cluster with probability 1. For example, a participant would only be assigned to one of k groups. Conversely, soft clustering assigns each data point with a probability or likelihood of belonging to each of the clusters considered. Given our previous example, each participant would be assigned a probability of belonging to each of the k clusters.

We attempt to cluster on weekly increments over the pre-competition period to capture nuanced behavior. Mean hourly energy usage per participant is normalized across each of the weekly periods to prevent the clustering algorithm from simply picking up on total energy consumption. Additionally, we extract auxiliary features in the form of peak characteristics from each participant's daily energy usage. These features include the number of peaks, average peak height, average peak width and average peak prominence per day.

We elect to evaluate time series k-means and k-shape clustering algorithms (both "hard" clustering methods) on the weekly pre-competition energy data and test k-means ("hard" clustering) and Gaussian Mixture Model (GMM) ("soft" clustering) on peak characteristic features. Hence, we considered six different clustering regimes.

For all clustering schemes, we used the "elbow" method to pick the number of clusters (Ketchen and Shook, 1996), which ended up being k=3 on all models. While clustering is performed on weekly intervals, we seek to prevent the clustering results from being influenced by the assignments from one specific week. Therefore, we first attempt to aggregate results using the mode of cluster assignments across all weeks of the pre-competition period. However, this leads to "undecided" results when a player is equally often assigned to two or more clusters. Instead, we elect to use the percentage of weeks a participant is assigned to each cluster to alleviate potential uncertainty (i.e. 50% of weeks in cluster 0, 20% of weeks in cluster 1, and 30% of weeks in cluster 2).

To select the best clustering regime, we diverge from traditional cluster quality metrics found in the optimization problems that clustering regimes solve, such as in-cluster homogeneity. We do not seek to test how well the clustering regimes can separate participant data, but rather how useful the clusters produced are for predicting post-competition energy consumption. Thus, we decide to measure this predictive power by evaluating a simple linear

Dependent variables	Independent variables
$\overline{E_D}$, Mean hourly energy usage during the game	Clustering regimes evaluated: 1. k-means 2. k-shape 3. GMM on peak characteristics 4. GMM on normalized peak characteristics 5. k-means on peak characteristics 6. k-means on normalized peak characteristics
$\overline{E_A}$, Mean hourly energy usage after the game	
W, Overall game performance	
R_{BD}, Ratio of energy used during vs before game	
R_{BA}, Ratio of energy used after vs before game	

Figure 4.9: Results of linear regression R^2 based on the clustering method and dependent variable (taw, 2020).

regression whereby cluster assignments are used to predict different dependent variables in terms of their coefficient of determination, R^2. The primary dependent variable of interest is mean hourly energy usage after the game ($\overline{E_A}$)– though other variables are tested– and the structure of the regression is of the following form:

$$\overline{E_A} \sim C_0 + C_1 \tag{4.5}$$

Here, mean hourly energy usage of a participant after the game is predicted by the percentage of pre-competition weeks a participant is assigned to cluster 0 (C_0) and cluster 1 (C_1) by a particular clustering regime. Cluster 2 is omitted from the regression to serve as a reference and to avoid multicollinearity.

Since each clustering regime is trained on pre-competition energy data, the regression aims to show how well observation of characteristics occurring before the energy competition can predict energy usage after the competition. An in-depth explanation on why the regression model is chosen and how dependent variables are derived can be found below.

4.4.2.2 Model Explanation

We elect to build a linear regression model to measure the predictive power of the cluster assignments produced by each of the six clustering regimes. We choose linear regression mainly due to its simplicity and interpretability. We find no apparent need to capture more complicated (e.g. non-linear) relationships in the data. To still account for the possibility of such relationships, we experiment with Support Vector Machine (SVM) and classification and regression tree (CART) models. However, the results do not contradict the simpler regression results, so the regression is presented to preserve readability. Figure 4.9 displays five different dependent variables used in the regressions.

Mean hourly energy is defined as:

$$\overline{E_D} = \frac{TE_D}{TH_D} \text{ and } \overline{E_A} = \frac{TE_A}{TH_A} \tag{4.6}$$

Dependent variable	k-means	k-shape	GMM peak charac.	Normalized GMM peak charac.	k-means peak charac.	Normalized k-means peak charac.
$\overline{E_D}$	0.63	0.06	0.14	0.02	0.39	0.39
$\overline{E_A}$	**0.67**	0.05	0.06	0.01	0.30	0.30
W	0.15	0.05	0.04	0.02	0.17	0.17
R_{BD}	0.10	0.18	0.42	0.02	0.05	0.05
R_{BA}	0.02	0.06	0.11	0.08	0.01	0.01
Mean R_2	0.31	0.08	0.15	0.03	0.18	0.18

Figure 4.10: Results of linear regression R_2 based on the clustering method and dependent variable (taw, 2020).

Where $TE_{D/A}$ is the total energy consumed by a player during (D) or after (A) the competition and $TH_{D/A}$ is the period duration in hours during or after the competition. Overall competition performance is measured by the total number of times the participant is in the top 5 winners throughout the two-week Social Game periods.

The ratio of energy savings is defined as:

$$R_{BD} = \frac{\overline{E_D} - \overline{E_B}}{\overline{E_B}} \text{ and } R_{BA} = \frac{\overline{E_A} - \overline{E_B}}{\overline{E_B}} \tag{4.7}$$

where $\overline{E_B}$ is the mean hourly energy usage before the game and $\overline{E}_{D/A}$ is the mean hourly energy usage of a participant during or after the competition.

4.4.3 Results

4.4.3.1 k-Means Clustering

: The k-means clustering regime (on weekly energy values) results in assignments with the highest R^2 value on the post-competition dependent variable of greatest interest ($\overline{E_B}$). An intuitive understanding on why the time-series variant of k-means performs best might be its simpler nature, which complements the low-dimensional (only hourly energy consumption) data well. Furthermore, k-means guarantees convergence (at least to a local optima) and is known to work best for the fixed-size data the experiment provides. Finally, a few of the clustering regimes have requirements (i.e. normally distributed features for GMMs) that may not be satisfied, thereby contributing to the gap in predictive performance.

We use the k-means clustering algorithm to group participants based on cluster assignment. The groups are used to partition the full Social Game dataset in preparation for persistence testing as described in further detail below.

Indicator name	Start date	End date	Abbreviation
Before Social Game 1	06-03-2018	06-30-2018	BSG_1
Before Social Game 2	07-01-2018	07-28-2018	BSG_2
During Social Game 1	07-29-2018	09-01-2018	DSG_1
During Social Game 2	09-02-2018	10-13-2018	DSG_2
After Social Game 1	10-14-2018	11-10-2018	ASG_1
After Social Game 2	11-11-2018	12-05-2018	ASG_2
After Social Game 3	12-20-2018	01-16-2019	ASG_3

Figure 4.11: List of indicators used to demarcate each period of the social game (taw, 2020).

4.4.3.2 Regression Analysis for Determining Persistence

This section describes the results of extensive data exploration and analysis on the energy competition dataset. First, we present a more granular display of the data that captures trends in energy consumption following the Social Game. Then, we set up and describe the regression outputs in three stages: (1) we validate several indicators and variables that predict energy use (2) we present results of the persistence regression without clustering (3) we subset the dataset into individual clusters and re-run the persistence regression on each cluster subset for greater insight into cluster significance.

We begin this section by describing and validating each of the indicators that are included in the regression to control for confounding effects. Then, we analyze outputs of the persistence regression without clustering followed by the persistence regression on subset cluster data.

- **WI**. We observe that consumption during the weekends is significantly lower compared to weekday usage with respect to energy use. We confirm this in a regression analysis predicting energy use with an indicator for weekends (p¡0.000).

- **Game period indicator**. We hypothesize that energy usage will change depending on whether a date falls before, during, or after the Social Game. We further predict that within these periods, energy usage may shift and should be analyzed. Therefore, we split the raw data into 7 periods of roughly equal number of days to preserve consistency (see Figure 4.11).

  We assign indicators for each period in the dataset and validate that energy use is statistically different before, during, and after the competition. We split periods with equivalence so they may be statistically comparable to each other in the regression.

- **Out of Office (OO) indicator.** We detect office workers are not present at their desk during certain periods. No log of office presence is made available, so we attempt to resolve this using participant energy signatures. Office presence is inferred from observing individual distributions of energy demand and marking a threshold for in-office versus out of office. Thus, every participant is assigned an indicator each day

with value zero if energy usage falls below the threshold for that day and value one if threshold is met or exceeded. We observe energy consumption when office workers are out of office is significantly lower than in-office readings ($p < 0.000$).

4.4.3.3 Persistence Regression without Clustering

We define Before Social Game (BSG), During Social Game (DSG) and After Social Game (ASG). We now test for a multiple regression given previously noted indicators of the following form:

$$E \sim BSG_1 + BSG_2 + ASG_1 + ASG_2 + ASG_3 + WI + OOI \tag{4.8}$$

Here, energy usage (E) across all periods is predicted by each period of the dataset (DSG_1 and DSG_2 are omitted to serve as a reference and to meet the assumption that there is no exact linear relation among the independent variables) along with the Weekend Indicator (WI) and an Out of Office (OO). The results of the regression (see Figure 4) agree with our hypothesis. Coefficients of energy usage during both before game periods are significant ($p < 0.05$) and similar in magnitude, indicating participants consumed 49 kilowatt hour (kWh) (BSG_1) and 54 kWh (BSG_2) more energy pre-competition versus during competition (8-9% savings). This result is expected as the literature review highlighted previous findings on the efficacy of energy saving competitions. ASG_1 and ASG_2 coefficients are insignificant ($p > 0.05$), suggesting that energy consumption in these two periods immediately following the energy competition did not deviate significantly from the reference value (DSG_1 and DSG_2). However, statistical significance is observed on the ASG_3 coefficient ($p < 0.05$). The coefficient relative to the reference value (DSG_1 and DSG_2) is positive and similar in magnitude to both before game periods, indicating a return to baseline consumption. Thus, energy savings from during the game only appear to persist through the first two periods immediately following the game.

4.4.3.4 Testing of Persistence effect on Data Subset by Cluster

Here, we test for a multiple regression with previously noted indicators on three subsets of data (one subset per possible cluster), each limited to data from the top seven participants with the greatest percentage of time spent in the cluster corresponding to the subset. The form of the multiple regression is identical to the prior persistence regression without clustering (see Equation 2). The results of all three multiple regressions are summarized in Figure 4.13 and highlight a couple points of interest. The magnitude of intercept in cluster 0 and cluster 1 regressions are much lower than the cluster 2 regression intercept, indicating those with higher baseline energy usage were placed in cluster 2. BSG_1 is insignificant across all three clusters for reasons that are not immediately clear but indicate that baseline energy consumption well before the game was not significantly different from during game energy usage levels. Significance is observed on the BSG_2 coefficient in cluster 0 and cluster 1 while the coefficient remains insignificant in cluster 2. Immediately following the game, cluster 2 exhibits a strong movement away from energy saving behavior, indicating a lack of persistence. This finding is replicated well after the game in ASG_3. Meanwhile, cluster 0 and cluster 1 demonstrate energy savings similar to in-game patterns in ASG_1; however, cluster 0 reverts

OLS Regression Results

Dep. Variable:	TotalEnergy	R-squared:	0.257
Model:	OLS	Adj. R-squared:	0.256

	coef	std err	t	P>|t|
Intercept	547.2781	11.960	45.759	0.000
BeforeGameIndicator_1	49.4911	21.526	2.299	0.022
BeforeGameIndicator_2	53.6618	21.545	2.491	0.013
AfterGameIndicator_1	20.5026	20.820	0.985	0.325
AfterGameIndicator_2	-9.2779	21.211	-0.437	0.662
AfterGameIndicator_3	48.5180	21.746	2.231	0.026
Weekend_Indicator	71.9578	18.561	3.877	0.000
OutOfOffice	-596.9108	18.242	-32.722	0.000

Figure 4.12: Summary of OLS persistence regression results without clustering (taw, 2020).

Table 4. Summary of OLS persistence regression results on subset data

Variable	Cluster 0	Cluster 1	Cluster 2
Intercept	187.79 (0.000*)	226.77 (0.000*)	1032.46 (0.000*)
BSG_1	-7.17 (0.335)	-20.48 (0.056)	40.47 (0.333)
BSG_2	20.99 (0.005*)	40.36 (0.000*)	24.83 (0.554)
ASG_1	7.36 (0.296)	**-4.07 (0.682)**	92.74 (0.027*)
ASG_2	15.23 (0.035*)	**6.65 (0.515)**	34.55 (0.422)
ASG_3	22.54 (0.002*)	**11.55 (0.266)**	154.23 (0.000*)
WI	-31.10 (0.000*)	-36.76 (0.000*)	-94.59 (0.004*)
OOI	-164.57 (0.000*)	-192.75 (0.000*)	-954.01 (0.000*)

*Denotes significance at the $p < 0.05$ level

Figure 4.13: Summary of OLS persistence regression results on subset data (taw, 2020).

to baseline level energy consumption during ASG_2 and ASG_3 periods, while cluster 1 exhibits persistence of energy savings at during game levels up to and including the ASG_3 period. Further analysis on cluster 1 shows that relative to BSG_2, participants maintained energy savings of 17% (44 kWh) in ASG_1, 13% (34 kWh) in ASG_2, and 11% (29 kWh) in ASG_3.

4.4.4 Conclusion

We conclude that in-game energy savings of roughly 8% are persistent for two months after the Social Game followed by a reversion to pre-competition baseline consumption during the third month. Furthermore, we conclude that clustering successfully isolates a group of participants who uniquely maintain energy savings of 17% immediately following the game followed by a decline to 11% savings during the third month.

Chapter 5

Conclusions

5.1 Summary

Climate change requires a radical and complex transition in the way our energy sector generates and uses energy. Solar and wind energy are leading carbon-free power sources, but they are non-dispatchable, meaning that an operator cannot control when their generation occurs. As a result, their growth may be stymied unless supporting technologies can change how the rest of our energy system responds to their generation. Demand response is an important tool in a suite of supporting technologies that help smooth the introduction of non-dispatchable energy into the grid. The more effective demand response signals (prices) are on responses (deferring of energy by appliances and building systems), the more quickly solar and wind may decarbonize our energy system.

Demand response signals may be amplified by advanced controls, of which reinforcement learning is a prime example. We present two different environments for testing demand response price-setting at different levels of the grid. The first environment is MicrogridLearn, an environment that transmits prices to a collection of buildings and learns to set prices in a way that better shapes the energy and lowers energy costs for all buildings in its purview. The second environment is OfficeLearn, an environment that simulates behavioral response to prices. Using these two environments, we identify six key challenges in moving RL from simulation to reality, and present RL strategies, some novel, to overcome them. First, RL controllers implemented in the real world need to be data efficient: we present Offline-Online RL, Surprise-Minimizing RL, and extrinsic and intrinsic planning as solutions. Second, RL in the real world should be robust and guarantee safe action: we present a novel method, the *guardrails* planning model, and demonstrate that using a conservative decision process with a distributional prediction can help learning. Third, RL may get stuck in local optima: we present meta-learning over domain randomization as a technique to ensure agent robustness. Fourth, agents may be attacked: we demonstrate a novel adversarial attack on RL and present a defense. Fifth, energy applications may require real-world RL to protect privacy and generalize to new subdomains easily; we present the first ever application of Personal Federated Hypernetworks (PFH) to RL to accomplish both tasks. Finally, hyperparameter sweeps may entail large data consumption; we present a regression analysis of hyperparameter sweep values to give a sense of hyperparameter-parameter strength.

Finally, we discuss how RL may be implemented in experiment. We give a prospective experiment plan. We present an API to connect the RL controller to the real world. We then discuss two prior experiments run in the same office setting: an A/B test of two different energy visualizations, and a measure of the persistence of the effects of energy reduction after the experimental period ended.

Our work contributes to societal knowledge in the following ways. We are the first to propose the use of RL for price-setting in energy systems, and we are the first to propose a Social Game as a mechanism to incentivize price sensitivity in an office setting. Of our RL methods, we are the first to propose adversarial poisoning during train time of algorithms. We are also the first to propose the use of personal federated hypernetworks for training new RL agents. Our other methods have been inspired by similar implementations in other fields, but are novel to the communities and application spaces in which we operate.

We hope that, from our work, the community may continue to iterate on RL architectures for price-setting, and may implement these techniques in experiment.

5.2 Future Work

5.2.1 Extensions on MicrogridLearn Control

Safety: An RL controller might generate prices that result in infeasible operation. We tackle this by using utility-set prices to enforce limits on the aggregator prices $\vec{p}_{\text{agg, s}}, \vec{p}_{\text{agg, b}} \in [\mathcal{P}_s, \mathcal{P}_b]$. An additional safety check would involve validating these prices through a single back-and-forth communication with prosumers to ensure that their operation does not exceed system limits. Probabilistic guarantees for RL controllers can be used (Bacci and Parker, 2020). Additionally, supervised RL can help guarantee safety at the outset (Rosenstein and Barto, 2002).

Efficiency: While the large number of training iterations represent a greater computational burden than iterative pricing methods, the RL controller will reduce the computational burden on each individual prosumer by eliminating the need to construct forecasts or demand/supply bids. Further, each prosumer's computations occur in parallel and adding new prosumers to the aggregation does not increase the problem complexity.

Robustness: Adversarial training is a useful method to construct robust RL controllers (Pattanaik et al., 2017), and can be extended to our setup.

Optimality: Research has been done in bounding the sub-optimality of RL policies using policy certificates (Dann et al., 2019), and research in this area can help provide guarantees for our RL controller as well.

Practical Implementation: The training iterations needed are a barrier to real-life use of an RL controller in a prosumer aggregation. However, there are numerous enhancements we propose to address this. First, *meta-learning:* a large part of the training can take place in a simulation environment which can use a rules-based heuristic as a starting point, and use exogenous parameters to create a distribution of unique systems to train on (i.e. domain randomization.) A technique like Model Agnostic Meta-Learning (MAML) can train in the different simulations to approximate a starting distribution for policy network weight initializations (Finn et al., 2017b). However, the accuracy of the simulation model will determine how effective the RL controller is when it transitions to an actual aggregation. Second, *planning*: a Dyna-like auxiliary model, either generative (i.e. GANs) or predictive (i.e. regression or neural nets), could train with the agent and help augment data (Sutton, 1991b). Third, *offline learning* can incorporate data from other microgrids using techniques in the causal methods literature to help adequately perform the data fusion necessary(Forney and Bareinboim, 2019). However, validating these ideas will require more work.

As prosumer aggregations grow larger, they may participate in wholesale energy markets instead of purchasing energy from utilities, and will have to adapt to variable energy market prices as well as having to generate estimates of their own consumption and production. RL algorithms have been previously used to optimize market participation, and these would be complementary to the model we present here.

5.2.2 Variants of the OfficeLearn MDP

We plan to offer the user the choice between a step size that is a day's length and a step size that is an hour's length. The alteration can provide a more efficient state space representation

that provides for a fully observable MDP for the agent, as well as a longer trajectory for action sequences (i.e., ten steps for every trajectory to determine the ten hours rather than a single step producing all ten hours), at which RL tends to excel.

5.2.3 Future Work in the Multi-MicrogridLearn Setup

5.2.3.1 Proposed Work: Transactive multi-agent RL for Distributed Energy Price Localization

Given that we have shown that a similar price-setting RL controller can be financially viable at both the grid level and the building level, i.e. two different levels of the grid hierarchy, we propose to expand the scope and complexity of our environment. What is the cumulative effect of a network of RL controllers creating, essentially, localizing prices?

We propose to study the multi-microgrid environment above. As a "peer-network" of controllers, a first solution would be to operate in parallel, with each receiving fixed input prices from the utility. However, there emerges a natural hierarchy: we propose to situate a meta-level RL controller on top of the level of microgrid controller; an aggregator of aggregators. This meta-level RL agent may set a global set of prices in addition to the utility's set of prices. In addition, we can populate a level of controllers below the leaf: each building, or at least one office building, may have a controller implementing a Social Game. In this way, we may achieve greater coordination by taking advantage of environmental structure.

5.2.3.2 Proposed Work: Fairness in Controls

Assuming that a microgrid controller covers a large enough area, we can be certain that different socioeconomic strata are covered. If our RL controller sees a non-response by a community to higher energy prices in an hour, it may be unable to determine whether the reason is that it was a wealthier community indifferent to paying higher prices, or if it was poor and had already deferred all the non-essential load. Thus, it would not know whether to continue to increase the price or decrease the price. Unfortunately, it is the poor community that stands to lose more if the controller makes a mistake. We propose now to simulate microgrids socioeconomic status as observable meta-data. We propose to study differences in behavior between controllers and intervene in architecture where necessary. For more reading on fairness in AI, please see (Mehrabi et al., 2021).

5.2.3.3 Proposed Work: Vehicle to Grid Simulation

With some modifications of the action space, we can simulate "vehicle to grid" interactions between the microgrids. A "vehicle to grid" deployment is a strategy in which electric vehicles charge in one microgrid, and then drive over to another microgrid to release energy in times of severe peak demand (Liu et al., 2013). We can easily extend our simulation to include an action to deploy the fleet between grids to trade energy. Vehicle deployment may have a significant effect on the prices that agents learn and thus have a significant effect on the reward that the agent is able to attain.

5.2.3.4 Proposed Work: Exotic Cryptocurrencies for Trading within a Cluster of Microgrids

Cryptocurrencies are a way to securely transact between buildings while preserving privacy, guaranteeing sales ("smart contracts") and keeping a record. However, new coins can create more unique features for better and more structured deployment of energy. For example, tranched cryptocurrencies provide a way to "tier" the energy that is produced, which can determine sale order when the production is volatile. Building A may opt to pay more per kilowatt hour than Building B, which guarantees A's place in line for the energy before Building B(Russo et al., [n. d.]).

We could imagine an extension to MicrogridLearn where an agent helms the demand response within each building and trades outwards. Financial benefit may accrue if other buildings are willing to buy energy reductions (or local generation) for more than the utility's sale price. Specifically, imagine that if the utility's sell and buy prices are \$2/kWh and \$4/kWh, and a building may be able to reduce (or produce) energy for \$1/kWh. It may then sell that energy to another building for $\$x$/kWh, with $2 \leq x \leq 4$, so that building can avoid paying more for the utility's energy. The agent may set a different price for each tranche of power, and transact with other buildings in a complex and nuanced way. The tranches create a natural "waterfall" in which more energy in one tranche overflows into another tranche. Energy sold in the most expensive tranches are most likely to filled to the full demand, and thus, the expensive tranches provide a service close to baseload power. In a way, the agent is selling not only power but also probability of delivering future power.

Please see Figure 5.1 for a depiction of how an energy-sales coin may work. On the top panel are tranches filled by the coin, with senior, middle, and junior tranches by customer priority. On the bottom are pro-rata, equal allocation which nonetheless proportionally allocates energy.

We may even imagine an aggregator coordinating the sale of a large amount of prosumer demand response from many buildings as a single entity in order to submit higher and more influential bids on the generation market.

5.2.4 Further Investigation into Guardrails

5.2.4.1 Ablations of the Planning Model to Help Understand How to Leverage Uncertainty

One of the important contributions of the guardrails project was the following: did the guardrails would *in spite of* or *because of* planning model uncertainty?

Preliminary evidence is mixed. We found that a planning model with fewer training examples was better for learning, which was a surprising finding. It was almost certainly due to the fact, however, that the fewer the datapoints, the more uncertain the model was in predicting, and so the more conservative the decision threshold was. A similar outcome may have been accomplished by decreasing the cost threshold in the planning model with more points: both would have required actions to be more certainly positive.

An ablation on training points and threshold would help us understand the role that the planning model plays in helping an agent avoid unsafe actions. Clearly uncertainty by itself may not hurt an agent's ability to predict safe actions; it seems likely, furthermore, than some

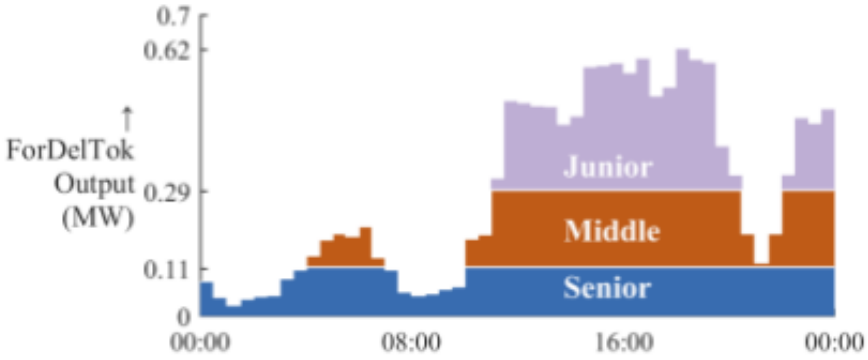


Fig. 1. An example of $\overline{CON}_{i,t}$ values for the tranched scheme.

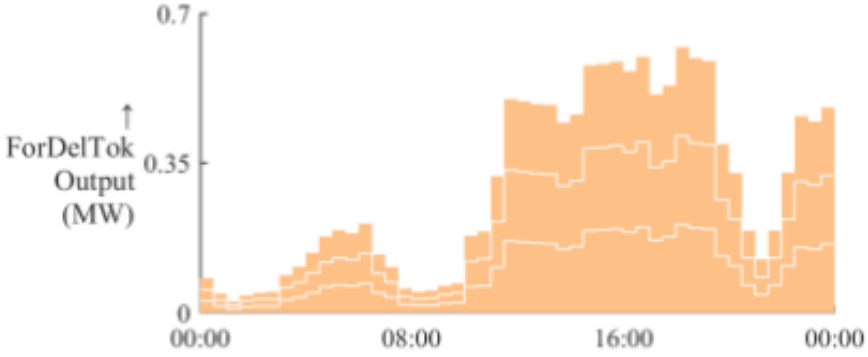


Fig. 2. An example of $\overline{CON}_{i,t}$ values for the pro-rata scheme.

Figure 5.1: Two metrics from a proposed tranched cryptocurrency for energy, from (Russo et al., [n. d.]).

uniform increase in uncertainty the state space is unlikely to hurt the agent. However, at what point does incorrectly assessed uncertainty hurt a controller's functioning? How might a decision rule accommodate uncertainty in uncertainty? These are questions for future work.

5.2.4.2 Dynamic Decision Thresholds

In our study, thresholds were constant as the baseline, TOU pricing, was constant. In another situation, where a baseline may also be moving (i.e. a rules based controller) it would be interesting to understand whether we could still elicit improvements in learning. We hypothesize yes, but perhaps the fixed threshold creates a rule that is easier for the agent to internalize (a piece-wise function in the implicit dynamics space.)

5.2.4.3 Including Offline Datasets in Planning Models

Methodologically, we seem to have stumbled upon an interesting use of offline data, but we do not know how to take into account data distribution shift. Incorporating data distribution shift may unlock the ability to incorporate larger offline datasets into planning model creation which may smoothly handle the difference in data distribution (similar, in fact, to geographic averaging in power systems.)

We propose to explore further how we can leverage larger offline datasets instead of slowing the learning of the RL price controller. Some suggestions of how to do so may involve exploring different distributions to sample actions for the initial offline dataset collection as well as guardrail strategies for planning model ensembles with smaller variance among the component neural networks. We may also use some form of Thompson sampling in decision rules (Bouneffouf et al., 2020).

5.2.4.4 Use of Guardrails in a Longer Trajectory

We study a situation in which the environment's state is a series of 1-step decisions. However, the guardrails method is easily extensible to a situation where a trajectory is important.

The simplest and perhaps most extreme way to extend the Guardrails to multi-step trajectories is to attempt to isolate single actions on the backdrop of greedy action in order to compare apples to apples. That is, assuming a trajectory τ of length T, exploratory action $a_1 \sim \pi_\theta$ would be the next action sampled from the policy and $a_{1,RBC} \sim \pi_{RBC}$ would be the action sampled from the rule-based controller, our baseline. Sample actions $a_2, ...a_T$ from greedily from the same policy (either $\pi_{RBC}(a_i|s_i)$ or $\pi_\theta(a_i|s_i)$ for $i = 2, ..., T-1$) thereafter, in both cases, to ensure that the effect of the first decision is what is being compared. We may call the trajectory $\tau_{\theta,b}$ where the experimental action is sampled from our policy and $\tau_{RBC,a}$ where the experimental action is sampled from our baseline, where $b \in (\theta, RBC)$ to correspond to the policy where the remainder of the trajectory is sampled.

What metrics should be used as the decision rule? An aggregate metric for each trajectory, such as the average return over τ, or the (single) reward at step T, may be appropriate. In following with our method, any aggregate metric would be distributional in prediction if a trajectory in our planning model is sampled. Thus, we may use distributional comparison metrics (KL divergence or Kolmogorov-Smirnov tests) to say whether the experimental action has had a significant effect.

Many implementation questions remain. If τ is long enough, would experimental actions ever have a large effect? If not, then we could imagine sampling a few ($t < T$) actions from each comparison policy and then sampling the rest from the same policy, in order to test a difference in short-term strategy. However, we believe that it is useful to answer such questions if we are to usefully extend the guardrails method to other domains.

5.2.5 Extension on Domain Randomization Driven Meta-Learning

5.2.5.1 Using Adversarially Compounding Complexity by Editing Levels (ACCEL) to Create Structured Auto-Curricula for Intelligent Environment-Side Evolution

Disconnects between simulation and real world may come in the form of variability and idiosyncrasy.

Some of idiosyncrasies may be failures in communication between the environment and the agent, such as the failure to communicate accurate (or any) state information, or the failure of the environment to respond to actions. For instance, a transactional microgrid environment may transmit biased, noisy, or flat readings of the power consumed along the network. The interface that displays prices to prosumers may display zero prices instead of the prices the agent intends to transmit (Sipple, 2020).

Other disconnects may come in forms of variability. For instance, the weather in the real world may be much more varied than the weather in simulation. Holidays may skew energy use. Summer months may bring difference in behavior (vacations, etc.)

Many of the disconnects mentioned above complicate the environment, but we may ideally like the agent to learn the correct environmental dynamics before it learns the complications. Some of the complications may be inline with existing dynamics (i.e., weather variability may correlate with expected changes in energy use.) Other complications may seem to obscure existing dynamics (i.e., if the environment does not receive a transmitted action, it may be unclear to the agent whether this was some fundamental response of the environment or simply a mistake) (Zhao et al., 2020). However, the agent in both cases, the agent may be better suited to learning the correct dynamics first before learning complications.

Curricula may be a great way to help RL agents adapt to real world idiosyncrasies. Curricula are sequences of environments that change in ways to produce valuable output in agents' functioning. Our work on domain randomization in meta-learning is a form of curriculum, albeit it a disordered and inefficient curriculum. We may easily imagine other types of curricula, such as manually incremented curricula, who build difficulty over time so that an agent is first led to local optima in an easier environment and then introduced to a novel dynamic.

However, if one's goal was to adjust an agent to several different types of idiosyncrasies at the same time (i.e. sensor failures, weather variability, and action sensitivity), it may be inefficient for a human to decide to increment the difficulty of all idiosyncrasies at the same time.

To this extent, *auto-curricula* may be important. Auto-curricula are sequences of environments that are determined with reference to the agent itself, rather than externally determined. There exists a growing literature of work (Parker-Holder et al., 2022; Dennis et al., 2020) interested in leveraging auto-curricula to help agents grow, which introduces a very interesting development in RL: an environment-side learning in addition to an agent-side learning.

Regret is an important measure for advancing autocurricula. It is defined as the difference between the reward an optimal policy could achieve and the reward a current policy achieves given a state:

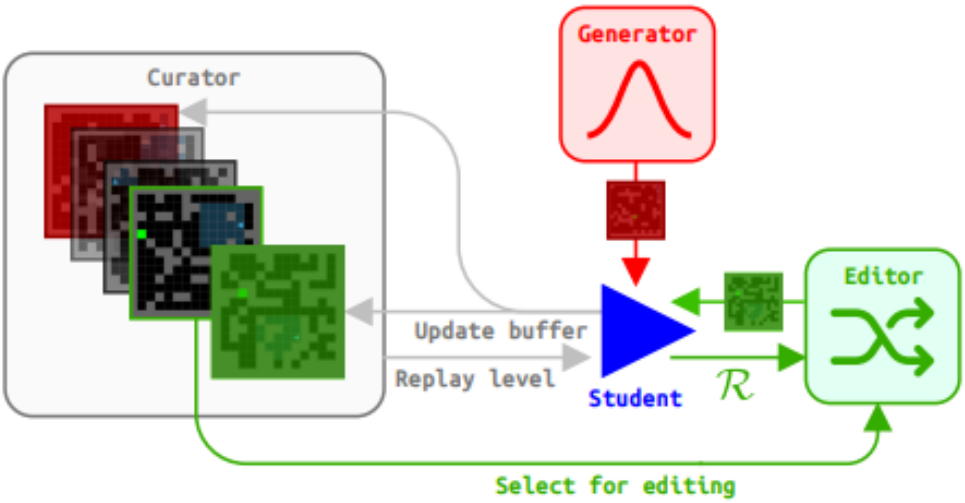


Figure 5.2: A depiction of the environment evolution process proposed by ACCEL. Image copied from (Parker-Holder et al., 2022).

$$\mathbb{U}_t(\pi(\theta), \theta) := \arg\max_{\pi(\theta)^* \in \Pi} V^(\pi^*(\theta)) - V^{\theta}(\pi(\theta)) \tag{5.1}$$

Regret is a great measure to create an autocurriculum because it identifies: (1) sane environments (2) useful environments for the agent at its current state. If an environment did not have a feasible optimum for an agent, the optimal policy would perform poorly, and so regret would be low. Similarly, if an environment were easy for the current agent, the current policy would perform near optimal.

Of course, it is impossible to exactly recover regret (otherwise one would already have the optimal policy.) Two attempts that recover regret are Positive Level Replay (PLR) (Jiang et al., 2021), which estimates regret from the temporal difference error, and PAIRED (Dennis et al., 2020), which optimizes an adversarial agent within an environment prior to evaluating the current agent. Of the two, it is the opinion of the author that PLR is preferable due to ease of implementation.

One auto-curricula strategy, ACCEL (Parker-Holder et al., 2022), uses PLR to estimate the regret in novel environments, and continually leads an agent through sequential environments that are estimated to have high regret. It evolves environment using genetic algorithms. Please see Figure 5.2 for a depiction of their evolution process within their test environment, a maze environment.

We are extremely interesting in applying (Parker-Holder et al., 2022) to Sim-to-Real applications where we could evolve environments to have successively more and more failures. Here, we believe that agents would be taught basic skills of well functioning environments and gradually led to learn more and more idiosyncrasies in order to anticipate well the demands of the real world.

5.2.6 Extensions on Adversarial Attacks to RL

The goal of our work in adversarial attacks to RL is to call attention to the threats made possible by adoption of RL in energy grid pricing. Towards this end, we focused on a narrow yet concrete setting, leaving much room for future work.

- The classical adversarial machine learning literature has an abundance of *targeted* attacks, in which the adversary is able to lead the agent towards a particular policy. One of our next objectives is to design an adversary to cause targeted damage, for example to lead the agent to learn a policy which puts the voltage constraints of the power grid at risk.
- Our proposed defense requires the RL agent to drop as many trajectories as could potentially be compromised. More sophisticated defenses could likely result in less dropped data and more robust learning.
- We would like to explore our attack in more environments. In parallel, we hope to achieve successful attacks with smaller settings of ρ and ϵ.

5.2.6.1 Weaker Attack Models

Our attack model considers a malicious entity that takes control over a small fraction of price controllers and causes them to slightly misreport prices. The attacker leverages the pricing scheme set by the provider to find small perturbations of prices that maximally mislead the agent.

In this work, we granted the attacker full (read-only) access to the neural network used by the RL agent. If the RL agent naively sets pricing schemes in each microgrid controller by uploading the pricing policy in full, then this is indeed a realistic assumption. However, a security-aware provider may instead upload an obfuscated pricing policy that retains the same functionality of the learned policy without exposing the parameters of the neural network. This precaution would prevent our attacker from being able to compute (and reverse) the gradient. An alternative attack could perhaps exploit the transferability of adversarial examples (e.g., (Papernot et al., 2016; Liu et al., 2017a)) to bypass this defense.

5.2.7 Intrinsic Motivation

A future exploration for our intrinsic motivation project may be to plot the visitation densities in order to more directly measure the extent of exploration.

One area of future work that may be interesting is to try competence-based intrinsic reward. This may cause different skills to be be learned for different areas of energy consumption and different prices. This may better capture similarities in the environment for different choices of price setting. Another extension of this project may be using multi-agent reinforcement learning. By deploying multiple controllers, each trying to maximize their reward, we can attempt to make each agent compete with each other to maximize their own reward. This may allow the controllers to set better prices and lead to a more drastic shift in demand.

5.2.8 Future Work in Personal Federated Hypernetworks Applied to RL

5.2.8.1 "Cost of Privacy"

We wish to further investigate the "cost" of privacy in terms of the negative impact it may have on training time and thus on cumulative aggregator profit. In order to create a true apples-to-apples comparison, we would need a mechanism that aggregates information across microgrids in a suitable way. Some ideas on this front include multi-agent RL that shares the critic but personalizes policies, and hierarchical RL with a global aggregator.

5.2.8.2 Vertical Integration of the Hierarchy

In the future, PFH may enable further exploitation of the hierarchical nature of price setting for energy demand response. The energy grid can be imagined as a hierarchical tree (Spangher, 2021), with buildings responding to energy prices set by microgrids, which respond to energy prices set by city utilities, which respond to energy prices set by state utilities, etc. In the future, we may have IoT devices adjusting demand to energy prices set at the building level. At any level of the energy grid, the task is the same: set prices for agents beneath you to elicit a demand response. In this work we have only looked at one level of this energy hierarchy, but the methods we have used could be applied to other layers of the hierarchy as well, and even multiple levels of the hierarchy. One could imagine a hypernetwork that learns from price-setting agents at every level of the hierarchy, and can be used to rapidly initialize agents to manage any new entrants to the energy grid, all while preserving privacy at different levels of the tree.

5.2.9 Future Work in Extrinsic Pretraining

Our exploration in heuristic based extrinsic pretraining has many opportunities for future work. One may bound the areas of the state space that it can explore, vary the structure of the planning environment, or cycle through several different extrinsic rewards according to their needs.

5.3 Limitations

5.3.1 Limitations on OfficeLearn Environment Setup

Other Common Resources Considered It should be noted that our results perhaps under-tell the story of office demand response: we lack a structured way to measure behavior towards air conditioning, lighting, and ventilation, including a comfort model to capture the interplay between price and perception of comfort. While our simulation might ambiguously include some of these demands as generic office worker energy demands, it does not do so explicitly. Indeed, according to EIA estimates Agency ([n. d.]), lighting accounts for 17% of energy use, ventilation 16%, and cooling 15%, whereas computers and office equipment, the categories that are best captured by our analysis, only account for a combined 14%. We assume therefore that the cost estimates we provide are a lower bound on the total cost savings that a pricing scheme within an office building.

Lack of Human Data We are limited by the inability to conduct an experiment on real people in an office building due to the pandemic interrupting in-person office work, and thus have resorted to reasonable behavioural simulations.

5.3.2 Limitations on Guardrails

5.3.2.1 Planning Model Data Needs

Our work assumes the ability to gather a cheap offline dataset of human responses to social game prices, perhaps by conducting the experiment on a single floor of the building to start with or a stratified sample of office workers.

5.3.2.2 Accurate Planning Model

Our neural net ensemble is an attempt to approximate the Oracle with a practical planning model. However, while it produces a less biased distribution than a single network, it likely still has some systemic bias in its predictions. The low frequency of the guardrail "trigger" in suggests that poorly trained ensembles tend to underestimate costs, which is followed by worse performance.

Therefore, we suggest that our results are taken in the following: given a limited amount of engineering time, one may balance the engineering required to increase the precision of their planning model with the engineering required to increase the normalcy of errors of their planning model. Gains in either direction will help the RL agent perform better and reduce demand response deployment costs.

5.3.3 Limitations on Adversarial Work

5.3.3.1 Practical Implementation and Subsequent Poisoning of a Pricing Aggregator

On both the positive and negative end, our work may suffer from high so-called "deployment costs"; i.e. the length of time that it may take a prosumer to train. Although within the context of RL our agent and attack actually happen quite quickly, a real-life investor may nevertheless balk at the notion of an RL algorithm leading microgrids on a twisting journey of price exploration for close to a year. Additionally, a real-life microgrid may be much more non-deterministic than even a complex simulation, making it difficult to forecast the strength of an adversary.

5.3.4 Limitations on Personal Federated Hypernetworks

What are potential negative societal effects of our work? Overall, negative effects to prosumers are limited, as the focus of our work is in protecting consumer information. Furthermore, prior work demonstrated that the presence of an aggregator consistently reduced energy costs for consumers.

However, a persistent danger of AI is that it is often through centralized profit-seekers that it is deployed. Our work is no different in this regard. Although our specific innovation protects prosumers, it may improve the economic viability of a profit-seeking entity whose scale may eventually enable it to further its own profit at the expense of prosumers.

Also, the act of setting prices in systems may raise fairness concerns. If initial training microgrids are biased towards wealthier residents, the PFH may initialize new policies with pricing that benefits consumption habits of wealthier clients but penalizes those of poorer clients. A vivid illustration may be seen in the types of prosumers who are best poised to benefit from economic aggregation: prosumers with large solar panels and batteries are able to shield themselves from or profit off of high prices by consuming their own energy, and may fully charge their batteries when prices are low. Prosumers with smaller or no storage capabilities do not have this luxury, and thus are more vulnerable to the negative effects of price fluctuation.

Technically, our work is limited in several ways. We present a "goldilocks" zone in which PFH outperforms other methods, but as our simulation reduces the complexity of the world by a significant degree, it is unclear how or where this goldilocks zone would appear in the real world. Second, we take care to make sure that information passed within our system is privacy-preserving, but any privacy is only as safe as the firewalls on the rest of the system.

Appendix A

Endnotes

A.1 Physical Basis for Climate Change

I wish to explain thoroughly the physical basis for climate change, as it is important for everyone to understand. Thus, I will take some time to explain why the burning of fossil fuels relates to climate change. To understand the former, we must understand a bit about fossil fuel formation.

Fossil fuels is an umbrella term encapsulating different fuels such as coal, oil, natural gas, and shale that are formed over millenia and are generally made up of molecules called "hydrocarbons". Generally, organic matter which, as it decomposes, is gradually transported deep under the earth's surface, is chemically transformed by geologic pressure and heat. Oxygen was almost completely absent in these conditions, and so a specific reaction occurred that turned the organic matter into a waxy substance called kerogen. As the geologic process continued, kerogen gradually underwent a process called *catagenesis*, and transformed into masses of substances with long hydrocarbon chains[XXXII].

The starting organic matter, geologic composition, and various amounts of heat and pressure all create different fuels. Oil and other petrofuels tend to result from the remains of algae that over many years sank to the bottom of shallow seas and were buried under layer after layer of sediment. Coal, meanwhile, tends to result from massive peat forests and trees which, over time, had some probability of cycling down in the massive ground carbon cycles(pet, [n. d.]). Natural gas generally arises from oil deposits which occur under some hard geologic layer, trapping the gas and allowing it to gather. Please see Figure A.1 for an elementary recap of the above.

Fossil fuels are all extremely useful for one reason: *combustion*. The long hydrocarbon chains are composed of a backbone of carbon to carbon atoms surrounded symmetrically by hydrogen atoms. Although stable at room temperature, these molecules are relatively unstable above certain thresholds and burn easily due to the long carbon chains. Carbon to carbon bonds surrounded by hydrogen molecules are relatively weak chemical bonds.. Perhaps the most consequential building blocks of modern society have been a series of inventions that harness the combustion of fossil fuels in increasingly sophisticated ways to harness their chemical energy, creating electricity, powering engines, and manufacturing advanced petrochemicals. When hydrocarbons are burned, their long hydrogen chains are split and the

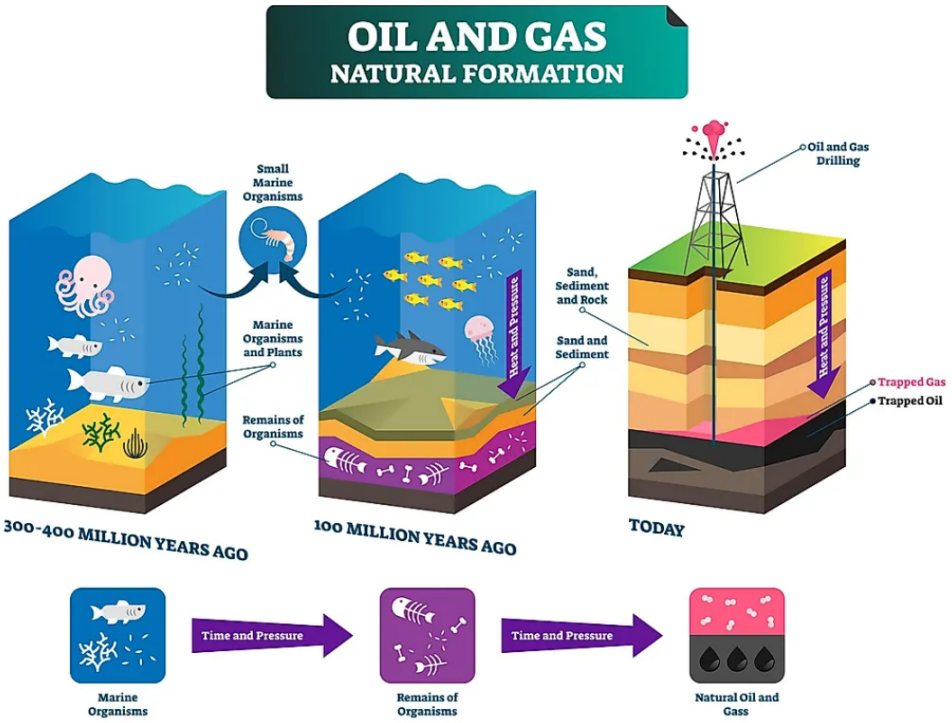


Figure A.1: An elementary diagram demonstrating fossil fuel formation (fos, [n. d.]b).

resulting cloud of carbon and hydrogen combines with atmospheric oxygen to form carbon dioxide[XXXIII] Please see Figure A.2 for a visualization of some common fossil fuel structures.

Combustion of fossil fuel fuels would not be an issue climatically were it not for a phenomenon known the *warming potential* of carbon dioxide and other greenhouse gases (GHGs)[XXXIV]. Light that enters the Earth's atmosphere enters as high energy photons in the visible spectrum. The atmosphere scatters the energy of the light, but much makes it to the ground which absorbs some of the energy. Some remaining energy, however, is reflected back towards space as *blackbody radiation*, i.e. lower energy photons below the visible spectrum. Carbon dioxide and other atmospheric molecules with high warming potentials re-reflect some of this radiation before it leaves the Earth, essentially harvesting more energy from the photons that originally entered. Earth's atmosphere and a stable supply of greenhouse gases are an incredibly intricate and finely tuned instrument; the process is essential to maintaining a temperature that can support life as we know it. Please see Figures A.3 and A.4 for demonstration of the blackbody effect and radiative forcing of common GHGs.

Regardless of the method of fossil fuel formation or fossil fuel combustion, the end result is on an abstract level the same: many atmospheric elements, including carbon dioxide, were at one point fixed into solid organic matter, and geologic processes buried them deep into the earth's core, preventing them from being easily reintergrated with the atmosphere. The process of geologic carbon fixing slowly decreased the warming potential of the atmosphere. However, as a result of burning the fuels, humans are causing a rapid reintegration with the atmosphere, changing the composition of the atmosphere and the warming potentials of the atmosphere at very fast rates relative to geologic timescales. This process is the process

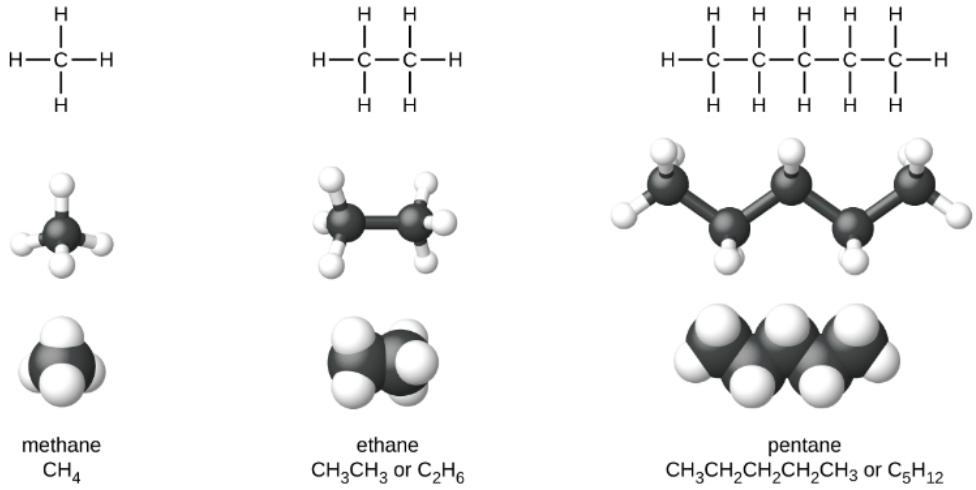


Figure A.2: Chemical structure of some common fossil fuels (fos, [n. d.]a).

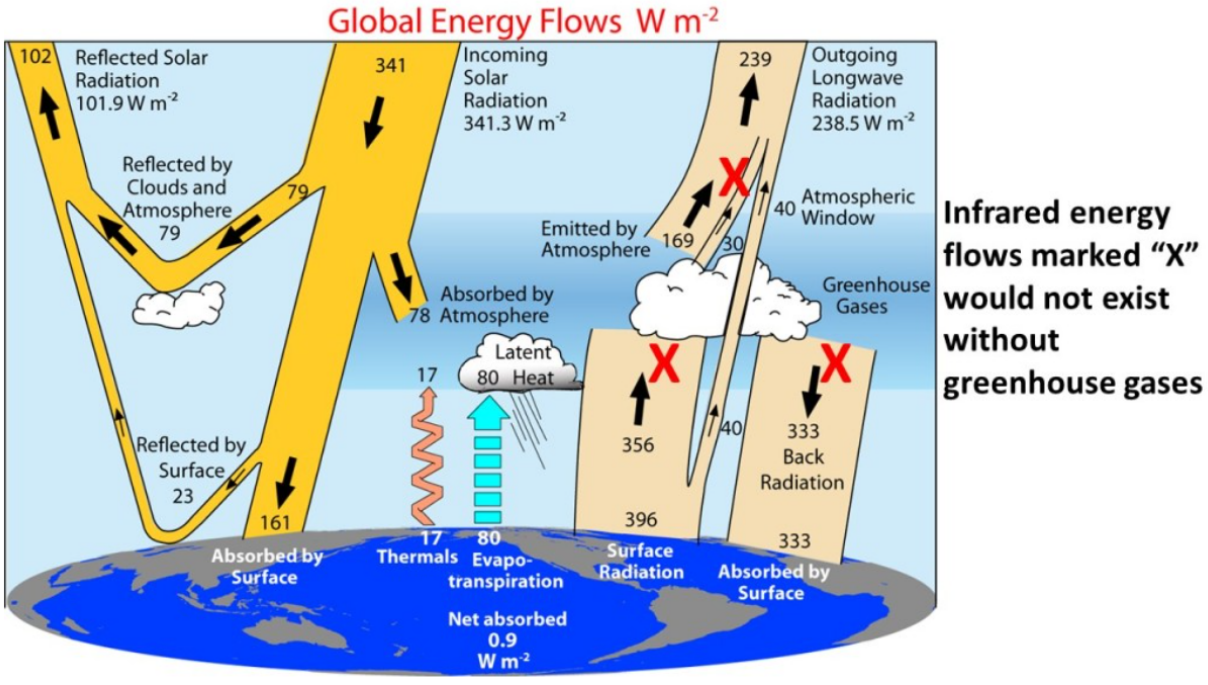


Figure A.3: Visible light and blackbody radiation through the Earth's atmosphere (bla, [n. d.]).

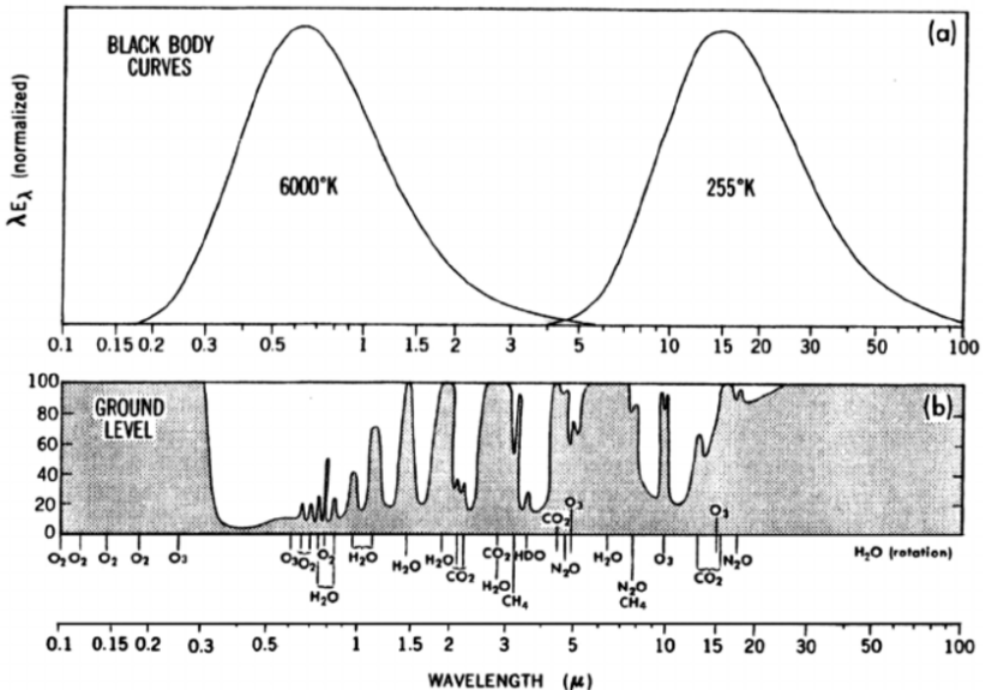


Figure A.4: Radiative forcing of common GHGs (Brunetti and Prodi, 2015).

known as "climate change" and could have profound impacts on nearly every system on the planet.

The earth has myriad positive feedback loops which, once triggered, may result in runaway climate change (Podesta and Ogden, 2008). The author looked into methods to address climate change once thought extreme, such as geoengineering (Keith et al., 2010) or carbon capture(Van Arsdale et al., 2022), but believes that because feedback loops must be accounted for in carbon emission, believes that the entire conversation on carbon capture is slightly impoverished. That said, positive social tipping points in climate change awareness are also important to consider (Moser and Dilling, 2007).

A.2 Climate Modeling

Climate change modeling is a labor and compute intensive task. Lewis Fry Richardson, in 1922, first published a book on computational climate modeling, imagining that if he had a "stadium full of 64,000 [human] computers", he could compute the weather by solving differential equations for gridded units dividing the earth's surface(Cli, [n. d.]). Indeed, some of the earliest computers, such as ENIAC at University at Pennsylvania, were deployed to run early climate models. Later, the first supercomputers were developed to run modern, large scale models known as Global Climate Models (GCMs).

Amazingly, even Richardson's speculation on climate computation was not the first to imagine and endeavor to compute potential climate change. Svante Arrhenius, the famous Swedish chemist behind the Arrhenius equation that we learn in high school, laboriously

completed numerical calculations necessary to estimate that decreasing carbon dioxide by half in the Earth's atmosphere would decrease global temperature by an average of four to five degrees Celsius, which is correct within some margin of error (although a calculation in the opposite direction, climate cooling, as the ones done today.) Guy Callendar was the first to rigorously estimate the warming potential, using a 1D radiative transfer model to show that rising CO2 levels are warming the atmosphere. His result is published in the Quarterly Journal of the Royal Meteorological Society in a paper entitled, "The artificial production of carbon dioxide and its influence on temperature".

Callendar's outcome shows the sheer length of time that the phenomenon was known to scientists, a fact that the oil industry has often labored to hide. These models tell us beyond all reasonable doubt that the atmospheric intuition explained in the previous section is indeed affecting the world; they also give predictions on what effects are happening where at given points in time.

Notes

[I] I will describe briefly the aspects that personally motivate me the most about climate change.

First and foremost, I am concerned with the forcible displacement of poorer people as agrarian climate dynamics change and farming becomes untenable. The UN has estimated that around 1.2 billion people to be official "climate refugees" by 2050(UNE, [n. d.]). The term "climate refugee" comes from 1985, when UN Environment Programme (UNEP) expert Essam El-Hinnawi defined climate – or environmental – refugees as people who have been "forced to leave their traditional habitat, temporarily or permanently, because of marked environmental disruption." The depths and widespread nature of the suffering are difficult to fathom, and the extent to which these masses of vulnerable people can be exploited by bad actors (through cheap labor, expensive loans, and other forms of human trafficking) is painful for me to imagine, especially when these people have contributed almost nothing to the cause. For these reasons, it feels clear to me that climate change is clearly a form of class warfare.

Prefacing the following with the disclaimer that I am not a trained ethicist: while the prospect and ramifications of ecosystem collapse worries me and of course I find the large-scale suffering and untimely deaths of many animals to be deeply tragic (not to mention intricate systems of animal-to-animal dependence which have developed over millennia - e.g., one report found that specific snake populations on a remote island were devastated by decreasing numbers of migratory birds who originate half a world away), their deaths feel like outcomes that are more similar to what their deaths would have been in a counterfactual world without climate change.

Instead, with displaced people, the effects of seems to moreso breach a social contract that they have already contributed to with their labor. It is essential to note that most climate refugees stay as internal refugees within their native countries, moving to cities and urban slums in the hopes of picking up industrial work. However, some do travel to other countries in order to find work and to be safe from climate-exacerbated conflict.

Thus, in response, political impacts arise related to the large-scale movement of poor people. A PNAS paper (Kelley et al., 2015), although criticized, recently linked climate change to the Syrian war, which caused mass displacement of Syrians, the speed of which triggered protectionist political movements across Europe. While nothing but a dilettante in political geography, I claim the Syrian mass migration event *fueled a political movement* that may have languished otherwise: it is my guess that slower movements of people would have created less political impact in societies. Indeed, more generally, we know generally that perception of swiftly changing contexts leads voters to lean conservative.

Drawing the argument back to my personal relationship with climate impacts: I ultimately want to live in a society that is culturally welcoming, politically liberal, and actively trying to improve its concept of safety nets for all. Thus, I fear that the rapid and numerous changes of climate change will contribute to a political climate that I do not wish. This, among with

myriad other reasons, motivates me to work in the area.

[II]Sometime from the years 2005-2020 it fell somewhat out of vogue to, when writing to communities such as energy efficiency engineers, re-introduce climate change or allude to it by describing one's work as being in "sustainability" or "environmentally friendly". I believe there may be multiple reasons for this. First, by the 2000's, generations of scientists had based lives and careers on the issue, and so to be overly explicit about sustainability motivating was to telegraph one's newly having entered the field. The mindset of needing to persuade one's audience on the necessity of sustainability was that of one who primarily communicated with audiences outside climate and energy a group that was growing increasingly large and professionally diverse. Second, the "energy transition" was starting to find legs in motivations other than altruistic concern for the environment. To focus work on those motivations was to look forward into what the energy transition could become, not to look backwards into the motivations that first powered it. Thus I will try not to inundate the reader with facts about climate change in my introduction.

[III]It is worthwhile to pause and note the philosophical goals of Energy Transition, which are to maintain the fossil-fuel-derived *promises of modernity* without incurring negative environmental externalities. promises of modernity can and should be defined is certainly up for debate. For the purposes of this dissertation, I argue that these promises can be encapsulated in three freedoms:

1. the **freedom of large-scale movement** arises from the use of petrofuels as engines of travel which enables us to transport ourselves fast and far. Freedom of movement dissolves the centuries-old practice of living where our familial roots are, and enables some of us to choose where we live depending on what location we feel is most in line with our values and personal aesthetics. One aspect of modernity is the fluid choosing of identifies based on choice of common location.
2. The **freedom of time** results from the harnessing of petro-derived electricity to perform labor that would have taken considerable human effort and time in the past. Freedom of time creates time for economic specialization or leisure activities. A diverse economy allows for an interesting multitude of job types and for the progression of academic knowledge generation.
3. The **freedom of material accumulation** may result from an increasingly diverse and petro-powered economy in which production per-capita is higher than at any point in human history. This material availability positively impacts the quality, comfort, and longevity of life that many are able to enjoy.

Energy Transition is an event that can and should happen, but it is the pragmatic belief of the author that it must occur in a way that retains most or all of the freedoms that comprise the promises of modernity.

There exists a field called "Energy Humanities" that seeks to train the critical lens developed in critical and poststructural theory on the Energy Transition. Of the previous,

Casey Williams (Williams, [n. d.]) takes a more assertive tone on how truly fossil-based the promises of modernity are, saying:

> It might be worth pointing out that the promises [this dissertation] identifies — freedom of movement, time, accumulation — really only become recognizable as promises as fossil-fueled industrialization begins to take off in Europe. In other words, the promises of modernity are bound up with the sorts of economic activity made possible by fossil fuels. This means that energy transition is really about replicating the social and cultural features of a fossil fueled world without fossil fuels. When we understand this, we can see that energy transition is more complicated than simply switching energy inputs; it's a process of reinforcing (or challenging, as the case may be) ways of life, and expectations about how life should be, that grew up alongside our expanding use of fossil fuels.

Will there be ways that the output of the green transition, i.e. a system composed of competing manufacturers of new materials? The answer is almost certainly "yes": batteries, wind turbines, solar panels, etc., all require some degree of material accumulation which may come from mining fraught with geopolitical implications. It is the opinion of the author that many fields will develop alternatives to material bottlenecks. Many of the most problematic minerals in green energy *support* as opposed to *compose* the energy generation, and so the market is more suited to help innovate around bottlenecks. Green energy lends itself to democratization of generation as well: rooftop and community based generators are possible to harness common fuels of sunlight and wind in a much more decentralized way than centrally mined fossil fuels.

In order to maintain the freedom on time that electricity provides, Energy Transition must create an alternative system of electricity generation that meets the dynamic demands of people.

[IV]While maintaining freedoms may seem like a constraint on the way Energy Transition occurs, it in some ways is liberating itself: it allows us to unleash creativity on ways we may retain freedoms under radically different forms. For instance, although not the focus of this dissertation, we may first constrain ourselves to replicate the freedom of vehicle transportation by creating an electric vehicle production economy; however, when realizing that vehicle ownership is only instrumental to the freedom of transporting oneself and one's materials quickly and efficiently, we may then liberate our strategy to create mass public transit systems that provide the same freedoms.

[V]I've grappled with the ethics of material accumulation for a long time: whether it was moral, i.e. whether it was ethically permissible given the knowledge that a study of Marxism gives one: that the production of material incurs some exploitation from the works by the owners of the means of production, that the system inherently drives greater production and greater consumption, and of the negative effects it may have on lower class labor and the planet. All of these are at odds with the fact that the Energy Transition currently both seeks to maintaining similar levels of material accumulation and is accomplished through material

production. The Energy Transition gives us the chance to imagine what alternate system might exist to help us achieve similar promises of modernity. However, it is the opinion of the authors that, practically, the speed at which the Energy Transition must occur means we must reconcile some of the ethical issues we have with the current system.

I once had the pleasure of asking a variant of the question to noted Marxist scholar Angela Davis: did she feel complicit with the capitalist system in taking a job in the UC system, which hegemonically reproduces capitalism around the world, and accruing material items? She said that she tried for decades to escape the "system", only to discover just how far the system actually reaches (i.e., that it is impossible to escape.) She furthermore acknowledged her own appreciation of the Apple computer she was speaking from, the car she drove, and other material items she accumulated. She reconciles her material appreciation with the knowledge of capitalist disparity by practicing constant detachment from her materials. She enjoys her materials while mentally ensuring that if she were called to give them up during a revolution, she would happily do so. Ultimately, the goal of any social utopia is not to decrease the material quality of life of any class of people, but raise the floor so that all may partake in the various freedoms.

One reaction that I got to this question was the following: "[I] should have asked Dr. David whether she is comfortable advocating for a political system that would not have provided her the same intellectual freedom that the current does to criticize it." Although an understandable critique, this may be responded to simply: one does not have to endorse all of the negatives of the U.S. government in order to advocate for democracy as a form of government. Indeed; one does not even need to advocate for certain ways in which U.S. democracy is implemented: through a two tiered republic, with two parties, with gerrymandering, etc., etc., in order to advocate for the ideal of providing *some way* for members to vote. In Dr. Davis' case, the ideals of non-materialism, worker equality, property abolition, collective ownership, etc., etc., may be advocated for without endorsing specific implementations of Marxism.

Technologists abstract away technological implementation concerns frequently. For instance, there are many instances in which A.I. may be used as a tool of power, to discriminate, and to increase the reach of authoritarian governments. It may also be used to increase the power of corporations unbeholden to the people. However, one may advocate for and believe in the principles of machine intelligence without endorsing every application.

[VI]The author would like to note that rare earth shortages may be circumstantial to our time period. Given the vast capital incentive, asteroid or moon mining may be a reality in the future, which may help the green transition perpetuate itself (for more information, please see an excellent survey of near-earth asteroids (Morales-Calderón et al., 2011) and main-belt asteroids (Marchis et al., 2006b,a), and a review of asteroid mining (Zacny et al., 2013). Of course, enhanced recycling campaigns and technologies are promising as well.

[VII]Specifically, the author has sat through numerous judging committees or paper reviews for battery development, and has seen firsthand the breadth of proposals to diversity battery chemistries. Although few are immediately available, it seems increasingly likely that emerging

battery companies will lean on anode and cathode alternatives as rare earth mineral mining becomes more competitive.

[VIII]Newer and different forms of hydropower may include small tidal energy generators that harvest energy from ocean currents. These have largely failed to produce a power density that justifies their upfront cost, and so we do not believe that they are an extremely promising long-term source of energy. They are, for the most part, non-dispatchable.

[IX]Here we pause briefly to explain what some call the "tyranny of LCOE". LCOE is a metric intended to contextualize the larger cost of an energy source by combining the cost of construction and operation (i.e., fuel and maintenance), into a single number. It "represents the average revenue per unit of electricity generated that would be required to recover the costs of building and operating a generating plant during an assumed financial life and duty cycle". It has been used almost to a fault to rate developing technologies to determine which should get grants and which shouldn't, and may be the cause of the untimely demise of many different technological developments in energy. Notoriously difficult to estimate, it makes certain assumptions about construction cost, siting, lifetime, and maintenance needs, and often ignores metrics such as environmental externalities, volatility of generation, or pace of technological development that may reduce maintenance costs.

[X]As a brief note for those who, unfamiliar with the physics of energy, may be forgiven for noting a similar between the *chemical energy* harvested from hydrocarbon combustion and the *nuclear energy* harvested from nuclear fission. The release of chemical energy results from the breaking up of bonds *between* the atoms of a single molecule. The release of nuclear energy, meanwhile, comes from breaking the *much* stronger bonds that bind subatomic particles (i.e. neutrons and protons) *within* the nucleus of a single (albeit much larger) atom.

[XI]Uranium is the classic heavy material used in fission reactors, with uranium-235 being the primary fissile atom present at rates of 3-5% in fissile uranium cakes which may consist of a range of uranium isotopes[39] Bombarding the uranium cakes with a beam of particles causes some atoms to gain protons (some uranium-235 turns into uranium-236) which make them unstable, and they then break apart, releasing energy which causes other atoms to undergo similar reaction. The resulting material is rich in plutonium and other radioactive elements and must be stored in a stable location for thousands of years. Nuclear products and techniques have many applications; please see Frame et al. (2022) for more details.

[XII]Newer generation fission reactors are able to process the waste plutonium and other trace outputs into other radioactive products that have even shorter half-lives, lessening the political and societal burden that nuclear fission poses. Newer reactors are also much smaller, allowing for modular construction.

[XIII]It is the strong opinion of the author that the safe handling of nuclear waste is a much easier societal problem to solve than climate change and local environmental toxicity caused

[39]An *isotope* is defined as a version of the same element that has a differing number of neutrons in their nucleus as another isotope but the same number of protons, thus different isotopes of the same element have the same nuclear charge.

by the burning of fossil fuels, and that risk of nuclear meltdowns is heavily overstated.

Exactly three nuclear reactor meltdowns have occurred (Chernobyl, Three-Mile Island, and Fukishima), and the number of people killed in any of them is under debate and may be as low as 2000-4000. Reactor safety has increased considerably since the 1960-1970's, when Chernobyl and TMI happened, so the author does not believe events like these to be future concerns. Fukishima was an extremely low probability event (an earthquake triggered a tsunami, and both hit the reactor relatively hard at the same time) and it would be fairly straightforward to site nuclear plants that reduce the chance of a Fukishima-like event to occur even further.

Plus, given recent innovations till the present cut the amount of time necessary to store spent fuel by orders of magnitude, we may expect future innovations to reduce storage time necessary by even more. It is unclear the extent to which we can reduce radioactive decay in fission waste products.

Of course, some of the risk estimation of nuclear fission is related to and bound up in lingering social trauma from nuclear bombs and a Cold War that threatened global safety. While that is fair, it is the strong belief of the author that the public is incorrect in its estimation of risk surrounding new nuclear.

[XIV]As an aside, the term "nuclear energy" could be an umbrella term encompassing *nuclear fusion* as well, if that were commercializable. Nuclear fusion is an opposite process to nuclear fission, and its goals are to combine lighter elements into heavier elements. The most common fusion reaction studied for power is the reaction of smashing two hydrogen isotopes (tritium, H3) together with enough force to break bonds, create a plasma, and form Helium. Doing so requires immense force (currently magnetic or inertial) because the plasma, once formed, creates extreme outward pressure.

Nuclear fusion is, notably, the reaction that powers the sun, and the sun is able to overcome the plasma pressure and sustain a fusion reaction by the immense gravity that its mass exerts inwards on itself. Lacking this type of mass on Earth, teams are currently trying to create similar conditions with complex manipulation of magnetic or inertial forces. We have already succeeded with creating an sustaining a plasma under laboratory conditions (Langmore et al., 2019; Granstedt et al., 2018).

However, we have yet not been able to harvest more energy from the reaction than we put in, which means that the technology stays in a bin of "theoretical possibility" rather than commercializable reality. The US Department of Energy's Advanced Research Projects Agency (ARPA-E) recently put out a Request for Information on alternative fusion methods, which may comprise of cold or room temperature fusion (Berlinguette et al., 2019).

Due to aforementioned perceptions around the safety of Nuclear Fission, however, I personally doubt that the industry will use "nuclear energy" to refer to both, even if it is semantically accurate.

[XV]Humans have harvested wind energy as early as they have risen sails against windy

waters, which may have happened as early as the 9th century BCE. Evidence of wind powered grain turbines in Iran, Afghanistan, and Pakistan as early as the 9th century CE and the first use for electricity generation is documented by Professor James Blyth in Glasgow, Scotland, in 1887. At the time of writing, the largest wind farm in the world is the Gansu farm in China, rated at $\tilde{8}$ gigawatt (GW) with thousands of turbines.

[XVI]A relatively large aside: Denmark and Holland make an interesting comparative case study in how wind technology development should be approached. The two countries happened to choose opposite directions in starting with small turbines and developing larger or vice versa, with different outcomes in success. Through a series of governmental grants, directives, and corporate decisions, wind energy generally started small in Denmark, and research efforts focused on solidifying understanding of the production process of small-scale turbines, and then scaling up size. Holland happened to choose the opposite route, opting for grants to fund large scale megaproject turbines. Unfortunately, all technological endeavors involve failures and associated learning, and when Holland's wind megaprojects failed, they were megafailures. When Denmark's small-scale turbines failed, barely anyone noticed – and scale of deployment increased as people bought them for their backyards. As a result, public perception in Holland soured towards wind as being costly, overpromised, and underdelivered, and public perception in Denmark generally improved as people formed community and cultural attachments to the local turbines. As a result, Denmark hosts some of the largest wind companies and innovators in the world (Vestas, Orsted, formerly Dong energy, etc.) and Holland has none.

[XVII]Arguably, wind power is also derived from the sun, as the sun differentially heats different parts of the earth and atmosphere to create wind – were there no sun, the fluid surrounding the earth would likely be relatively solid. (It may be noted that, while there is some debate about whether a "rogue planet" - i.e. a planet that does not orbit a star - could trap heat, and the answer seems to be "almost certainly not over astronomical periods of time.") By this train of logic, fossil fuels were too derived from the sun. The author does not think that entertaining this line of thought is particularly useful (or even interesting), and so we carefully contain our definitions to "direct sources of energy."

[XVIII]Flow batteries are an exciting technology composed of large tanks that cycle hundreds of gallons of electrolyte to relatively small reaction sites (i.e. anode and cathode.) They make a common economic pain point for batteries, the mineral-rich anode and cathode, small parts of the overall machine and thus save on cost per kilowatt hour (kWh). They can also last many of cycles as there are less physical-on-physical contacts. However, it is difficult to achieve necessary efficiencies or discharge rates.

[XIX]Solid-state batteries, named in contrast to the fluid electrolyte of Lithium-ion, were proposed as early as the 1930's when Faraday demonstrated a glass-based battery functionality. John Goodenough, the co-inventor of Li-ion batteries with a hilarious last name, unveiled a solid-state battery with a glass electrolyte and an alkali-metal anode consisting of lithium, sodium or potassium in 2017. Solid state batteries are lighter, more energy dense, feature faster recharge times, and are safer than lithium-ions, which makes them great for vehicular applications. However, solid state batteries are plagued by cost concerns (much of the

manufacturing process requires vacuums) and chemical concerns such as dendritic formation in the anode piercing the separator and reaching the cathode.

[XX]Large-scale implementation of pumped hydro is underway in Norway and Switzerland, both mountainous regions with governments committed to climate innovation. Switzerland just unveiled a massive 900MW pumped hydro facility called "Nant de Drance", which took 14 years to build and increased Switzerland's entire storage capacity by 33%.

[XXI]I attended the RL and Decision Making (RLDM) 2022 conference, and Sutton actually happened to give a poster in the poster presentation right next to me. During that conference, I met an undergrad who stayed in a shared dorm and was surprised to be assigned Rich Sutton as his roommate for the week.

[XXII]Witnessing the success of the SB research proposal has suggested to the author that RL can be an incredibly good sales tool: multiple competing stakeholder objections can be lumped together in one reward function.

[XXIII]The weighting determined by the reward function was manual, and was tuned after carefully examining the output behavior of the agent in simulation. There are ways to incorporate multiple objectives into control, generally through multi-objective pareto fronts (Ngatchou et al., 2005). Here, iso-reward surfaces provide numerous points where the objectives are computed, and in many implementations, hand the decision off to a "decision maker" to determine which solution is appropriate.

[XXIV]Assuming unvarying TOU prices as a pricing innovation is appropriate for the real world. In general, pricing electricity differently throughout the day is a relatively new practice, and although more utilities price differently with commercial buildings, very few do it with residential consumers. Pacific Gas and Electric (PG&E), for example, widely deploys time-varying prices for both commercial and residential in the form of different TOU schemes (PG&E, 2021). As another example, Illinois defines customers based on the size of their demand, with 2MW as a threshold between "non-competitive retail" (residential and commercial below a size) and "competitive retail". Competitive retail may be subject to TOU pricing but non-competitive is entirely fixed rate (Commission, 2021).

[XXV]Let us provide an example of an office worker who behaves according to Curtail and Shift Behavior. Imagine that an office worker may consume a total of 1000 Wh (280 kJ) throughout the day, 300 Wh (83 kJ) of which are for printing documents that could be curtailed, 300 Wh (83 kJ) of which are for presenting at a meeting whose time can be shifted, and 400 Wh (110 kJ) of which are the minimum required to run a PC. Upon receiving a price signal with high prices from 11am-2pm, this simulated worker would curtail their printing away from the 3 hours with the highest energy price, and schedule their meeting at the hour within $[h - T_{\text{shift}}, h + T_{\text{shift}}]$ with the lowest energy price.

[XXVI]As an aside, the author would like to note a curious tendency of the field to anthropomorphize various technical terms. To direct this example directly, the author notes that instead of calling "curiousity" something more native to statistics, such as "error-seeking" we call it a term that invokes a human brain process. Indeed, even the old terms of "uncertainty",

which may have overlap with more technical terms such as "error", and "neural networks", which replaced "perceptron", seem to be convenient anthropomorphizations. "Intrinsic" and "extrinsic" motivation follow this model too. If neural networks had retained the title of "perceptron" or a worse technical name such as "arithmetic-node-based-layer functions", would they have attracted the legion on researchers that, successively refining and advancing the technique, forged it into the clear frontrunner amongst universal function approximators to be used across Machine Learning? There existed other candidates for investigation at the time of takeoff, including Support Vector Machines (SVMs), Deep Random Forests, and Gaussian processes that may have been kernelized and refined for widespread use. We may never know the answer to this question, and such questions are perhaps better addressed by a Kuhnian critique of the social construction of science on the whole.

[xxvii]The reader may note from our problem setup that the workers only respond to prices given to them that day; thus, a bandit architecture may be appropriate. We tried a Upper Confidence Bound (UCB) bandit, but found that it was not as performant as SAC. Furthermore, because the problem can be generalized to one in which multi-day trajectories are considered, with weekly energy demands, we continue our exploration with SAC as it is flexible to these generalizations.

[xxviii]We tried alternative reward functions: we used simply $-\log\left[\sum_i p^{\intercal}_{\text{util},t} d_i\right]$ as the reward, but we found that when the agent was incentivized only to reduce energy cost, it tended to converge to uniformly high prices for all hours, which was not a very useful control scheme. We also tried optimally solving energy demand distribution at each step after total energy use d_i was observed, i.e. the output of the equation $d_i^* = \min_e(p^d_{\text{util},t\,i})$ with inequality bounds around some d_{min} and d_{max} and treating the reward as the distance $(d_i - d_i^*)$, but the extra computation required made the environment quite bulky, and the optimization output was not particularly interesting or realistic. We thus chose the reward we present in the main text for its interpretability, computational speed, and agent-behavior shaping.

[xxix]

A.2.1 Other Guardrail Rules Considered

Here we list some other guardrail rules we considered that were in general less effective than the quantile approach. We define $b(p_t) := \mathbb{E}^{\text{NN}}[c(p_t)] - c^{\text{tou}}$

- Hard: $\theta(p_t) = \mathbb{1}[\mathbb{E}^{\text{NN}}[c(p_t)] < c^{\text{tou}}]$. Any price with a mean predicted cost smaller than c^{tou} is sent to the real world with probability one, while prices with a higher cost are sent to the planning model.
- Convex: $\theta(p_t) = \min(1, e^{-b(p_t)})$. We wish to test a threshold shape that simulates a quick decrease from the initial bound, and then a gradual asymptote to 0. A convex shape is appropriate for testing this, wherein the decrease after the boundary is quick and the decrease thereafter becomes less and less.

- S-curve: $\theta(p_t) = \min(1, 1 - \frac{1}{1+e^{-b(p_t)/2}})$. We desire to test a threshold shape that simulates slow initial growth and then an exponential leveling off in probability. The S-curve is a common model for such an endeavor. It is common in fields such as epidemiology and technology spread, and helps us test a more "real-world" model in our suite of thresholds.
- Concave: $\theta(p_t) = \min(1, 1 - b(p_t)^2/100)$ if $b(p_t) > 0$ else 1. Here, we seek to model a threshold which initially rises gradually and then increasingly approaches one with greater speed.

Compared to the Hard guardrail, Convex, S-Curve, and Concave, all experiments with different shapes of "soft" thresholds that allow for some probability of a risky price being sent to the real world. This is under the hypothesis that some real-world data on the consequences of risky behavior is necessary for the controller to learn correctly. Quantile is the only guardrail that takes a data-driven approach via the ensemble of neural networks to assess the risk of a price.

[xxx]Privacy concerns may exist when local models have so few weights (i.e. 4 or 5) that only a few architectures are possible. An attacker may intercept the weights and instantiate their own model, even using it to predict some of the original data. However, we use large enough local models that we do not believe that is a concern.

[xxxi]The most common non-RL methods for microgrid price-setting are iterative pricing methods (IP) (Liu et al., 2017b; Wang and Huang, 2016) in which buildings "bargain" with microgrids to reach equilibrium prices. We exclude these baselines because they require *each building* to develop their own demand forecasts. This requirement raises the computational barrier for entry by an order of magnitude. For comparison, if we had 10 microgrids with 10 buildings each, local RL requires training 10 models (10 microgrids), PFH and FedAvg requires 11 (10 microgrids + 1 central model), and IP requires 100 (10 buildings x 10 microgrids). Agwan et al. (2021a) also showed RL results in less volatile pricing curves and better performance compared to IP.

[xxxii]The relatively long and unique process of fossil fuel formation and the origin of the organisms that compose them, i.e. prehistoric plants that have journeyed through the Earth's mantle to arrive in their current state, would seem special enough to deserve some sort of piety or austere deference. At the risk of stating a tautology, fossil fuels are truly "fossil" fuels. I wonder if, given awareness, many would think twice about using the remnants of ancient plant bodies to fuel such frivolities as a summer weekend trip to Puerto Vallarta.

[xxxiii]The atmospheric reaction described, i.e. the gassification of fixed carbon into carbon dioxide after fossil fuel combustion, is of course an oversimplification, as fossil fuels are often complex materials containing many trace impurities. Many of these result in quite toxic products of combustion, and the ones of most concern at the time of the writing are sulfur dioxides, nitrous oxides, and ground level ozone. All of these gases are dangerous for a variety of acute and chronic health as well as climactic concerns. The field of study, known as environmental justice, which seeks to understand patterns of environmental disruption

that relate to the health of humans and how these are distributed relative to race, wealth, and other societal hierarchices, is important for understanding how acute environmental impacts are distributed. At the time of the writing, some level of "climate justice" has entered political discourse around climate solutions.

[xxxiv]We and much of the climate change community may use "carbon dioxide", GHG, and sometimes even "carbon" somewhat interchangeably, unfortunately. This notational fluidity is largely through convenience but partially due to the centrality of "carbon dioxide" to the entire climate change lexicon. A GHG's warming potential is defined in terms of carbon dioxide; specifically, it is a measure of how much energy the emissions of 1 ton of a gas will absorb over a given period of time, relative to the emissions of 1 ton of carbon dioxide. Furthermore, carbon dioxide is arguably the most physically impactful of the GHG suite and almost certainly the largest emitted by gaseous quantity and breadth of emitting sources. Thus, to an extent, to use "carbon dioxide" as a more abstract reference to the suite of GHGs under concern is somewhat fair. To use "carbon" in place of "carbon dioxide" (i.e., "carbon-free fuels" or "carbon-intense lifestyle") is somewhat less accurate and more a verbal convenience, given the overlap with elemental carbon which may be solid and thus unrelated to atmospheric carbon dioxide. It is worth noting that the entire stock of atmospheric carbon exists as carbon dioxide.

Bibliography

[n. d.]a. Chem Alkalines. https://commons.wikimedia.org/wiki/File:CNX_Chem_20_01_alkanes.png

[n. d.]. Climate Modeling History. https://www.carbonbrief.org/timeline-history-climate-modelling/.

[n. d.]. Energy Flows. https://serc.carleton.edu/integrate/teaching_materials/global_energy/activity3.html

[n. d.]b. Non Renewable Energy Resources. https://www.shalom-education.com/courses/ks3-physics/lessons/energy/topic/non-renewable-energy-resources/

[n. d.]. Petrofuel Formation. https://education.nationalgeographic.org/resource/petroleum.

[n. d.]. There could be 1.2 billion climate refugees by 2050. Here's what you need to know. https://www.zurich.com/en/media/magazine/2022/there-could-be-1-2-billion-climate-refugees-by-2050-here-s-what-you-need-to-know.

2020. Persistence of Energy Reduction Behaviors Following a Competition: A Quantitative Approach with Regression and Cluster-Based Learning. In *ACEEE Summer Study on Energy Efficiency in Buildings*.

Durmus Alp Emre Acar, Yue Zhao, Ramon Matas Navarro, Matthew Mattina, Paul N. Whatmough, and Venkatesh Saligrama. 2021. Federated Learning Based on Dynamic Regularization. https://doi.org/10.48550/ARXIV.2111.04263

Daniel Adelman and Canan Uçkun. 2019. Dynamic electricity pricing to smart homes. *Operations Research* 67, 6 (2019), 1520–1542.

Abdul Afram and Farrokh Janabi-Sharifi. 2014. Theory and applications of HVAC control systems–A review of model predictive control (MPC). *Building and Environment* 72 (2014), 343–355.

Energy Information Agency. [n. d.]. Energy Use Explained. ([n. d.]). https://www.eia.gov/energyexplained/use-of-energy/commercial-buildings.php

Utkarsha Agwan. 2020. Optimal Prosumer Aggregations: Design and Modeling. (2020).

Utkarsha Agwan, Lucas Spangher, William Arnold, Tarang Srivastava, Kameshwar Poolla, and Costas J Spanos. 2021a. Pricing in Prosumer Aggregations using Reinforcement Learning. In *Proceedings of the Twelfth ACM International Conference on Future Energy Systems*. 220–224.

Utkarsha Agwan, Lucas Spangher, William Arnold, Tarang Srivastava, Kameshwar Poolla, and Costas J. Spanos. 2021b. Pricing in Prosumer Aggregations using Reinforcement Learning. In *e-Energy '21: The Twelfth ACM International Conference on Future Energy Systems, Virtual Event, Torino, Italy, 28 June - 2 July, 2021*, Herman de Meer and Michela Meo (Eds.). ACM, 220–224.

Mohamed H Albadi and Ehab F El-Saadany. 2007. Demand response in electricity markets: An overview. In *2007 IEEE power engineering society general meeting*. IEEE, 1–5.

Aqeel Anwar and Arijit Raychowdhury. 2021. Multi-task federated reinforcement learning with adversaries. *arXiv preprint arXiv:2103.06473* (2021).

Manoj Ghuhan Arivazhagan, Vinay Aggarwal, Aaditya Kumar Singh, and Sunav Choudhary. 2019. Federated learning with personalization layers. *arXiv preprint arXiv:1912.00818* (2019).

William Arnold, Tarang Srivastava, Lucas Spangher, Utkarsha Agwan, and Costas Spanos. 2021a. Adapting Surprise Minimizing Reinforcement Learning Techniques for Transactive Control. In *Proceedings of the Twelfth ACM International Conference on Future Energy Systems* (Virtual Event, Italy) *(e-Energy '21)*. Association for Computing Machinery, New York, NY, USA, 488–492. https://doi.org/10.1145/3447555.3466590

William Arnold, Tarang Srivastava, Lucas Spangher, Utkarsha Agwan, and Costas Spanos. 2021b. Adapting Surprise Minimizing Reinforcement Learning Techniques for Transactive Control. In *Proceedings of the Twelfth ACM International Conference on Future Energy Systems*. 488–492.

Ailin Asadinejad, Alireza Rahimpour, Kevin Tomsovic, Hairong Qi, and Chien-fei Chen. 2018. Evaluation of residential customer elasticity for incentive based demand response programs. *Electric Power Systems Research* 158 (2018), 26–36.

Seyed Mohammad Attaran, Rubiyah Yusof, and Hazlina Selamat. 2014. Short review on HVAC components, mathematical model of HVAC system and different PID controllers. *International Review of Automatic Control* 7, 3 (2014), 263–70.

Ian Ayres, Sophie Raseman, and Alice Shih. 2012. Evidence from Two Large Field Experiments that Peer Comparison Feedback Can Reduce Residential Energy Usage. *The Journal of Law, Economics, and Organization* 29, 5 (08 2012), 992–1022. https://doi.org/10.1093/jleo/ews020 arXiv:http://oup.prod.sis.lan/jleo/article-pdf/29/5/992/2987656/ews020.pdf

Donald Azuatalam, Wee-Lih Lee, Frits de Nijs, and Ariel Liebman. 2020. Reinforcement learning for whole-building HVAC control and demand response. *Energy and AI* 2 (2020), 100020.

Edoardo Bacci and David Parker. 2020. Probabilistic guarantees for safe deep reinforcement learning. In *International Conference on Formal Modeling and Analysis of Timed Systems*. Springer, 231–248.

Leemon C Baird. 1994. Reinforcement learning in continuous time: Advantage updating. In *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94)*, Vol. 4. IEEE, 2448–2453.

Wele Gedara Chaminda Bandara, Naman Patel, Ali Gholami, Mehdi Nikkhah, Motilal Agrawal, and Vishal M Patel. 2022. AdaMAE: Adaptive Masking for Efficient Spatiotemporal Learning with Masked Autoencoders. *arXiv preprint arXiv:2211.09120* (2022).

Leonard E. Baum and Ted Petrie. 1966. Statistical Inference for Probabilistic Functions of Finite State Markov Chains. *Ann. Math. Statist.* 37, 6 (12 1966), 1554–1563. `https://doi.org/10.1214/aoms/1177699147`

Gary S Becker and Kevin M Murphy. 1988. A theory of rational addiction. *Journal of political Economy* 96, 4 (1988), 675–700.

Philipp Beiter, Aubryn Cooperman, Eric Lantz, Tyler Stehly, Matt Shields, Ryan Wiser, Thomas Telsnig, Lena Kitzing, Volker Berkhout, and Yuka Kikuchi. 2021. Wind power costs driven by innovation and experience with further reductions on the horizon. *Wiley Interdisciplinary Reviews: Energy and Environment* 10, 5 (2021), e398.

Curtis P Berlinguette, Yet-Ming Chiang, Jeremy N Munday, Thomas Schenkel, David K Fork, Ross Koningstein, and Matthew D Trevithick. 2019. Revisiting the cold case of cold fusion. *Nature* 570, 7759 (2019), 45–51.

Glen Berseth, Daniel Geng, Coline Devin, Chelsea Finn, Dinesh Jayaraman, and Sergey Levine. 2019. SMiRL: Surprise Minimizing RL in Dynamic Environments. *CoRR* abs/1912.05510 (2019). arXiv:1912.05510 `http://arxiv.org/abs/1912.05510`

Lori Bird, Debra Lew, Michael Milligan, E Maria Carlini, Ana Estanqueiro, Damian Flynn, Emilio Gomez-Lazaro, Hannele Holttinen, Nickie Menemenlis, Antje Orths, et al. 2016. Wind and solar energy curtailment: A review of international experience. *Renewable and Sustainable Energy Reviews* 65 (2016), 577–586.

Aaron Bloom, Josh Novacheck, Greg Brinkman, James McCalley, Armando Figueroa-Acevedo, Ali Jahanbani-Ardakani, Hussam Nosair, Abhinav Venkatraman, Jay Caspary, Dale Osborn, et al. 2021. The value of increased HVDC capacity between eastern and western US grids: The interconnections seam study. *IEEE Transactions on Power Systems* 37, 3 (2021), 1760–1769.

Rakesh P Borase, DK Maghade, SY Sondkar, and SN Pawar. 2021. A review of PID control, tuning methods and applications. *International Journal of Dynamics and Control* 9, 2 (2021), 818–827.

Dirk Börner, Jeroen Storm, Marco Kalz, and Marcus Specht. 2012. Energy Awareness Displays: Prototype for Personalised Energy Consumption Feedback. In *Proceedings of the 7th European Conference on Technology Enhanced Learning* (Saarbrücken, Germany) *(EC-TEL'12)*. Springer-Verlag, Berlin, Heidelberg, 471–476. https://doi.org/10.1007/978-3-642-33263-0_45

Djallel Bouneffouf, Irina Rish, and Charu Aggarwal. 2020. Survey on applications of multi-armed and contextual bandits. In *2020 IEEE Congress on Evolutionary Computation (CEC)*. IEEE, 1–8.

Gregory Brinkman, Dominique Bain, Grant Buster, Caroline Draxl, Paritosh Das, Jonathan Ho, Eduardo Ibanez, Ryan Jones, Sam Koebrich, Sinnott Murphy, et al. 2021. *The North American Renewable Integration Study (NARIS): A Canadian Perspective*. Technical Report. National Renewable Energy Lab.(NREL), Golden, CO (United States).

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. 2016. Openai gym. *arXiv preprint arXiv:1606.01540* (2016).

Michele Brunetti and F. Prodi. 2015. The climate system. *EPJ Web of Conferences* 98 (01 2015), 02001. https://doi.org/10.1051/epjconf/20159802001

Christoph Brunner, Gerda Deac, Sebastian Braun, and Christoph Zöphel. 2020. The future need for flexibility and the impact of fluctuating renewable power generation. *Renewable Energy* 149 (2020), 1314–1324.

Ian Raymond Cameron. 2012. *Nuclear fission reactors*. Springer Science & Business Media.

Josep G Canadell and Robert B Jackson. 2021. *Ecosystem collapse and climate change*. Springer.

Leonel Magalhães Carvalho, Ricardo Jorge Ferreira, Mauro Rosa, and Vladimiro Miranda. 2011. A chronological composite system adequacy assessment considering non-dispathable renewable energy sources and their integration strategies. (2011).

Bipartisan Policy Center. 2020. Annual energy outlook 2020. *Energy Information Administration, Washington, DC* 12 (2020), 1672–1679.

Michael M Cernea. 1997. *Hydropower dams and social impacts: a sociological perspective*. The World Bank.

Boxiao Chen, Selvaprabu Nadarajah, Parshan Pakiman, and Stefanus Jasin. 2020. Self-adapting Robustness in Demand Learning. *Available at SSRN 3734591* (2020).

Yujiao Chen, Leslie K Norford, Holly W Samuelson, and Ali Malkawi. 2018. Optimal control of HVAC and window systems for natural ventilation through reinforcement learning. *Energy and Buildings* 169 (2018), 195–205.

Marshini Chetty, A.J. Bernheim Brush, Brian R. Meyers, and Paul Johns. 2009. It's Not Easy Being Green: Understanding Home Computer Power Management. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Boston, MA, USA) *(CHI '09)*. ACM, New York, NY, USA, 1033–1042. https://doi.org/10.1145/1518701.1518860

Smart Energy Demand Coalition. 2014. Mapping demand response in Europe today. *Tracking Compliance with Article* 15 (2014).

Illinois Energy Commission. 2021. Rates and tarrifs. https://www.energybot.com/shop.html#/search-results-v3

Claire L Cooper, Graeme T Swindles, Ivan P Savov, Anja Schmidt, and Karen L Bacon. 2018. Evaluating the relationship between climate change and volcanism. *Earth-Science Reviews* 177 (2018), 238–247.

Francisco Costa and François Gerard. 2021. Hysteresis and the welfare effect of corrective policies: Theory and evidence from an energy-saving program. *Journal of Political Economy* 129, 6 (2021), 1705–1743.

Ben Cowley, Jose Luiz Moutinho, Chris Bateman, and Alvaro Oliveira. 2011. Learning principles and interaction design for "Green My Place": A massively multiplayer serious game. *Entertainment Computing* 2, 2 (2011), 103 – 113. https://doi.org/10.1016/j.entcom.2011.01.001 Serious Games Development and Applications.

Felix Creutzig, Jan Christoph Goldschmidt, Paul Lehmann, Eva Schmid, Felix von Blücher, Christian Breyer, Blanca Fernandez, Michael Jakob, Brigitte Knopf, Steffen Lohrey, et al. 2014. Catching two European birds with one renewable stone: Mitigating climate change and Eurozone crisis by an energy transition. *Renewable and Sustainable Energy Reviews* 38 (2014), 1015–1028.

Christoph Dann, Lihong Li, Wei Wei, and Emma Brunskill. 2019. Policy certificates: Towards accountable reinforcement learning. In *International Conference on Machine Learning*. PMLR, 1507–1516.

Hari Prasanna Das, Ioannis Konstantakopoulos, Aummul Baneen Manasawala, Tanya Veeravalli, Huihan Liu, and Costas J Spanos. 2020. Do Occupants in a Building exhibit patterns in Energy Consumption? Analyzing Clusters in Energy Social Games. (2020).

Laurence L Delina. 2017. *Accelerating Sustainable Energy Transition (s) in Developing Countries: The challenges of climate change and sustainable development*. Routledge.

Paul Denholm, Matthew O'Connell, Gregory Brinkman, and Jennie Jorgenson. 2015. *Overgeneration from solar energy in california. a field guide to the duck chart*. Technical Report. National Renewable Energy Lab.(NREL), Golden, CO (United States).

Michael Dennis, Natasha Jaques, Eugene Vinitsky, Alexandre Bayen, Stuart Russell, Andrew Critch, and Sergey Levine. 2020. Emergent complexity and zero-shot transfer via unsupervised environment design. *Advances in neural information processing systems* 33 (2020), 13049–13061.

Nikos Dimitriou, Anastasia Garbi, Kostas Vasilakis, Anthony Schoofs, Amr Taha, Manolis Nikiforakis, Sarantis Kotsilitis, Thanasis G Papaioannou, Dimosthenis Kotsopoulos, Cleopatra Bardaki, et al. 2018. ChArGED: Implementing a framework for improving energy efficiency in public buildings through IoTenabled energy disaggregation and serious games. In *2018 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*. IEEE, 65–70.

Xianzhong Ding, Wan Du, and Alberto Cerpa. 2019a. OCTOPUS: Deep Reinforcement Learning for Holistic Smart Building Control. In *Proceedings of the 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation* (New York, NY, USA) *(BuildSys '19)*. Association for Computing Machinery, New York, NY, USA, 326–335. https://doi.org/10.1145/3360322.3360857

Xiaohan Ding, Xiangxin Zhou, Yuchen Guo, Jungong Han, Ji Liu, et al. 2019b. Global sparse momentum sgd for pruning very deep neural networks. *Advances in Neural Information Processing Systems* 32 (2019).

Alexander W Dowling, Tian Zheng, and Victor M Zavala. 2017. Economic assessment of concentrated solar power technologies: A review. *Renewable and Sustainable Energy Reviews* 72 (2017), 1019–1032.

Mildred Spiewak Dresselhaus and IL Thomas. 2001. Alternative energy technologies. *Nature* 414, 6861 (2001), 332–337.

L El Chaar, N El Zein, et al. 2011. Review of photovoltaic technologies. *Renewable and sustainable energy reviews* 15, 5 (2011), 2165–2175.

Eiman Y. Elbanhawy, Andrew F. G. Smith, and John Moore. 2016. Towards an Ambient Awareness Interface for Home Battery Storage System. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct* (Heidelberg, Germany) *(UbiComp '16)*. ACM, New York, NY, USA, 1608–1613. https://doi.org/10.1145/2968219.2968557

Michael Emmanuel, Kate Doubleday, Burcin Cakir, Marija Marković, and Bri-Mathias Hodge. 2020. A review of power system planning and operational models for flexibility assessment in high solar energy penetration scenarios. *Solar Energy* 210 (2020), 169–180.

Mohammad Esrafilian-Najafabadi and Fariborz Haghighat. 2021. Occupancy-based HVAC control systems in buildings: A state-of-the-art review. *Building and Environment* 197 (2021), 107810.

Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. 2020. Personalized federated learning: A meta-learning approach. *arXiv preprint arXiv:2002.07948* (2020).

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017a. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*. PMLR, 1126–1135.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017b. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. *CoRR* abs/1703.03400 (2017). arXiv:1703.03400 http://arxiv.org/abs/1703.03400

David Fork and Ross Koningstein. 2021. How Engineers Can Disrupt Climate Change. *IEEE Spectrum* 58, 7 (2021), 24–29.

Andrew Forney and Elias Bareinboim. 2019. Counterfactual Randomization: Rescuing Experimental Studies from Obscured Confounding. *Proceedings of the AAAI Conference on Artificial Intelligence* 33, 01 (Jul. 2019), 2454–2461. https://doi.org/10.1609/aaai.v33i01.33012454

Emily Frame, Ross Barnowski, Donald Gunter, Lucian Mihailescu, and Kai Vetter. 2022. A Dual-Modality Volumetric Gamma-Ray Imager for Near-Field Applications. *IEEE Transactions on Nuclear Science* (2022).

Mikhail Frank, Jürgen Leitner, Marijn Stollenga, Alexander Förster, and Jürgen Schmidhuber. 2014. Curiosity driven reinforcement learning for motion planning on humanoids. *Frontiers in neurorobotics* 7 (2014), 25.

Jon Froehlich, Tawanna Dillahunt, Predrag Klasnja, Jennifer Mankoff, Sunny Consolvo, Beverly Harrison, and James A Landay. 2009. UbiGreen: Investigating a Mobile Tool for Tracking and Supporting Green Transportation Habits. (2009). http://www.ecorio.org/

Dian-ce Gao, Yongjun Sun, and Yuehong Lu. 2015. A robust demand response control of commercial buildings for smart grid under load prediction uncertainty. *Energy* 93 (2015), 275–283.

Daphne Geelen, David Keyson, Stella Boess, and Han Brezet. 2012. Exploring the use of a game to stimulate energy saving in households. *Journal of Design Research 14* 10, 1-2 (2012), 102–120.

Guang Geng and GM Geary. 1993. On performance and tuning of PID controllers in HVAC systems. In *Proceedings of IEEE International Conference on Control and Applications*. IEEE, 819–824.

Alan S. Gerber, Donald P. Green, and Christopher W. Larimer. 2008. Social Pressure and Voter Turnout: Evidence from a Large-Scale Field Experiment. *American Political Science Review* 102, 1 (2008), 33–48. https://doi.org/10.1017/S000305540808009X

Kenneth Gillingham, Richard Newell, and Karen Palmer. 2006. Energy efficiency policies: a retrospective examination. *Annu. Rev. Environ. Resour.* 31 (2006), 161–192.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep learning*. MIT press.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572* (2014).

Erik Granstedt, Erik Trask, Hiroshi Gota, Ian Langmore, Jesus Romero, Matthew Thompson, Michael Dikovsky, Peter Norgaard, Roberto Mendoza, Ted Baltz, et al. 2018. The Plasma Debugger. (2018).

Peter J Gregory, John SI Ingram, and Michael Brklacich. 2005. Climate change and food security. *Philosophical Transactions of the Royal Society B: Biological Sciences* 360, 1463 (2005), 2139–2148.

Sam Gunn, Doseok Jang, Orr Paradise, Lucas Spangher, and Costas J Spanos. [n. d.]. Adversarial poisoning attacks on reinforcement learning-driven energy pricing. In *Proceedings of the 8th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation (BuildSys '22)*.

Li Guoqiang, Jin Fei, Liu Jingli, Liu Xiaoliang, Lu Xiaohui, Tang Min, Song Zhanhui, and Liu Zhonghui. 2021. A Multi-Source Dispatching Model with Considering the Nuclear Power Plants Dispatching and Wind Power Accommodation. In *2021 IEEE 4th International Electrical and Energy Conference (CIEEC)*. IEEE, 1–6.

Anton Gustafsson and Magnus Gyllenswärd. 2005. The Power-aware Cord: Energy Awareness Through Ambient Information Display. In *CHI '05 Extended Abstracts on Human Factors in Computing Systems* (Portland, OR, USA) *(CHI EA '05)*. ACM, New York, NY, USA, 1423–1426. https://doi.org/10.1145/1056808.1056932

Samuel Gyamfi, Susan Krumdieck, and Tania Urmee. 2013. Residential peak electricity demand response—Highlights of some behavioural issues. *Renewable and Sustainable Energy Reviews* 25 (2013), 71–77.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. 2018. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*. PMLR, 1861–1870.

G. Han, S. Lee, J. Lee, K. Lee, and J. Bae. 2021. Deep-learning- and reinforcement-learning-based profitable strategy of a grid-level energy storage system for the smart grid. *Journal of Energy Storage* 41 (2021). https://doi.org/10.1016/j.est.2021.102868 cited By 1.

Mengjie Han, Ross May, Xingxing Zhang, Xinru Wang, Song Pan, Yan Da, and Yuan Jin. 2020. A novel reinforcement learning method for improving occupant comfort via window opening and closing. *Sustainable Cities and Society* 61 (2020), 102247.

Cullen S Hendrix and Idean Salehyan. 2012. Climate change, rainfall, and social conflict in Africa. *Journal of peace research* 49, 1 (2012), 35–50.

Rodrigo Henriquez-Auba, Patricia Pauli, Dileep Kalathil, Duncan S Callaway, and Kameshwar Poolla. 2018. The Sharing Economy for Residential Solar Generation. In *2018 IEEE Conference on Decision and Control (CDC)*. IEEE, 7322–7329.

Ashley Hill, Antonin Raffin, Maximilian Ernestus, Adam Gleave, Anssi Kanervisto, Rene Traore, Prafulla Dhariwal, Christopher Hesse, Oleg Klimov, Alex Nichol, Matthias Plappert, Alec Radford, John Schulman, Szymon Sidor, and Yuhuai Wu. 2018. Stable Baselines. `https://github.com/hill-a/stable-baselines`.

Briland Hitaj, Giuseppe Ateniese, and Fernando Perez-Cruz. 2017. Deep Models Under the GAN: Information Leakage from Collaborative Deep Learning. `https://doi.org/10.48550/ARXIV.1702.07464`

Carl-Johan Hoel, Katherine Driggs-Campbell, Krister Wolff, Leo Laine, and Mykel J Kochenderfer. 2019. Combining planning and deep reinforcement learning in tactical decision making for autonomous driving. *IEEE transactions on intelligent vehicles* 5, 2 (2019), 294–305.

Ian M Hoffman, Gregory Rybka, Greg Leventis, Charles A Goldman, Lisa Schwartz, Megan Billingsley, and Steven Schiller. 2015. The total cost of saving electricity through utility customer-funded energy efficiency programs: estimates at the national, state, sector and program level. *Berkeley Lab Technical Brief* (2015).

Tiffany Grace Holmes. 2007. Eco-visualization: Combining Art and Technology to Reduce Energy Consumption. In *Proceedings of the 6th ACM SIGCHI Conference on Creativity &Amp; Cognition* (Washington, DC, USA) *(C&C '07)*. ACM, New York, NY, USA, 153–162. `https://doi.org/10.1145/1254960.1254982`

José Iria, Nuno Fonseca, Fernando Cassola, António Barbosa, Filipe Soares, António Coelho, and Aydogan Ozdemir. 2020. A gamification platform to foster energy efficiency in office buildings. *Energy and Buildings* 222 (2020), 110101.

Sofiah Jamil. 2017. Fighting climate change from Singapore. (2017).

Doseok Jang, Lucas Spangher, Manan Khattar, Utkarsha Agwan, and Costas Spanos. 2021a. Using Meta Reinforcement Learning to Bridge the Gap between Simulation and Experiment in Energy Demand Response. In *Proceedings of the Twelfth ACM International Conference on Future Energy Systems* (Virtual Event, Italy) *(e-Energy '21)*. Association for Computing Machinery, New York, NY, USA, 483–487. `https://doi.org/10.1145/3447555.3466589`

Doseok Jang, Lucas Spangher, Manan Khattar, Utkarsha Agwan, and Costas Spanos. 2021b. Using Meta Reinforcement Learning to Bridge the Gap between Simulation and Experiment in Energy Demand Response. In *Proceedings of the Twelfth ACM International Conference on Future Energy Systems*. 483–487.

Doseok Jang, Lucas Spangher, Selvaprabu Nadarajah, and Costas Spanos. 2022a. Decarbonizing Buildings via Energy Demand Response and Deep Reinforcement Learning: The Deployment Value of Supervisory Planning and Guardrails. *Available at SSRN 4078206* (2022).

Doseok Jang, Lucas Spangher, Tarang Srivistava, Manan Khattar, Utkarsha Agwan, Selvaprabu Nadarajah, and Costas Spanos. 2021c. Offline-online reinforcement learning for

energy pricing in office demand response: lowering energy and data costs. In *Proceedings of the 8th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*. 131–139.

Doseok Jang, Lucas Spangher, Tarang Srivistava, Manan Khattar, Utkarsha Agwan, Selvaprabu Nadarajah, and Costas Spanos. 2021d. Offline-Online Reinforcement Learning for Energy Pricing in Office Demand Response: Lowering Energy and Data Costs. In *Proceedings of the 8th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation* (Coimbra, Portugal) *(BuildSys '21)*. Association for Computing Machinery, New York, NY, USA, 131–139. https://doi.org/10.1145/3486611.3486668

Doseok Jang, Larry Yan, Lucas Spangher, and Costas J. Spanos. 2022b. Personalized Federated Hypernetworks for Privacy Preservation in Multi-Task Reinforcement Learning. https://doi.org/10.48550/ARXIV.2210.06820

Liyan Jia, Qing Zhao, and Lang Tong. 2013. Retail pricing for stochastic demand with unknown parameters: An online machine learning approach. In *2013 51st Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 1353–1358.

Minqi Jiang, Michael Dennis, Jack Parker-Holder, Jakob Foerster, Edward Grefenstette, and Tim Rocktäschel. 2021. Replay-guided adversarial environment design. *Advances in Neural Information Processing Systems* 34 (2021), 1884–1897.

Javier Jiménez-Raboso, Alejandro Campoy-Nieves, Antonio Manjavacas-Lucas, Juan Gómez-Romero, and Miguel Molina-Solana. 2021. Sinergym: a building simulation and control framework for training reinforcement learning agents. In *Proceedings of the 8th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*. 319–323.

Brandon J Johnson, Michael R Starke, Omar A Abdelaziz, Roderick K Jackson, and Leon M Tolbert. 2015. A dynamic simulation tool for estimating demand response potential from residential loads. In *2015 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT)*. IEEE, 1–5.

Masato Kasahara, Tadahiko Matsuba, Yoshiaki Kuzuu, Takanori Yamazaki, et al. 1999. Design and tuning of robust PID controller for HVAC systems. *AsHRAE Transactions* 105 (1999), 154.

Anjukan Kathirgamanathan, Eleni Mangina, and Donal P Finn. 2021. Development of a Soft Actor Critic deep reinforcement learning approach for harnessing energy flexibility in a Large Office building. *Energy and AI* 5 (2021), 100101.

Rachneet Kaur, Clara Schaye, Kevin Thompson, Daniel C Yee, Rachel Zilz, RS Sreenivas, and Richard B Sowers. 2021. Machine learning and price-based load scheduling for an optimal IoT control in the smart and frugal home. *Energy and AI* 3 (2021), 100042.

David W Keith, KENTON Heidel, and Robert Cherry. 2010. Capturing CO2 from the atmosphere: rationale and process design considerations. , 107–126 pages.

Colin P Kelley, Shahrzad Mohtadi, Mark A Cane, Richard Seager, and Yochanan Kushnir. 2015. Climate change in the Fertile Crescent and implications of the recent Syrian drought. *Proceedings of the national Academy of Sciences* 112, 11 (2015), 3241–3246.

David J Ketchen and Christopher L Shook. 1996. The application of cluster analysis in strategic management research: an analysis and critique. *Strategic management journal* 17, 6 (1996), 441–458.

Rahul Kidambi, Aravind Rajeswaran, Praneeth Netrapalli, and Thorsten Joachims. 2020. Morel: Model-based offline reinforcement learning. *Advances in neural information processing systems* 33 (2020), 21810–21823.

Ånund Killingtveit. 2019. Hydropower. In *Managing Global Warming*. Elsevier, 265–315.

Seung-Jun Kim and Geogios B Giannakis. 2016. An online convex optimization approach to real-time energy pricing for demand response. *IEEE Transactions on Smart Grid* 8, 6 (2016), 2784–2793.

Tanyoung Kim, Hwajung Hong, and Brian Magerko. 2009. Coralog: Use-aware Visualization Connecting Human Micro-activities to Environmental Change. In *CHI '09 Extended Abstracts on Human Factors in Computing Systems* (Boston, MA, USA) *(CHI EA '09)*. ACM, New York, NY, USA, 4303–4308. `https://doi.org/10.1145/1520340.1520657`

Taehoon Kim, Wentao Song, Dae-Yong Son, Luis K Ono, and Yabing Qi. 2019. Lithium-ion batteries: outlook on present, future, and hybridized technologies. *Journal of materials chemistry A* 7, 7 (2019), 2942–2964.

Y. Kim. 2018. Optimal Price Based Demand Response of HVAC Systems in Multizone Office Buildings Considering Thermal Preferences of Individual Occupants Buildings. *IEEE Transactions on Industrial Informatics* 14, 11 (2018), 5060–5073.

Nicolas Knudde, Joachim van der Herten, Tom Dhaene, and Ivo Couckuyt. 2017. GPflowOpt: A Bayesian optimization library using tensorflow. *arXiv preprint arXiv:1711.03845* (2017).

Ross Koningstein. 2022. Conversation on Energy Cost. Personal Conversation.

Ross Koningstein and David K Fork. 2014. What it would really take to reverse climate change. (2014).

Ioannis C Konstantakopoulos, Andrew R Barkan, Shiying He, Tanya Veeravalli, Huihan Liu, and Costas Spanos. 2019a. A deep learning and gamification approach to improving human-building interaction and energy efficiency in smart infrastructure. *Applied energy* 237 (2019), 810–821.

Ioannis C Konstantakopoulos, Hari Prasanna Das, Andrew R Barkan, Shiying He, Tanya Veeravalli, Huihan Liu, Aummul Baneen Manasawala, Yu-Wen Lin, and Costas J Spanos. 2019b. Design, benchmarking and explainability analysis of a game-theoretic framework towards energy efficiency in smart infrastructure. *arXiv preprint arXiv:1910.07899* (2019).

I. C. Konstantakopoulos, L. J. Ratliff, M. Jin, and C. J. Spanos. 2017. Leveraging correlations in utility learning. In *2017 American Control Conference (ACC)*. 5249–5256. https://doi.org/10.23919/ACC.2017.7963770

Vikram Krishnamurthy. 2016. *Partially observed Markov decision processes*. Cambridge university press.

Nir Kshetri and Jeffrey M. Voas. 2017. Hacking Power Grids: A Current Problem. *Computer* 50, 12 (2017), 91–95.

Sang Gyu Kwak and Jong Hae Kim. 2017. Central limit theorem: the cornerstone of modern statistics. *Korean journal of anesthesiology* 70, 2 (2017), 144–156.

Dohyun Kwon, Joohyung Jeon, Soohyun Park, Joongheon Kim, and Sungrae Cho. 2020. Multiagent DDPG-based deep learning for smart ocean federated learning IoT networks. *IEEE Internet of Things Journal* 7, 10 (2020), 9895–9903.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems* 30 (2017).

Ian Langmore, John Platt, Michael Dikovsky, Nathan Neibauer, Peter Norgaard, Rob von Behren, Scott Davidson Geraedts, and Ted Baltz. 2019. Fusion Plasma Reconstruction. (2019).

Siddique Latif, Heriberto Cuayáhuitl, Farrukh Pervez, Fahad Shamshad, Hafiz Shehbaz Ali, and Erik Cambria. 2022. A survey on deep reinforcement learning for audio-based applications. *Artificial Intelligence Review* (2022), 1–48.

Henry N Le Houérou. 1996. Climate change, drought and desertification. *Journal of arid Environments* 34, 2 (1996), 133–185.

Chaojie Li, Chen Liu, Xinghuo Yu, Ke Deng, Tingwen Huang, and Liangchen Liu. 2018. Integrating Demand Response and Renewable Energy In Wholesale Market.. In *IJCAI*. 382–388.

Tian Li, Maziar Sanjabi, Ahmad Beirami, and Virginia Smith. 2019. Fair Resource Allocation in Federated Learning. https://doi.org/10.48550/ARXIV.1905.10497

Eric Liang, Richard Liaw, Robert Nishihara, Philipp Moritz, Roy Fox, Ken Goldberg, Joseph Gonzalez, Michael Jordan, and Ion Stoica. 2018. RLlib: Abstractions for distributed reinforcement learning. In *International Conference on Machine Learning*. PMLR, 3053–3062.

Rebecca Lindsey. 2020. Climate change: global sea level. *Climate. gov, last modified August* 14 (2020).

Paulo Lissa, Conor Deane, Michael Schukat, Federico Seri, Marcus Keane, and Enda Barrett. 2021. Deep reinforcement learning for home energy management system control. *Energy and AI* 3 (2021), 100043. `https://www.sciencedirect.com/science/article/pii/S2666546820300434`

Chunhua Liu, KT Chau, Diyun Wu, and Shuang Gao. 2013. Opportunities and challenges of vehicle-to-home, vehicle-to-vehicle, and vehicle-to-grid technologies. *Proc. IEEE* 101, 11 (2013), 2409–2427.

Nian Liu, Xinghuo Yu, Cheng Wang, Chaojie Li, Li Ma, and Jinyong Lei. 2017b. Energy-sharing model with price-based demand response for microgrids of peer-to-peer prosumers. *IEEE Transactions on Power Systems* 32, 5 (2017), 3569–3583.

Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. 2017a. Delving into Transferable Adversarial Examples and Black-box Attacks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Ryan Lowe, Yi I Wu, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. 2017. Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in neural information processing systems* 30 (2017).

Kai Ma, Guoqiang Hu, and Costas J Spanos. 2015. A cooperative demand response scheme using punishment mechanism and application to industrial refrigerated warehouses. *IEEE Transactions on Industrial Informatics* 11, 6 (2015), 1520–1531.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Batta Mahesh. 2020. Machine learning algorithms-a review. *International Journal of Science and Research (IJSR).[Internet]* 9 (2020), 381–386.

Trieu T Mai, Paige Jadun, Jeffrey S Logan, Colin A McMillan, Matteo Muratori, Daniel C Steinberg, Laura J Vimmerstedt, Benjamin Haley, Ryan Jones, and Brent Nelson. 2018. *Electrification futures study: scenarios of electric technology adoption and power consumption for the United States*. Technical Report. National Renewable Energy Lab.(NREL), Golden, CO (United States).

Franck Marchis, Daniel Hestroffer, Pascal Descamps, Jérôme Berthier, Antonin H Bouchez, Randall D Campbell, Jason CY Chin, Marcos A Van Dam, Scott K Hartman, Erik M Johansson, et al. 2006a. A low density of 0.8 g cm-3 for the Trojan binary asteroid 617 Patroclus. *Nature* 439, 7076 (2006), 565–567.

F Marchis, Mikko Kaasalainen, EFY Hom, J Berthier, J Enriquez, D Hestroffer, D Le Mignant, and I De Pater. 2006b. Shape, size and multiplicity of main-belt asteroids: I. Keck Adaptive Optics survey. *Icarus* 185, 1 (2006), 39–63.

Jochen Markard. 2018. The next phase of the energy transition and its implications for research and policy. *Nature Energy* 3, 8 (2018), 628–633.

Valérie Masson-Delmotte, Panmao Zhai, Anna Pirani, Sarah L Connors, Clotilde Péan, Sophie Berger, Nada Caud, Y Chen, L Goldfarb, MI Gomis, et al. 2021. Climate change 2021: the physical science basis. *Contribution of working group I to the sixth assessment report of the intergovernmental panel on climate change* 2 (2021).

Doug McKenzie-Mohr and P Wesley Schultz. 2014. Choosing effective behavior change tools. *Social Marketing Quarterly* 20, 1 (2014), 35–46.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*. PMLR, 1273–1282.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)* 54, 6 (2021), 1–35.

Luca Melis, Congzheng Song, Emiliano De Cristofaro, and Vitaly Shmatikov. 2018. Exploiting Unintended Feature Leakage in Collaborative Learning. `https://doi.org/10.48550/ARXIV.1805.04049`

Efstathios E Stathis Michaelides. 2012. *Alternative energy sources*. Springer Science & Business Media.

Jerzy Mikulik. 2018. Energy demand patterns in an office building: a case study in Kraków (Southern Poland). *Sustainability* 10, 8 (2018), 2901.

Clayton Miller and Forrest Meggers. 2017. The Building Data Genome Project: An open, public data set from non-residential building electrical meters. *Energy Procedia* 122 (2017), 439–444.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. 2013. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602* (2013).

Chris Mooney. 2014. There'sa surprisingly strong link between climate change and violence. *The Washington Post* 22 (2014).

Maria Morales-Calderón, John Richard Stauffer, Lynne Anne Hillenbrand, R Gutermuth, Inseok Song, Luisa M Rebull, P Plavchan, John Michael Carpenter, Barbara Anne Whitney, K Covey, et al. 2011. YSOVAR: The first sensitive, wide-area, mid-infrared photometric monitoring of the Orion Nebula Cluster. *The Astrophysical Journal* 733, 1 (2011), 50.

Ted Mortenson. 2022. `https://realpars.com/pid-controller/`

Susanne C Moser and Lisa Dilling. 2007. Toward the social tipping point: Creating a climate for change. *Creating a climate for change: Communicating climate change and facilitating social change* (2007), 491–516.

Raghuraman Mudumbai, Soura Dasgupta, and Brian B Cho. 2012. Distributed control for optimal economic dispatch of a network of heterogeneous power generators. *IEEE Transactions on Power Systems* 27, 4 (2012), 1750–1760.

Latha Karthigaa Murugesan, Rashina Hoda, and Zoran Salcic. 2013. Investigating Visualization of Energy Consumption. In *Proceedings of the 14th Annual ACM SIGCHI NZ* (Christchurch, New Zealand) *(CHINZ '13)*. ACM, New York, NY, USA, Article 12, 1 pages. https://doi.org/10.1145/2542242.2542255

Kari C Nadeau, Ioana Agache, Marek Jutel, Isabella Annesi Maesano, Mübeccel Akdis, Vanitha Sampath, Gennaro d'Amato, Lorenzo Cecchi, Claudia Traidl-Hoffmann, and Cezmi A Akdis. 2022. Climate change: A call to action for the united nations. , 1087–1090 pages.

Chetan Nadiger, Anil Kumar, and Sherine Abdelhak. 2019. Federated reinforcement learning for fast personalization. In *2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*. IEEE, 123–127.

Patrick Ngatchou, Anahita Zarei, and A El-Sharkawi. 2005. Pareto multi objective optimization. In *Proceedings of the 13th international conference on, intelligent systems application to power systems*. IEEE, 84–91.

NREL. 2012. Wind Systems Integration Basics. (2012).

OpenEI. 2017. Time of Use pricing. https://openei.org/apps/USURDB/rate/view/5cbf78b25457a34e40671081#3__Energy

Eric O'Shaughnessy, Jesse R Cruce, and Kaifeng Xu. 2020. Too much of a good thing? Global trends in the curtailment of solar PV. *Solar Energy* 208 (2020), 1068–1077.

Bikash Pal and Balarko Chaudhuri. 2006. *Robust control in power systems*. Springer Science & Business Media.

T.G. Papaioannou and G.D. Stamoulis. 2018. Teaming and competition for demand-side management in office buildings. *2017 IEEE International Conference on Smart Grid Communications, SmartGridComm 2017* 2018-January (2018), 332–337. https://doi.org/10.1109/SmartGridComm.2017.8340734

Thanasis G Papaioannou, Nikos Dimitriou, Kostas Vasilakis, Anthony Schoofs, Manolis Nikiforakis, Fabian Pursche, Nikolay Deliyski, Amr Taha, Dimosthenis Kotsopoulos, Cleopatra Bardaki, et al. 2018. An IoT-based gamified approach for reducing occupants' energy wastage in public buildings. *Sensors* 18, 2 (2018), 537.

Nicolas Papernot, Patrick D. McDaniel, and Ian J. Goodfellow. 2016. Transferability in Machine Learning: from Phenomena to Black-Box Attacks using Adversarial Samples. *CoRR* abs/1605.07277 (2016). arXiv:1605.07277

Jack Parker-Holder, Minqi Jiang, Michael Dennis, Mikayel Samvelyan, Jakob Foerster, Edward Grefenstette, and Tim Rocktäschel. 2022. Evolving Curricula with Regret-Based Environment Design. https://doi.org/10.48550/ARXIV.2203.01302

Prachi Patel. 2018. New battery tech launches in drones [news]. *IEEE Spectrum* 55, 7 (2018), 7–9.

Anay Pattanaik, Zhenyi Tang, Shuijing Liu, Gautham Bommannan, and Girish Chowdhary. 2017. Robust deep reinforcement learning with adversarial attacks. *arXiv preprint arXiv:1712.03632* (2017).

Heather Payne. 2018. Sharing Negawatts: Property Law, Electricity Data, and Facilitating the Energy Sharing Economy. *Penn St. L. Rev.* 123 (2018), 355.

Thijs Peirelinck, Hussain Kazmi, Brida V Mbuwir, Chris Hermans, Fred Spiessens, Johan Suykens, and Geert Deconinck. 2022. Transfer learning in demand response: A review of algorithms for data-efficient modelling and control. *Energy and AI* 7 (2022), 100126.

PG&E. 2021. Rates and tariffs. `https://www.pge.com/tariffs/electric.shtml#RESELEC`

Lara S. G. Piccolo, Cecília Baranauskas, Miriam Fernandez, Harith Alani, and Anna de Liddo. 2014. Energy Consumption Awareness in the Workplace: Technical Artefacts and Practices. In *Proceedings of the 13th Brazilian Symposium on Human Factors in Computing Systems* (Foz do Iguau, Brazil) *(IHC '14)*. Sociedade Brasileira de Computação, Porto Alegre, Brazil, Brazil, 41–50. `http://dl.acm.org/citation.cfm?id=2738055.2738065`

Etienne Piguet, Antoine Pécoud, and Paul De Guchteneire. 2011. Migration and climate change: An overview. *Refugee Survey Quarterly* 30, 3 (2011), 1–23.

John Podesta and Peter Ogden. 2008. The security implications of climate change. *Washington Quarterly* 31, 1 (2008), 115–138.

Deb Polson and Cassandra Selin. 2012. The ECOS Green Buildings Project: Data Dramatization, Visualization and Manipulation. In *Proceedings of the Second International Conference on ICT As Key Technology Against Global Warming* (Vienna, Austria) *(ICT-GLOW'12)*. Springer-Verlag, Berlin, Heidelberg, 33–43. `https://doi.org/10.1007/978-3-642-32606-6_3`

LA Prashanth, Michael C Fu, et al. 2022. Risk-Sensitive Reinforcement Learning via Policy Gradient Search. *Foundations and Trends® in Machine Learning* 15, 5 (2022), 537–693.

William Prindle and Scott Finlinson. 2011. How organizations can drive behavior-based energy efficiency. In *Energy, sustainability and the environment*. Elsevier, 305–335.

Martin L Puterman. 2014. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons.

Jiaju Qi, Qihao Zhou, Lei Lei, and Kan Zheng. 2021. Federated reinforcement learning: Techniques, applications, and open challenges. *arXiv preprint arXiv:2108.11887* (2021).

Filipe Quintal, Nuno J. Nunes, Adrian Ocneanu, and Mario Berges. 2010. SINAIS: Home Consumption Package: A Low-cost Eco-feedback Energy-monitoring Research Platform. In *Proceedings of the 8th ACM Conference on Designing Interactive Systems* (Aarhus,

Denmark) *(DIS '10)*. ACM, New York, NY, USA, 419–421. https://doi.org/10.1145/1858171.1858252

Ana Radovanovic, Ross Koningstein, Ian Schneider, Bokan Chen, Alexandre Duarte, Binz Roy, Diyue Xiao, Maya Haridasan, Patrick Hung, Nick Care, et al. 2021. Carbon-aware computing for datacenters. *arXiv preprint arXiv:2106.11750* (2021).

Samuel Rahardian, Bentang Arief Budiman, Poetro Lebdo Sambegoro, and Ignatius Pulung Nurprasetio. 2019. Review of solid-state battery technology progress. In *2019 6th International Conference on Electric Vehicular Technology (ICEVT)*. IEEE, 310–315.

Amin Rakhsha, Xuezhou Zhang, Xiaojin Zhu, and Adish Singla. 2021. Reward Poisoning in Reinforcement Learning: Attacks Against Unknown Learners in Unknown Environments. *CoRR* abs/2102.08492 (2021). arXiv:2102.08492

Tabish Rashid, Mikayel Samvelyan, Christian Schroeder, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. 2018. Qmix: Monotonic value function factorisation for deep multi-agent reinforcement learning. In *International Conference on Machine Learning*. PMLR, 4295–4304.

Lillian J. Ratliff, Ming Jin, Ioannis C. Konstantakopoulos, Costas Spanos, and S. Shankar Sastry. 2014. Social game for building energy efficiency: Incentive design. In *2014 52nd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. 1011–1018.

L. J. Ratliff, M. Jin, I. C. Konstantakopoulos, C. Spanos, and S. S. Sastry. 2014. Social game for building energy efficiency: Incentive design. In *2014 52nd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. 1011–1018. https://doi.org/10.1109/ALLERTON.2014.7028565

Shafiqur Rehman, Luai M Al-Hadhrami, and Md Mahbub Alam. 2015. Pumped hydro energy storage system: A technological review. *Renewable and Sustainable Energy Reviews* 44 (2015), 586–598.

Jianji Ren, Haichao Wang, Tingting Hou, Shuai Zheng, and Chaosheng Tang. 2019. Federated learning-based computation offloading optimization in edge computing-supported internet of things. *IEEE Access* 7 (2019), 69194–69201.

Timothy J Richards and Gareth P Green. 2003. Economic hysteresis in variety selection. *Journal of Agricultural and Applied Economics* 35, 1 (2003), 1–14.

John G Rodgers. 2011. *Residential resource use feedback: exploring ambient and artistic approaches*. Ph. D. Dissertation. Communication, Art & Technology: School of Interactive Arts and Technology.

Jason R Rohr, Andrew P Dobson, Pieter TJ Johnson, A Marm Kilpatrick, Sara H Paull, Thomas R Raffel, Diego Ruiz-Moreno, and Matthew B Thomas. 2011. Frontiers in climate change–disease research. *Trends in ecology & evolution* 26, 6 (2011), 270–277.

David Rolnick, Priya L Donti, Lynn H Kaack, Kelly Kochanski, Alexandre Lacoste, Kris Sankaran, Andrew Slavin Ross, Nikola Milojevic-Dupont, Natasha Jaques, Anna Waldman-Brown, et al. 2022. Tackling climate change with machine learning. *ACM Computing Surveys (CSUR)* 55, 2 (2022), 1–96.

M. Rosenstein and A. Barto. 2002. *Supervised Learning Combined with an Actor-Critic Architecture TITLE2:*. Technical Report. USA.

Sonny Rosenthal, Edmund Lee, Shirley Ho, and Benjamin Detenber. 2013. Perceptions of climate change in Singapore and the United States.

Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. 2011. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 627–635.

Marianna Russo, Mel T Devine, and Paul Cuffe. [n. d.]. Blockchain Trading of Tokenized Electricity Using Tranched Power Delivery Contracts. *Available at SSRN 4146283* ([n. d.]).

Idean Salehyan. 2014. Climate change and conflict: Making sense of disparate findings. , 5 pages.

Luísa Schmidt, Ana Horta, Sérgio Pereira, and Ana Delicado. 2015. The Fukushima nuclear disaster and its effects on media framing of fission and fusion energy technologies. In *2015 4th International Conference on Advancements in Nuclear Instrumentation Measurement Methods and their Applications (ANIMMA)*. IEEE, 1–11.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* (2017).

Peter M Schwarz and Joseph A Cochran. 2013. Renaissance or Requiem: Is Nuclear Energy Cost Effective in a Post-Fukushima World? *Contemporary Economic Policy* 31, 4 (2013), 691–707.

Thayer Ted Scudder. 2012. *The future of large dams: Dealing with social, environmental, institutional and political costs*. Routledge.

Tim Seyde, Igor Gilitschenski, Wilko Schwarting, Bartolomeo Stellato, Martin Riedmiller, Markus Wulfmeier, and Daniela Rus. 2021. Is bang-bang control all you need? solving continuous control with bernoulli policies. *Advances in Neural Information Processing Systems* 34 (2021), 27209–27221.

Aviv Shamsian, Aviv Navon, Ethan Fetaya, and Gal Chechik. 2021. Personalized federated learning using hypernetworks. In *International Conference on Machine Learning*. PMLR, 9489–9502.

Eugene Yu Shchetinin. 2019. Modeling the energy consumption of smart buildings using artificial intelligence.. In *ITTMM (Selected Papers)*. 130–140.

Amin Shojaeighadikolaei, Arman Ghasemi, Kailani R Jones, Alexandru G Bardas, Morteza Hashemi, and Reza Ahmadi. 2021. Demand responsive dynamic pricing framework for prosumer dominated microgrids using multiagent reinforcement learning. In *2020 52nd North American Power Symposium (NAPS)*. IEEE, 1–6.

Reza Shokri and Vitaly Shmatikov. 2015. Privacy-Preserving Deep Learning. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security* (Denver, Colorado, USA) *(CCS '15)*. Association for Computing Machinery, New York, NY, USA, 1310–1321. https://doi.org/10.1145/2810103.2813687

Pierluigi Siano. 2014. Demand response and smart grids: A survey. *Renewable and Sustainable Energy Reviews* 30 (2014), 461–478.

John Sipple. 2020. Interpretable, multidimensional, multimodal anomaly detection with negative sampling for detection of device failure. In *International Conference on Machine Learning*. PMLR, 9016–9025.

John Sipple. 2022. From Principles to Prototypes to Products for Intelligent Energy Systems. Keynote Talk. In *International Workshop on Applied Machine Learning for Intelligent Energy Systems (AMLIES) 2022*.

Vaclav Smil. 2010. *Energy transitions: history, requirements, prospects*. ABC-CLIO.

Alexander Spangher. 2018. How Does This Article Make You Feel. *Times Open on Medium* (2018).

Lucas Spangher. 2021. Transactive Multi-Agent Reinforcement Learning for Distributed Energy Price Localization. In *Proceedings of the 8th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation* (Coimbra, Portugal) *(BuildSys '21)*. Association for Computing Machinery, New York, NY, USA, 244–245. https://doi.org/10.1145/3486611.3492387

Lucas Spangher, Utkarsha Agwan, William Arnold, and Tarang Srivastava. 2021. . https://github.com/utkarshapets/microgrid-RL

Lucas Spangher, Akash Gokul, Manan Khattar, Joseph Palakapilly, Utkarsha Agwan, Akaash Tawade, and Costas Spanos. 2020a. Augmenting Reinforcement Learning with a Planning Model for Optimizing Energy Demand Response. In *Proceedings of the 1st International Workshop on Reinforcement Learning for Energy Management in Buildings amp; Cities* (Virtual Event, Japan) *(RLEM'20)*. Association for Computing Machinery, New York, NY, USA, 39–42. https://doi.org/10.1145/3427773.3427863

Lucas Spangher, Akash Gokul, Manan Khattar, Joseph Palakapilly, Utkarsha Agwan, Akaash Tawade, and Costas Spanos. 2020b. Augmenting Reinforcement Learning with a Planning Model for Optimizing Energy Demand Response. In *Proceedings of the 1st International Workshop on Reinforcement Learning for Energy Management in Buildings & Cities*. 39–42.

Lucas Spangher, Akash Gokul, Manan Khattar, Joseph Palakapilly, Akaash Tawade, Adam Bouyamourn, Alex Devonport, and Costas Spanos. 2020c. Prospective experiment for reinforcement learning on demand response in a social game framework. In *Proceedings of the Eleventh ACM International Conference on Future Energy Systems*. 438–444.

Lucas Spangher, Akash Gokul, Joseph Palakapilly, Utkarsha Agwan, Manan Khattar, Wann-Jiun Ma, and Costas Spanos. [n. d.]. OfficeLearn: An OpenAI Gym Environment for Reinforcement Learning on Occupant-Level Building's Energy Demand Response. In *Tackling Climate Change with Artificial Intelligence Workshop at NeurIPS, 2020*.

Lucas Spangher, Akash Gokul, Joseph Palakapilly, Utkarsha Agwan, Manan Khattar, Wann-Jiun Ma, and Costas Spanos. 2020d. OfficeLearn: An OpenAI Gym Environment for Reinforcement Learning on Occupant-Level Building's Energy Demand Response. In *Tackling Climate Change with Artificial Intelligence Workshop at NeurIPS*.

Lucas Spangher, Akaash Tawade, Alex Devonport, and Costas Spanos. 2019a. Engineering vs. Ambient Type Visualizations: Quantifying Effects of Different Data Visualizations on Energy Consumption. In *Proceedings of the 1st ACM International Workshop on Urban Building Energy Sensing, Controls, Big Data Analysis, and Visualization* (New York, NY, USA) *(UrbSys'19)*. Association for Computing Machinery, New York, NY, USA, 14–22.

Lucas Spangher, Akaash Tawade, Alex Devonport, and Costas Spanos. 2019b. Engineering vs. Ambient Type Visualizations: Quantifying Effects of Different Data Visualizations on Energy Consumption. In *Proceedings of the 1st ACM International Workshop on Urban Building Energy Sensing, Controls, Big Data Analysis, and Visualization* (New York, NY, USA) *(UrbSys'19)*. Association for Computing Machinery, New York, NY, USA, 14–22. https://doi.org/10.1145/3363459.3363527

Lucas Spangher, Akaash Tawade, Alex Devonport, and Costas Spanos. 2019c. Engineering vs. Ambient Type Visualizations: Quantifying Effects of Different Data Visualizations on Energy Consumption. In *Proceedings of the 1st ACM International Workshop on Urban Building Energy Sensing, Controls, Big Data Analysis, and Visualization* (New York, NY, USA) *(UrbSys'19)*. Association for Computing Machinery, New York, NY, USA, 14–22. https://doi.org/10.1145/3363459.3363527

Brady Stoll, Elizabeth Buechler, and Elaine Hale. 2017. The value of demand response in Florida. *The Electricity Journal* 30, 9 (2017), 57–64.

Frost Sullivan. [n. d.]. Is the Asia-Pacific Region Demand Response Ready? — frost.com. https://www.frost.com/frost-perspectives/asia-pacific-region-demand-response-ready/. [Accessed 12-Aug-2022].

Peter Sunehag, Guy Lever, Audrunas Gruslys, Wojciech Marian Czarnecki, Vinicius Zambaldi, Max Jaderberg, Marc Lanctot, Nicolas Sonnerat, Joel Z Leibo, Karl Tuyls, et al. 2017. Value-decomposition networks for cooperative multi-agent learning. *arXiv preprint arXiv:1706.05296* (2017).

Richard S Sutton. 1991a. Dyna, an integrated architecture for learning, planning, and reacting. *ACM Sigart Bulletin* 2, 4 (1991), 160–163.

Richard S. Sutton. 1991b. Dyna, an Integrated Architecture for Learning, Planning, and Reacting. *SIGART Bull.* 2, 4 (July 1991), 160–163. https://doi.org/10.1145/122344.122377

Richard S. Sutton. 2022. The Quest for a Common Model of the Intelligent Decision Maker. https://doi.org/10.48550/ARXIV.2202.13252

Richard S Sutton and Andrew G Barto. 2018. *Reinforcement learning: An introduction*. MIT press.

Richard S Sutton, Andrew G Barto, et al. 1999a. Reinforcement learning. *Journal of Cognitive Neuroscience* 11, 1 (1999), 126–134.

Richard S Sutton, Doina Precup, and Satinder Singh. 1999b. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence* 112, 1-2 (1999), 181–211.

Canh T Dinh, Nguyen Tran, and Josh Nguyen. 2020. Personalized federated learning with moreau envelopes. *Advances in Neural Information Processing Systems* 33 (2020), 21394–21405.

Gerald Tesauro. 1994. TD-Gammon, a self-teaching backgammon program, achieves master-level play. *Neural computation* 6, 2 (1994), 215–219.

Zhongda Tian. 2021. A state-of-the-art review on wind power deterministic prediction. *Wind Engineering* 45, 5 (2021), 1374–1392.

Jonas Togler, Fabian Hemmert, and Reto Wettach. 2009. Living Interfaces: The Thrifty Faucet. In *Proceedings of the 3rd International Conference on Tangible and Embedded Interaction* (Cambridge, United Kingdom) *(TEI '09)*. ACM, New York, NY, USA, 43–44. https://doi.org/10.1145/1517664.1517680

Falko Ueckerdt, Lion Hirth, Gunnar Luderer, and Ottmar Edenhofer. 2013. System LCOE: What are the costs of variable renewables? *Energy* 63 (2013), 61–75.

Christopher H Van Arsdale, John Platt, and Mike Tyka. 2022. CO2 capture by pumping surface acidity to the deep ocean. (2022).

Edward Vine and Christopher M Jones. 2015. A review of energy reduction competitions: What have we learned? (2015).

Edward L Vine and Christopher M Jones. 2016. Competition, carbon, and conservation: Assessing the energy savings potential of energy efficiency competitions. *Energy Research & Social Science* 19 (2016), 158–176.

Alexandra Von Meier. 2006. *Electric power systems: a conceptual introduction*. John Wiley & Sons.

Zhiqiang Wan, Hepeng Li, Hang Shuai, Yan Lindsay Sun, and Haibo He. 2021. Adversarial Attack for Deep Reinforcement Learning Based Demand Response. In *2021 IEEE Power & Energy Society General Meeting (PESGM)*. IEEE, 1–5.

Angelina Wang, Thanard Kurutach, Kara Liu, Pieter Abbeel, and Aviv Tamar. 2019a. Learning robotic manipulation through visual planning and acting. *arXiv preprint arXiv:1905.04411* (2019).

Hao Wang and Jianwei Huang. 2016. Incentivizing energy trading for interconnected microgrids. *IEEE Transactions on Smart Grid* 9, 4 (2016), 2647–2657.

Hao Wang, Zakhary Kaplan, Di Niu, and Baochun Li. 2020a. Optimizing federated learning on non-iid data with reinforcement learning. In *IEEE INFOCOM 2020-IEEE Conference on Computer Communications*. IEEE, 1698–1707.

Hongyi Wang, Mikhail Yurochkin, Yuekai Sun, Dimitris Papailiopoulos, and Yasaman Khazaeni. 2020c. Federated Learning with Matched Averaging. `https://doi.org/10.48550/ARXIV.2002.06440`

Xiaofei Wang, Chenyang Wang, Xiuhua Li, Victor CM Leung, and Tarik Taleb. 2020b. Federated deep reinforcement learning for Internet of Things with decentralized cooperative edge caching. *IEEE Internet of Things Journal* 7, 10 (2020), 9441–9455.

Yu Wang, Shiwen Mao, and R Mark Nelms. 2015. On hierarchical power scheduling for the macrogrid and cooperative microgrids. *IEEE Transactions on Industrial Informatics* 11, 6 (2015), 1574–1584.

Yu Wang, Wotao Yin, and Jinshan Zeng. 2019b. Global convergence of ADMM in nonconvex nonsmooth optimization. *Journal of Scientific Computing* 78, 1 (2019), 29–63.

Zhe Wang and Tianzhen Hong. 2020. Reinforcement learning for building controls: The opportunities and challenges. *Applied Energy* 269 (2020), 115036.

Lulu Wen, Kaile Zhou, Jun Li, and Shanyong Wang. 2020. Modified deep learning and reinforcement learning for an incentive-based demand response model. *Energy* 205 (2020), 118019. `https://doi.org/10.1016/j.energy.2020.118019`

Casey Williams. [n. d.]. Energy Humanities. *The Johns Hopkins Guide to Critical Theory and Cultural Studies* ([n. d.]).

Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning* 8, 3 (1992), 229–256.

Damon Wischik, Mark Handley, and Marcelo Bagnulo Braun. 2008. The Resource Pooling Principle. *SIGCOMM Comput. Commun. Rev.* 38, 5 (sep 2008), 47–52. `https://doi.org/10.1145/1452335.1452342`

Alistair J Woodward and Jonathan M Samet. 2018. Climate change, hurricanes, and health. , 33–35 pages.

Chulin Xie, Keli Huang, Pin-Yu Chen, and Bo Li. 2019. Dba: Distributed backdoor attacks against federated learning. In *International Conference on Learning Representations*.

Xing Xu, Rongpeng Li, Zhifeng Zhao, and Honggang Zhang. 2021. The gradient convergence bound of federated multi-agent reinforcement learning with efficient communication. *arXiv preprint arXiv:2103.13026* (2021).

Bo Yang, Selvaprabu Nadarajah, and Nicola Secomandi. 2021. Least Squares Monte Carlo and Pathwise Optimization for Merchant Energy Production. *Available at SSRN 3900797* (2021).

Chi-Jen Yang and Robert B Jackson. 2011. Opportunities and barriers to pumped-hydro energy storage in the United States. *Renewable and Sustainable Energy Reviews* 15, 1 (2011), 839–844.

Ji Hoon Yoon, Ross Baldick, and Atila Novoselac. 2014a. Dynamic demand response controller based on real-time retail price for residential buildings. *IEEE Transactions on Smart Grid* 5, 1 (2014), 121–129.

Ji Hoon Yoon, Ross Bladick, and Atila Novoselac. 2014b. Demand response for residential buildings based on dynamic price of electricity. *Energy and Buildings* 80 (2014), 531–541.

Woo-Seung Yun, Won-Hwa Hong, and Hyuncheol Seo. 2021. A data-driven fault detection and diagnosis scheme for air handling units in building HVAC systems considering undefined states. *Journal of Building Engineering* 35 (2021), 102111.

Kris Zacny, Marc M Cohen, Warren W James, and Brent Hilscher. 2013. Asteroid mining. In *AIAA Space 2013 Conference and Exposition*. 5304.

David D Zhang, Peter Brecke, Harry F Lee, Yuan-Qing He, and Jane Zhang. 2007. Global climate change, war, and population decline in recent human history. *Proceedings of the National Academy of Sciences* 104, 49 (2007), 19214–19219.

Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. 2021a. Decentralized multi-agent reinforcement learning with networked agents: Recent advances. *Frontiers of Information Technology & Electronic Engineering* 22, 6 (2021), 802–814.

Kaiqing Zhang, Zhuoran Yang, Han Liu, Tong Zhang, and Tamer Basar. 2018. Fully decentralized multi-agent reinforcement learning with networked agents. In *International Conference on Machine Learning*. PMLR, 5872–5881.

Michael Zhang, Karan Sapra, Sanja Fidler, Serena Yeung, and Jose M. Alvarez. 2020. Personalized Federated Learning with First Order Model Optimization. `https://doi.org/10.48550/ARXIV.2012.08565`

Weiting Zhang, Dong Yang, Wen Wu, Haixia Peng, Ning Zhang, Hongke Zhang, and Xuemin Shen. 2021b. Optimizing federated learning in distributed industrial IoT: A multi-agent approach. *IEEE Journal on Selected Areas in Communications* 39, 12 (2021), 3688–3703.

Wenshuai Zhao, Jorge Peña Queralta, and Tomi Westerlund. 2020. Sim-to-real transfer in deep reinforcement learning for robotics: a survey. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, 737–744.

Peng Zhou, RY Jin, and LW Fan. 2016. Reliability and economic evaluation of power system with renewables: A review. *Renewable and Sustainable Energy Reviews* 58 (2016), 537–547.