

A generic self-learning emotional framework for Reinforcement Learning agents



**UNIVERSIDAD
DE GRANADA**

DOCTORAL THESIS

Submitted for the degree of Philosophiae Doctor (PhD) by

Alberto Hernández Marcos

Doctoral Programme in Information and Communication Technologies
Department of Computer Engineering, Automation, and Robotics (ICAR)
University of Granada

Supervisor
Dr. Eduardo Ros Vidal

Granada, 2024

A generic self-learning emotional framework for Reinforcement Learning agents



**UNIVERSIDAD
DE GRANADA**

DOCTORAL THESIS

Submitted for the degree of Philosophiae Doctor (PhD) by

Alberto Hernández Marcos

Doctoral Programme in Information and Communication Technologies
Department of Computer Engineering, Automation, and Robotics (ICAR)
University of Granada

Supervisor
Dr. Eduardo Ros Vidal

Granada, 2024

Editor: Universidad de Granada. Tesis Doctorales
Autor: Alberto Hernández Marcos
ISBN: 978-84-1195-778-6
URI: <https://hdl.handle.net/10481/103595>

Title: A generic self-learning emotional framework for Reinforcement Learning agents
Author: Alberto Hernández Marcos
Doctoral Programme in Information and Communication Technologies
School Of Technology And Telecommunications Engineering (ETSIIT)
University of Granada

Thesis Supervision:
Dr. Eduardo Ros Vidal, Research Centre for Information and Communications Technologies
(CITIC-UGR) - Department of Computer Engineering, Automation, and Robotics (ICAR),
University of Granada, Spain

Funding:
Contribution of Eduardo Ros partially funded by Grant PID2022-140095NB-I00 funded by
MCIN/AEI /10.13039/501100011033/ and FEDER Una manera de hacer Europa.

*The question is not whether intelligent machines can have any emotions,
but whether machines can be intelligent without any emotions.*

Marvin Minsky, 1986

Acknowledgements

Reaching the end of the challenging journey I formally embarked on in 2021, from a collection of unfinished notes and questions I wasn't sure I could answer, would not have been possible without the selfless help of many dear people. I am deeply grateful to Eduardo Ros Vidal, the supervisor of this thesis, for his dedicated assistance in focusing and addressing the idea and formal validation methods, but no less for his constant advice and availability throughout these three years. I am also indebted to Antonio Cándido Ortiz and Andrés Catena Martínez, from the University of Granada, for their guidance and collaboration in conducting the emotional attribution survey, one of the outcomes from this research that merits a deeper, separate continuation.

The initial push toward this academic journey is owed to the greatness of several lifelong colleagues and friends: Pascual de Juan Núñez, whose enthusiasm, example, and unwavering confidence in me inspired me to pursue this unexpected venture; David Suárez Caro and Raimundo Alegría Llorente, experts in the field and the first to review and validate the initial raw idea; and Manuel Gómez Olmedo, from the University of Granada, whose doors I would never have known how to find without his guidance and encouragement.

But none of this would have been possible without the unconditional and daily faith and support of my Katrin, Laura, Marcos, and Nicolás; my parents Alberto and Carmina; and my German parents Alfons and Renate, none of whom it ever occurred to that I couldn't achieve whatever I set my mind to.

Agradecimientos (Spanish)

Llegar al final de un camino tan complicado como el que inicié formalmente en 2021, desde un montón de notas inacabadas y preguntas que no estaba seguro de poder responder, no habría sido posible sin la desinteresada ayuda de muchas personas muy cercanas. Agradezco profundamente a Eduardo Ros Vidal, supervisor de esta tesis, su atenta ayuda para enfocar y abordar la idea, así como los métodos formales de validación, pero no menos por su constante asesoramiento y disponibilidad durante estos tres años. Estoy asimismo en deuda con Antonio Cándido Ortiz y Andrés Catena Martínez, de la Universidad de Granada, por su orientación y colaboración en la ejecución de la encuesta de atribución emocional, uno de los resultados de esta investigación que amerita una continuación por separado más profunda.

El impulso inicial hacia esta aventura académica lo debo a la grandeza de varios compañeros y amigos de por vida, como lo son Pascual de Juan Núñez, cuyo entusiasmo, ejemplo y absoluta confianza en mí me inspiraron a iniciar este inopinado camino; David Suárez Caro y Raimundo Alegría Llorente, expertos en la materia y primeros en revisar y validar la idea inicial en crudo; y Manuel Gómez Olmedo, de la propia Universidad de Granada, cuyas puertas nunca habría sabido encontrar sin su orientación y aliento.

Pero nada de esto me habría sido posible sin el apoyo y fe incondicional y diaria de mis Katrin, Laura, Marcos, y Nicolás; mis padres Alberto y Carmina; y mis padres alemanes Alfons y Renate, a quienes jamás se les ocurrió que no pudiera alcanzar lo que me propusiera.

Abstract

In nature, intelligent living beings have developed emotions to modulate their behavior as a fundamental evolutionary advantage. However, researchers seeking to endow machines with this advantage lack a clear theory from cognitive neuroscience describing emotional elicitation from first principles, namely, from raw observations to concrete affects. As a result, they often resort to case-specific solutions and arbitrary or hard-coded models that fail to generalize well to other agents and tasks.

Here we propose that emotions correspond to distinct temporal patterns perceived in crucial values for living beings in their environment (like recent rewards, expected future rewards or anticipated world states). Based on this foundation, we introduce a fully self-learning emotional framework for Artificial Intelligence agents convincingly associating said patterns to natural emotions documented in the scientific literature.

Applied in a case study, an artificial neural network trained on unlabeled agent’s experiences successfully learned and identified eight basic emotional patterns that are situationally coherent and reproduce natural emotional dynamics. Validation through an emotional attribution study, where human observers rated their pleasure-arousal-dominance dimensions, showed high statistical agreement, distinguishability, and strong alignment with experimental psychology accounts, demonstrating that human observers can infer the internal “emotional state” of artificial agents, aligning closely with the agents’ internal status variables.

We believe that the generality of the framework and the cross-disciplinary language defined, grounded on first principles from Reinforcement Learning, may lay the foundations for further research and applications, leading us toward emotional machines that think and act more like us.

Resumen

En la naturaleza, los seres vivos dotados de inteligencia han desarrollado emociones para modular su comportamiento como una ventaja evolutiva fundamental. Sin embargo, en su aspiración de dotar a las máquinas de esta ventaja, los investigadores carecen de una teoría clara en la neurociencia cognitiva que describa la elicitación emocional desde principios fundamentales, es decir, desde observaciones crudas hasta afectos concretos. Como resultado, se recurre a menudo a soluciones específicas para cada caso y a modelos arbitrarios o codificados de manera fija que no generalizan bien a otros agentes y tareas.

El presente estudio propone la hipótesis de que las emociones corresponden a distintos patrones temporales percibidos en valores cruciales para los seres vivos en su entorno (como recompensas recientes, recompensas futuras esperadas o estados anticipados del mundo). Sobre esta base, se introduce un marco emocional completamente autoaprendido para agentes con Inteligencia Artificial que asocia convincentemente dichos patrones con las emociones naturales documentadas en la literatura científica.

Aplicado a un caso práctico, una red neuronal artificial entrenada con experiencias no etiquetadas del agente identificó con éxito ocho patrones emocionales básicos que son coherentes situacionalmente y reproducen dinámicas emocionales naturales. La validación mediante un estudio de atribución emocional, en el que observadores humanos evaluaron sus dimensiones de placer-activación-dominancia, demostró un alto nivel de acuerdo estadístico, distinguibilidad y un fuerte alineamiento con los registros de la psicología experimental, demostrando que los observadores humanos pueden inferir el “estado emocional” interno de agentes artificiales, alineándose estrechamente con las variables internas de estado del agente.

Creemos que la generalidad de este marco de trabajo y el lenguaje interdisciplinario que define, basado en principios fundamentales del Aprendizaje por Refuerzo, pueden sentar las bases para futuras investigaciones y aplicaciones que conduzcan hacia máquinas emocionales que piensen y actúen más como nosotros.

Nota: Un resumen más amplio en español de este trabajo se encuentra al final del documento, como Resumen de la tesis.

Table of Contents

Acknowledgements	v
Abstract	vi
Resumen	vii
List of Figures	xii
List of Tables	xiv
Abbreviations and acronyms	xvi
1 Introduction	1
1.1 Emotions in Nature: Biology, Neuroscience and Psychology	1
1.1.1 Origins of the concept	1
1.1.2 Overview of the main emotion theories across fields	2
1.1.3 The biological perspective	3
1.1.4 The psychological labyrinth	4
Embodiment Theories	5
Discrete theories	6
Dimensional or Continuous theories	6
Appraisal theories	8
Motivational theories	9
Constructivist theories	9
Emotional Intelligence	10
1.1.5 The neuroscience of emotions	10
1.2 Emotions in AI: Related work in Affective Computing and Reinforcement Learning	11
1.2.1 Artificial intelligence is born without artificial emotion	12
1.2.2 Early skepticism about emotion in machines	13
1.2.3 Increasing interest but little material progress	14
1.2.4 The birth of Affective Computing and related fields	16
1.2.5 Non-learned models of emotion synthesis: State-of-the art and limitations	20
Minsky and the “emotion machine”	20
The OCC model: A computational framework for emotion synthesis .	21
The TDRL theory of emotion and related approaches	22
1.2.6 Efforts toward self-learned emotion	24
The IBIA architecture	24
Emotions and the Free-energy principle	25

	Learning intrinsic emotion-based rewards	26
	Emotion-driven robotic path planning	27
1.3	Concluding remarks	27
1.4	Overview of this work	28
2	Objectives	29
2.1	Overall objectives and scope of this research	29
2.1.1	Hypotheses	29
2.1.2	General objectives	29
2.1.3	Specific objectives	30
2.2	Possible future objectives	31
3	A generic self-learning emotional framework	33
3.1	Inspirational background	33
3.2	Introduction to the proposed framework	34
3.3	Theoretical framework	38
3.3.1	Foundational concepts from Reinforcement Learning	38
3.3.2	Primary definitions for the framework	39
3.4	Methodology	42
3.4.1	Learning emotions from experience	42
	Training of an emotional encoder	42
	Input values	43
	Model architecture	44
3.4.2	Elicitation of emotions and integration within an RL architecture . .	44
3.4.3	Interpretation of the learned emotions	45
	Clustering of the emotional spectrum	45
	Selection and validation of the interpretability mapping	45
	Attribution of emotion terms	46
	Real-time emotional interpretation	47
3.4.4	Extensions of the actor-critic method	48
	Training of a classic actor-critic agent (Order 0 - Non-emotional agent)	48
	Training of an emotional agent (Order III - Anticipation)	49
	Training of an emotional agent (Order IV - World-knowledgeability) .	49
4	Results	53
4.1	Application of the framework on a practical case study	53
4.1.1	Learning emotions from experience	55
	Pre-training of a conventional RL agent	55
	Dataset generation. Selection of input values and emotional window .	55
	Training of the emotional model from dataset sequences	55
4.1.2	Elicitation of emotions	56
4.1.3	Interpretation of the learned emotions	56
	Clustering of the emotional spectrum	56
	Selection and validation of the interpretability mapping	56
	Attribution of emotion terms	57

Visualization	59
4.2 Experimental validation of the learned emotions	62
4.2.1 Emotional attribution test with humans	62
4.2.2 Mapping versus documented experimental accounts	65
5 Methods and tools	69
5.1 Application of the framework on a practical case study	69
5.1.1 Learning emotions from experience	69
Pre-training of a conventional RL agent	69
Dataset generation. Selection of input values and emotional window .	70
Training of the emotional model from dataset sequences	71
5.1.2 Elicitation of emotions	72
5.1.3 Interpretation of the learned emotions	72
Clustering of the emotional spectrum	72
Selection and validation of the interpretability mapping	74
Attribution of emotion terms	74
Visualization	77
5.2 Theoretical validation of LOVE profile terms	77
5.3 Experimental validation of learned emotions with humans	78
5.3.1 Emotional attribution test with humans	78
Dataset	78
Tests with Lang’s SAM manikin	78
Statistical significance	81
Test reliability	81
PAD values attributed to the 48 videos	81
PAD values attributed to the eight learned emotions	81
Distinguishability of the learned emotions	81
5.3.2 Mapping versus documented experimental accounts	81
6 Conclusions	85
6.1 Discussion	85
6.1.1 Key achievements	85
6.1.2 Limitations and future work	87
6.1.3 Ethical considerations	88
6.2 Final thoughts	88
A Extended data	89
A.1 Code and data	89
A.2 Application of the framework on a practical case study: Extended data . . .	90
A.2.1 Alternative conventional RL agents trained	90
A.2.2 Alternative clustering results	92
A.2.3 3D visualization of the emotional spectrum	93
A.3 Theoretical Validation of LOVE Profile Terms: Extended data	94
A.4 Experimental validation of learned emotions with humans: Extended data .	100
A.4.1 Likert scales used for emotion attribution in Spanish	100

A.4.2	PAD values attributed to the 48 videos	102
B	Research publications	103
B.1	Journal papers	103
B.2	Other dissemination activities	104
C	Resumen de la tesis	105
	References	113

List of Figures

1.1	Illustrations from Darwin’s <i>The Expression of the Emotions in Man and Animals</i>	4
1.2	Ekman’s six basic emotions	6
1.3	Plutchik’s “wheel of emotions”	7
1.4	Russell’s circumplex model	8
1.5	The Woggles in <i>Edge of Intention</i>	16
1.6	Electroencephalograma (EEG) in Affective Computing	17
1.7	<i>Kismet</i> (1998)	18
1.8	<i>Pepper</i> (2014)	18
1.9	<i>Sophia</i> (2016)	19
1.10	World Robot Conference (2024)	19
1.11	The OCC model	21
3.1	Overview of the framework: How emotions can be learned from experiences, then elicited and interpreted	35
3.2	Elicitation of learned emotions	37
3.3	Interpretation of the learned emotions	37
3.4	How cognitive abilities determine the emotional spectrum	41
4.1	The environment LunarLander-v2 used in the case study	53
4.2	Learning emotions from experiences in a practical case study	54
4.3	Trajectory of a full episode run with the trained agent	55
4.4	The eight classes identified by the emotional interpreter.	56
4.5	The eight emotional patterns learned	57
4.6	Order III - LOVE 2:5x6 Interpretability mapping	57
4.7	Interpretation of the learned emotions	58
4.8	Emotional transition matrix and graph	59
4.9	Interpretation of the instantaneous emotion	60
4.10	Interpreting live emotions during a successful landing	60
4.11	Interpreting live emotions during a failed landing	61
4.12	Spontaneous emergence of documented emotion dimensions	62
4.13	Results of the emotional attribution study	63
4.14	Average Pleasure value attributed to sequences of each emotion	64
5.1	Training of the conventional RL agent chosen	70
5.2	All the values registered during one trajectory	71

5.3	Comparison of different encoding dimensions	72
5.4	Test sequences reproduced by the autoencoder.	72
5.5	The custom-made tool “ <i>RL Emotion Lab</i> ”	73
5.6	The eight emotional patterns learned	74
5.7	Information in a frame from one of the 48 rated sequences	79
5.8	SAM, the classic graphic layout used for the rating of sequences	80
A.1	Quick initial exploration of hyperparameters	90
A.2	Alternative architectures tested	91
A.3	Tests with several random seeds on the final architecture	91
A.4	Alternative clustering options evaluated	92
A.5	3D representation of the eight identified classes.	93
A.6	Sequence coherence test: Series 1.1.	95
A.7	Sequence coherence test: Series 1.2.	96
A.8	Sequence coherence test: Series 2.1.	97
A.9	Sequence coherence test: Series 3.1.	98
A.10	Sequence coherence test: Series 3.2.	98
A.11	Sequence coherence test: Series 3.3.	99
A.12	Likert scale for the “Pleasure” dimension.	100
A.13	Likert scale for the “Activation” dimension.	100
A.14	Likert scale for the “Dominance” dimension.	101

List of Tables

1.1	Classification of emotion theories according to the aspects of emotions they focus on (adapted from Zennaro, 2013).	3
3.1	Association of some emotions to latest observed values.	36
4.1	PAD rates of each emotion in range [1, 9] aggregated over all raters from their six corresponding videos.	64
4.2	Correlation among Pleasure / Arousal / Dominance across all the videos (Pearson Two-sided).	65
4.3	Correlation among Pleasure / Arousal / Dominance across the eight emotions (Pearson Two-sided).	65
4.4	ICC2k Intraclass Correlation Coefficients obtained from the study.	65
4.5	p-value of the Hotelling's T-squared statistical test for all emotion pairs. . .	66
4.6	Results from mapping PAD values from the survey to five documented experimental accounts.	67
4.7	Top pleasure-arousal-dominance (PAD) matches across authors for each learned emotion.	67
5.1	Global statistics of all the original MTS trajectories by 20-step sequences. . .	75
5.2	Recent-trend classification criterion of a variable's slope.	75
5.3	Recent-trend classification criterion of a variable's mean value.	75
5.4	Local statistics and classification of the eight individual average sequences for <i>Reward</i>	76
5.5	Local statistics and classification of the eight individual average sequences for <i>State-value</i>	76
5.6	Final emotion term attribution to learned clusters.	76
A.1	PAD rates for each video in range [1, 9] aggregated over all raters.	102

Abbreviations and acronyms

AAAI	Association for the Advancement of Artificial Intelligence
AI	Artificial Intelligence
ANN	Artificial Neural Network
BCE	Before Current Era
BIC	Bayes Information Criterion
DAE	Deep Autoencoder
EEG	Electroencephalography
EI	Emotional Intelligence
fMRI	Functional Magnetic Resonance Imaging
GMM	Gaussian Mixture Model
IBIA	Integrated Biologically-Inspired Architecture
ICC2k	Intraclass Correlation Coefficient, two-way random effects model
KL	Kullback-Leibler divergence
LIDAR	Light Detection and Ranging
LOVE	Latest Observed Values Encoding
ML	Machine Learning
MTS	Multivariate Time Series
OCC	Ortony, Collins, and Clore model
PAD	Pleasure-Arousal-Dominance
PET	Positron Emission Tomography
PPO	Proximal Policy Optimization
ReLU	Rectified Linear Unit
RL	Reinforcement Learning
RMSE	Root Mean Squared Error
SAM	Lang’s Self-Assessment Manikin
SARSA	State-Action-Reward-State-Action
tanh	Hyperbolic tangent
TD	Temporal Difference
TDRL	Temporal Difference Reinforcement Learning model
t-SNE	T-distributed Stochastic Neighbor Embedding
XCS	eXtended Classifier System

Chapter 1

Introduction

1.1 Emotions in Nature: Biology, Neuroscience and Psychology

There is no fundamental difference between man and animals in their ability to feel pleasure and pain, happiness, and misery.

Charles Darwin, *The Expression of the Emotions in Man and Animals* (1872)

It is difficult to overemphasize the significance of emotions in nature. Indeed, growing evidence indicates that many living beings, particularly those displaying intelligent behaviors, have evolved the capacity for emotions of varying complexity as an intrinsic and essential constituent of their mental processes [1][2]. The reason is well supported by established disciplines such as neuroscience [3][4][5], psychology [6][7][8], and biology [1][9]: contrary to western philosophical tradition, which views them as detached from reason and hindering rational thought, emotions—alongside related phenomena like affects, feelings or sentiments—are currently understood as an evolutionary advantage, as first-order psychodynamic forces enhancing an organism’s adaptability to its environment, supporting the ultimate goal of survival. They play a crucial role in learning and social behavior, and evidence suggests that damages to emotional components of the brain severely impairs decision making [3].

1.1.1 Origins of the concept

The very etymology of the word *emotion*, from the Latin *emotio* (*e-* (out) and *movere* (to move), meaning “a movement or agitation”), reflects an early understanding of emotions as dynamic forces within our inner selves, driving thought and behavior.

Among the earliest comprehensive accounts, Aristotle’s *Rhetoric* (around 350 BCE) examines emotions like anger, calmness, fear, shame, pity, and envy—to which he often referred to with the term *pathē* (πάθη), meaning “passions” or “affections”—, detailing their causes and

effects. He emphasizes their role as powerful forces influencing persuasion and decision-making, describing them as *movements* of the soul, associated with pleasure and pain, which aligns with the later concept of *emotio* as outward-moving forces.

But interest in emotions appeared early across various cultures. For example, the *Li Chi* (Book of Rites), compiled in the first century BCE in China, enumerates emotions such as joy, anger, sadness, fear, love, disliking, and liking as innate human feelings. Similarly, ancient Indian texts like the *Natyashastra* explore emotions in the context of drama and aesthetics. A more comprehensive historical account falls beyond the scope of this work; for further exploration, see Lindholm’s *An Anthropology of Emotion* [10].

One of the most influential voices in the early study of emotions, René Descartes, introduced the French term *émotion* as an alternative to *passion*, whose principal effect was to “*move and dispose the soul to want the things for which they prepare the body.*” In his work *The Passions of the Soul* (1649), he defined them as perceptions or disturbances of the soul, identifying six primary passions: wonder, love, hatred, desire, joy, and sadness [11]. Descartes viewed these as fundamental emotions that shape human experience, emphasizing their role in the mind-body connection.

Building on Descartes’ foundational work, David Hume emphasized the significance of emotions in human reasoning, suggesting that our passions drive moral judgments: “*Reason alone can never be a motive to any action of the will*” [12]. Immanuel Kant further differentiated emotions from passions, focusing on their impact on rational decision-making [13]. Later, William James proposed that emotions are deeply connected to physiological responses [14], laying the groundwork for modern psychological theories.

1.1.2 Overview of the main emotion theories across fields

These philosophical perspectives set the stage for various influential theories that attempt to describe emotional phenomena. Despite the absence of a unified model broadly accepted to clearly and comprehensively describe emotions, numerous theories have been developed across fields, each capturing significant aspects of emotions and garnering considerable support.

Among the numerous attempts to organize and compare these heterogeneous theories, we follow here a taxonomy suggested by Fabio Massimo Zennaro in his *Theories of Emotion*, probably one of the most lucid and clarifying classifications [15]. In it, theories are classified by the different aspects of emotions they focus on:

- *Anthropological origin of emotions*: How emotions evolved in humans, their role in the development of the human species, and their impact on survival.
- *Actual development of emotions*: How emotions are experienced in everyday life, what elicits them, and how they unfold over time.
- *Characterization and categorization of emotions*: How different emotions can be described, differentiated, and classified based on their common and peculiar traits.
- *Behavioral and cognitive effects of emotions*: How emotions influence human behavior and cognitive processes, and what the consequences of these emotional experiences are.

Table 1.1 shows how different types of theories from the fields of biology, neuroscience, and psychology consider these four aspects. Adaptive theories, rooted in biology, explore the evolutionary aspects of emotions, emphasizing their role in enhancing biological fitness. Neuroscientific theories aim to explain emotions by identifying the brain circuits and neural patterns underlying emotional responses, thereby providing a bridge between physiological processes and emotional experiences. In contrast, psychology encompasses a broader range of theories, including embodiment, discrete, dimensional, appraisal, motivational, and constructivist theories, each focusing on different aspects and processes of emotional experiences.

Table 1.1: Classification of emotion theories according to the aspects of emotions they focus on (adapted from Zennaro, 2013).

Field	Theory type	Anthropological origins	Actual development	Characterization and categorization	Behavioural and cognitive effects
Biology	<i>Adaptive theories</i>	✓			
Neuroscience	<i>Neuroscientific theories</i>		✓		
Psychology	<i>Embodiment theories</i>		✓	✓	
	<i>Discrete theories</i>			✓	
	<i>Dimensional theories</i>			✓	
	<i>Appraisal theories</i>		✓	✓	
	<i>Motivational theories</i>			✓	✓
	<i>Constructivist theories</i>	✓	✓		

In the following sections, we will delve deeper into these fields, examining the principal theories and their contributions to our understanding of emotions.

1.1.3 The biological perspective

Pre-empting modern psychology by a century, Charles Darwin’s exploration of emotions in a biological context marks a pivotal moment in the understanding of emotional expressions. In his seminal work *The Expression of the Emotions in Man and Animals* (1872) [1], he introduced and developed the idea that emotions have evolutionary significance, linking human and animal emotional expressions forever with painstakingly detailed examples.

Darwin wrote, “*The community of certain expressions in distinct though allied species, as in the movements of the same facial muscles during laughter by man and by various monkeys, is rendered somewhat more intelligible, if we believe in their descent from a common progenitor.*”

Without ever departing from his all-encompassing conception of natural species, he acutely described thirty six human emotions, or “*expressions exhibited by Man under various states of the mind,*” which he classified into eight main groups. Originally conceived as part of *On the Origin of Species*, Darwin’s insights in this groundbreaking work laid the foundation for

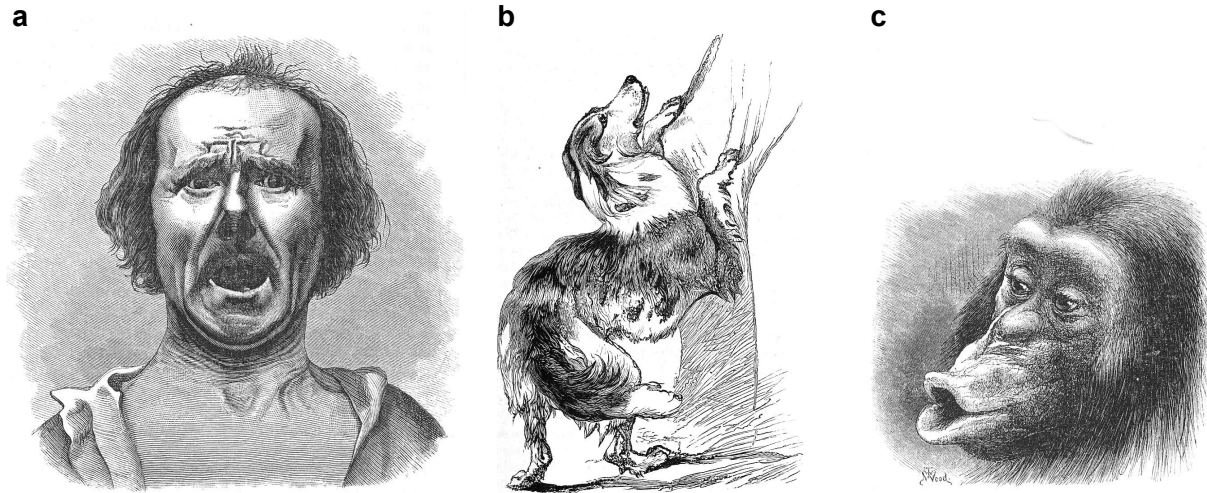


Figure 1.1: Illustrations from Darwin’s *The Expression of the Emotions in Man and Animals*. Darwin’s seminal studies on emotions (1872) anticipated foundational ideas of modern psychology, biology and neuroscience. **a**, Human in horror. **b**, Dog in an affectionate frame of mind. **c**, Chimpanzee disappointed and sulky.

understanding emotions as biological and adaptive phenomena, significantly influencing later research in psychology and ethology.

The evolutionary value of emotions is no longer questioned within the field. As Tooby and Cosmides summarize [9]: “*Within an evolutionary framework, it is assumed that emotions enable individuals to meet specific problems and opportunities that increase their chances of physical survival, reproduction, and gene replication.*”

Building on Darwin’s foundation, contemporary research continues to explore the biological underpinnings of emotions. For example, Marc Bekoff, in his article *Animal Emotions: Exploring Passionate Natures* [2], provides compelling evidence that many animals experience emotions such as joy, fear, love, despair, and grief, demonstrating that humans are not alone in their emotional experiences.

1.1.4 The psychological labyrinth

In contrast to the biological perspectives, the field of psychology presents a more convoluted path, marked by very diverse and often discordant attempts to define and capture the nature of emotions. As we will show, minimal consensus exists on numerous fundamental aspects, including what constitutes an emotion, its types or categories, and even the basic terminology—emotion, affect, sentiment, or feeling [16]. Disagreements extend to the nature of emotions (discrete [6][17], dimensional [7], or mixed [18]), their triggers and physiological responses [8], and whether they are universal [6][5] or culturally dependent [19].

However, from the agitations of this historical division, a number of influential theories have successfully emerged over time, contributing partial but meaningful perspectives to our understanding of the emotional phenomena. The following represent the most influential and widely referenced ones:

- *Embodiment theories* (late 19th century): Emotions as the sum of physiological changes inside the body, turning upside-down the traditional conception that emotions cause these changes.
- *Discrete theories* (mid-20th century): Emotions as specific instances or combinations of a few universal core emotions, within a finite and limited dictionary of basic kinds.
- *Dimensional or Continuous theories* (mid-20th century): Emotions as points in an n-dimensional space of certain continuous dimensions.
- *Appraisal or Cognitivist theories* (late 20th century): Emotions as the result of an unconscious cognitive evaluation of an event.
- *Motivational theories* (late 20th century): Emotions as related to the goals of a subject, whose behavior is determined and explained by them.
- *Constructivist theories* (21st century): Emotions as relative socio-cultural artifacts shaped by social and cultural processes.

Without attempting a comprehensive comparison of this complex landscape, we briefly describe below the key contributions of each.

Embodiment Theories

The divergences in the field can be traced back to its very origins, as evidenced by one of the earliest theories by William James, the founder of the “embodiment theories” [20]. His hypothesis contradicts the natural conception that the perception of a fact excites mental affection (emotion), which in turn causes bodily reactions. James argues the opposite: *“My thesis on the contrary is that the bodily changes follow directly the PERCEPTION of the exciting fact, and that our feeling of the same changes as they occur IS the emotion.”* (Emphasis in the original text.)

This theory was received with widespread disbelief (despite its otherwise correct association between physiological changes and emotions). Worse yet, far from eventually converging toward a general consensus, these early disagreements extended over time. As Scherer laments: *“A particularly unfortunate example is William James’s asking the question ‘What is an emotion?’ when he really meant ‘feeling’, a misnomer that started a debate which is still ongoing, more than a century later”* [21].

Indeed, disagreement rapidly extended to an increasing list of matters, from the most basic concepts to more intricate psychological dynamics, leading to what Ross Buck described as *“the conceptual and definitional chaos that characterizes this area of research”* [22].

Consequently, statements of pessimism and skepticism frequently arise concerning the very definability or soundness of its core concept. DeLancey comes to fear that *“there probably is no scientifically appropriate class of things referred to by our term emotion. Such disparate phenomena—fear, guilt, shame, melancholy, and so on—are grouped under this term that it is dubious that they share anything but a family resemblance”* [23]. As ironically summarized by observers from the specialized media, *“the only thing certain in the emotion field is that no one agrees on how to define emotion”* [24].

Discrete theories

A much higher degree of support has been garnered by the so-called discrete theories of emotions. One of the most influential theories within the field is due to Paul Ekman, who posits that several basic emotions, such as happiness, sadness, anger, fear, disgust and surprise (Fig. 1.2), clearly differing from one another, are biologically universal to all humans, having evolved from their adaptive value in dealing with “*fundamental life tasks*” [25]. Ekman suggests though that these basic emotions combine to form more complex or compound emotions; for example, smugness (Spanish: “petulancia”) would be a combination of happiness and contempt.

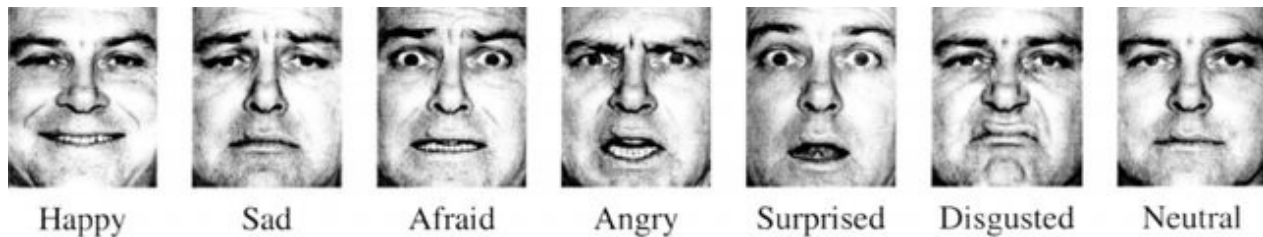


Figure 1.2: Ekman’s six basic emotions. Prototypical facial expression of Ekman’s six basic emotions and a neutral face for actor “J. J.”.

Unsurprisingly though, the list of emotions considered “basic” differs sensibly across researchers. For example, Ekman’s original framework from 1969, “*most influenced by Darwin and Tomkins,*” discarded two of Tomkins’ original basic “affects” (shame and dismissal) [26][27], and substituted enjoyment with happiness [6]. Later on, he discarded interest as well in 1971 (and regularly introduced changes to his own initial lists over different publications [28][29]).

Another reference along the lines of these psychoevolutionary perspectives is Carroll Izard, whose list emotions include many from Tomkins’ and Ekman’s, but substitutes sadness with distress, and differentiates two levels of intensity to each (for example, distress / anguish, anger / rage, etc.) [17].

As a final example of the diversity of approaches across discrete theories, Robert Plutchik proposed his own eight primary emotions in his “wheel of emotions” (Fig. 1.3), conceptualized in terms of pairs of polar opposites: joy / sadness, anger / fear, trust / disgust and surprise / anticipation [18]. However, his model also introduces the idea of different intensities and complexities to his basic list, and explores how these emotions combine to form complex feelings. For this reason, his model is often viewed as a hybrid or “mixed” model, combining elements of both discrete and dimensional theories, which we discuss next.

Dimensional or Continuous theories

In apparent opposition with the discrete theories, dimensional (or continuous) theories consider emotions to consist of continuous dimensions, corresponding to fundamental properties and qualities of human emotions. Thus, the emotion at a certain moment would correspond to a point in the n -dimensional space defined by the theory, which generally includes a neutral state to reflect situations where no emotion prevails.

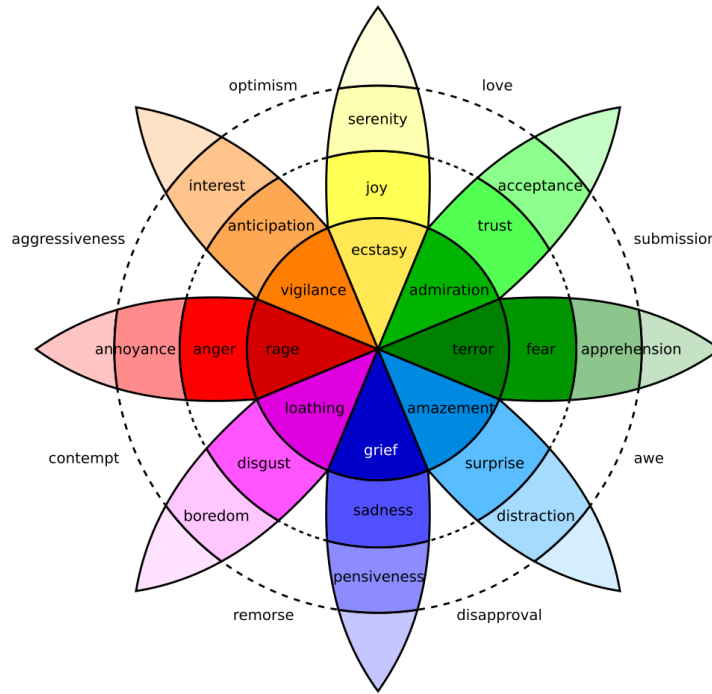


Figure 1.3: Plutchik’s “wheel of emotions”. Often referred to as a “mixed model”, primary emotions are defined for different intensities (inner = more intense / basic; outer = less intense / complex), and can mix into different, more complex emotions.

Their foundational concept can be traced back to Wilhelm Max Wundt [30], whose deeply influential research proposed three dimensions, spanning between *pleasurable* versus *unpleasurable* (currently known as hedonic valence), *arousing* versus *subduing* (or physiological activation value), and *strain* versus *relaxation* (as a temporal aspect of the emotional event).

The most prominent voice in this subfield though is James A. Russell, whose work evolved significantly from an early three-dimensional model to his later, widely recognized two-dimensional circumplex model of affect. In his early work with Albert Mehrabian, Russell proposed a three-factor theory of emotions including three independent and bipolar dimensions [31], very similar to Wundt’s:

1. Pleasure-Displeasure: Reflecting the valence or positivity/negativity of the emotion.
2. Degree of Arousal: Indicating the level of activation or energy.
3. Dominance-Submissiveness: Representing the degree of perceived control.

The introduction of dominance was justified through two studies providing evidence that “*three independent and bipolar dimensions, pleasure-displeasure, degree of arousal, and dominance-submissiveness, are both necessary and sufficient to adequately define emotional states.*”

However, while the first two dimensions gained and maintained general support over time, counterarguments have been historically raised against dominance, considered a weak predictor of the variance of affective judgments (by Bradley, Lang, and Cuthbert [32][33]), and strongly

correlated with valence (by Warriner, Kuperman, and Brysbaert [34]). As a matter of fact, James Russell himself eventually eliminated it from his model, defining the circumplex model of affect [7], a robust framework that continues to inform studies on emotional experience and expression (Fig. 1.4).

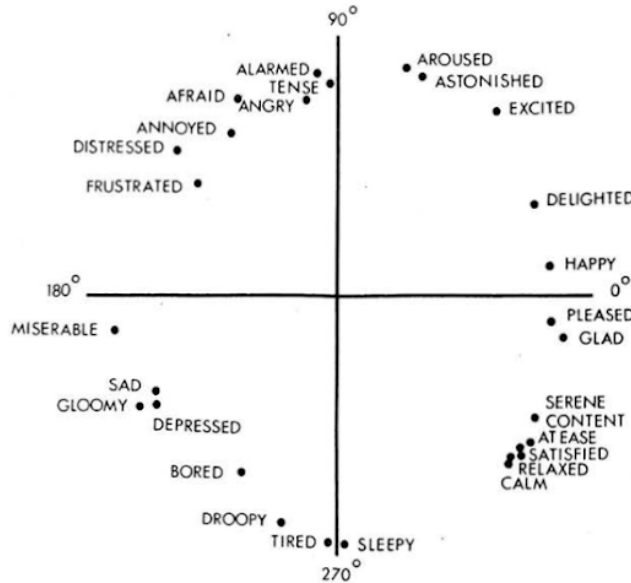


Figure 1.4: Russell’s circumplex model. Direct circular scaling coordinates obtained for 28 affect words, according to the circumplex model of affect.

Appraisal theories

A different perspective to the emotional phenomena focuses on their actual development, elicitation and temporal dynamics. Appraisal theories propose that emotions result from people’s interpretation and explanations of their circumstances, even in the absence of physiological arousal [35]. According to this perspective, emotions complement perceptual information, coloring our sensations and guiding our decision-making [36].

Two prominent approaches within appraisal theories are the *structural model* and the *process model*. Richard Lazarus’ structural model [8], for instance, emphasizes that different emotions are elicited based on the relational, motivational, and cognitive aspects of an individual’s appraisal of a situation. The relational aspect concerns the relationship between a person and their environment. The motivational aspect involves the assessment of the status of one’s goals, and the cognitive aspect pertains to the appraisal of the situation’s relevance to one’s life.

The appraisal process can be divided into primary and secondary appraisals. In the primary appraisal, individuals evaluate motivational relevance (“*How relevant is this situation to my needs?*”), which determines the intensity of the emotion, and motivational congruence (“*Is this situation consistent with my goals?*”). The secondary appraisal involves evaluating resources and options for coping, and determining accountability (“*Who should be held accountable?*”).

For example, if a situation is motivationally relevant and incongruent with an individual’s

goals, and another person is held accountable, the resultant emotion is anger. Conversely, if there is no obvious person to hold accountable, the emotion may be anxiety. Roseman's theory of appraisal further elaborates on these aspects by specifying the conditions under which specific emotions are elicited [37].

Motivational theories

A different group of theories has focused on what *moves* an individual, namely, their motivations, which is the core concept to which emotions had been linked from their very definition. The field of motivational theories explain emotions as responses that arise from the evaluation of events concerning personal goals, describing how these emotions influence behavior related to such goals in order to achieve them.

A pivotal figure associated with motivational theories, Bernard Weiner [38], explains how individuals attribute causes to events and how these attributions influence their emotional state. For example, if a student attributes their failure in an exam to lack of ability, they will likely feel hopeless and unmotivated, whereas an attribution to a lack of effort might elicit feelings of guilt.

Another prominent contributor, Richard Lazarus, broadens these perspectives by integrating cognitive appraisal and motivational aspects of emotions, to explain how they arise from the evaluation of personal significance and coping potential in relation to one's goals [39].

Constructivist theories

Concluding this overview of emotion theories, we turn to the constructivist approach, which stands in contrast to the evolutionary framework that underpins most theories discussed thus far. While classic perspectives describe emotions as universal, genetically encoded biological processes that emerged during evolution as adaptations to environmental challenges or opportunities, the constructivist approach emphasizes their pronounced differences across cultures as a proof of their socio-cultural specificity. Thus, constructivist theories define emotions as relative constructs or forms of discourse that emerge within specific cultural or social contexts.

Lisa Feldman Barret, the most influential researcher in the field, formulated what is known as the *Theory of Constructed Emotion* [19]. She proposes that emotions, including basic emotions, are prototypical instances of affectively founded, contextually specific, goal-based concepts that are constructed by the brain to ultimately serve allostasis (or "body budgeting", which is the optimal use and distribution of bodily functions for survival and well-being). From this neuroscientifically-grounded perspective, therefore, emotions are not pre-wired or universal, but are constructed as situationally-dependent concepts.

Currently, more balanced perspectives affirm that while some aspects of emotions are indeed universal and biologically grounded, different cultural contexts influence how they are expressed, experienced, and understood by individuals. As suggested by Matsumoto and Hwang, the relative contributions of biology and culture may have varying impacts across different types of emotion [40].

Emotional Intelligence

Finally, shifting from theories focused on the origins and nature of emotions, we turn to the framework of *Emotional Intelligence* (EI), which approaches emotions as essential components of cognitive and interpersonal skills. Defined by Peter Salovey and John D. Mayer in 1990 [41], EI is described as an individual’s ability to perceive, understand, manage, and use emotions—both their own and those of others—to facilitate thinking. This model identifies four essential dimensions:

- Perceiving emotions: The ability to accurately recognize emotions in oneself and others.
- Using emotions to facilitate thought: Leveraging emotions to prioritize and guide cognitive activities.
- Understanding emotions: Analyzing emotions, their trajectories, and interactions with other emotions.
- Managing emotions: Regulating emotions effectively to achieve personal and relational goals.

This framework expands the understanding of emotions beyond reactive phenomena, underscoring the functional role emotions play in cognition and decision-making. As a psychological construct, EI provides a skill-based interpretation of emotionality that has influenced diverse fields, including education, clinical psychology, and, later, Affective Computing and AI, as we will see in subsequent sections.

1.1.5 The neuroscience of emotions

The field of affective neuroscience provides the bridge between psychological theories and their biological underpinnings, or what Jaak Panksepp described as “*the view of emotions from the inside*” [42]. This interdisciplinary approach has made significant contributions to our understanding of the neural mechanisms and brain structures (the neurophysiological substrate) involved in the emotional processes, offering insights into the complex interconnection among cognition, physiology, and behavior that characterizes emotional experiences.

Neuroscientific research has identified how key brain areas such as the amygdala, prefrontal cortex, and hippocampus play pivotal roles in emotional regulation, processing and memory. For example, LeDoux’s research, in one of the seminal works in affective neuroscience, demonstrated the crucial role of the amygdala in processing fear-related stimuli and coordinating fear responses [43].

One of the most relevant contributions to the field is the *somatic marker hypothesis*, proposed by Antonio Damásio, which posits that emotional processes guide, or bias, the decision-making processes through bodily sensations associated with previous experiences that arise in bioregulatory processes, like emotions and feelings. For example, when deciding about a financial investment, somatic markers related to past negative experiences may produce a feeling of unease, prompting a more careful consideration of the outcomes [3].

Providing direct evidence of basic or primordial emotions, Jaak Panksepp has taken our

understanding of emotions ever closer to their neural foundations. His groundbreaking work identified seven distinct emotional systems, each characterized not only by differentiated associated behaviors but, most remarkably, by distinct neural circuits and neurochemicals within the mammalian brain. These biologically inherited primary affective systems are divided into positive and negative categories. The positive systems include *seeking* (expectancy), *care* (nurturance), *play* (social joy), and *lust* (sexual excitement), while the negative systems consist of *fear* (anxiety), *rage* (anger), and *panic/grief* (sadness), generated by what he describes as the “core-self” [44].

Finally, the role of neurotransmitters and neuromodulators in emotional processes, by facilitating or inhibiting neuronal activity and influencing learning and memory mechanisms, has been a significant focus of research in the field. It is now well known that dopamine plays a critical role in emotions and motivation [45], influencing reward-seeking, reward learning and decision-making processes, thus guiding animal behavior by reinforcing actions that lead to rewarding outcomes. Similarly, serotonin has been found to relate to mood regulation and anxiety [46]. It has also been associated with self-confidence, inner strength, and satisfaction, highlighting its broad influence on emotional well-being and behavior [47].

The field of neuroscientific research continues to evolve, incorporating and refining technologies such as functional magnetic resonance imaging (fMRI) to map the neural correlates of emotions, as well as positron emission tomography (PET) and electroencephalography (EEG), which have enhanced our ability to visualize brain activity with unprecedented detail. The use of these advanced techniques provide critical experimental substrates that validate or challenge psychological and biological theories, bridging gaps between theoretical models and empirical evidence by providing an increasingly accurate and interpretable view of the very neural basis of emotions.

1.2 Emotions in AI: Related work in Affective Computing and Reinforcement Learning

HELENA: But they're so intelligent.

HALLEMEIER: Immensely, but they're nothing else. They've no will of their own.

No soul. No passion.

HELENA: No love?

HALLEMEIER: No. Absolutely not. Robots don't love. Not even themselves.

Karel Čapek, *R.U.R. (Rossum's Universal Robots)* (1923)

To date, the evolutionary advantages registered and studied by the fields of biology, neuroscience and psychology, have not been effectively utilized in the field of Artificial Intelligence (AI), in which the study of emotions and feelings, vigorously developed since the 19th century, has had limited resonance. This stands in stark contrast to the profound influence exerted on AI (and vice versa) by breakthroughs from neuroscience [48], such as artificial neural networks [49][50][51], bioinspired neural architectures [52][53][54], and attention mechanisms [55][56];

biology, with examples like evolutionary algorithms [57][58], multi-agent systems [59] or swarm intelligence [60]; and psychology, for example goal-oriented behavior or reinforcement learning (RL) [61]. Instead, AI research has historically focused on emulating human reasoning, namely, the rational performance of our natural minds, relegating their emotional dimension to a second plane, or most frequently neglecting it. The field concentrated in the ascription of reason, or “cold logic”, to machines [62][63].

Different factors may have influenced this late and slow integration of synthetic emotions in AI, like the difficulty to define and measure them, historic-cultural hyper-rationalism and dichotomy of reason and emotion [3], and the early results obtained from classic reason-based symbolic AI [64]. However, one major deterrent may be the absence of a universally accepted psychological framework describing how emotions develop from first principles—raw inputs and prior learning—that specialists can apply or adapt to machines [65], as discussed in section *The psychological labyrinth*.

The work presented here, building upon key breakthroughs in these fields, aims to bridge this gap by defining such a framework from a computational perspective.

1.2.1 Artificial intelligence is born without artificial emotion

The early history of Artificial Intelligence, as the field was first named in 1955 by John McCarthy [66], has been that of the imitation of the rational, that is, the intellectual, achievements of the human mind. This focus was heavily influenced by Western hyper-rationalism and the profound dichotomy of reason and emotion which can be traced back to Descartes and ancient Greek philosophy. It may be self-revealing that the very name “Artificial Intelligence” relates to the “intellect” (linked to knowledge, reasoning and understanding), extirpated from broader natural systems such as the “mind” or the “psyche”, to which the emotional phenomena had already been associated at the time.

This early departure from the emotional side of natural minds may be justified by the classic belief that emotions interfere with the desirable rational performance of the mind in these relevant matters, and should therefore be eradicated from our mental processes [67][68].

In Western philosophy, emotion was often regarded as a less advanced or refined mode of understanding, in contrast with more elevated and rational forms of thought [69]. As described by Calhoun and Solomon (1984), “*many [great thinkers] neglected it altogether. Some treated the emotions with disdain, as the ‘lower’ part of the soul*” [70].

The origins of this “disdain” can be found in Plato, an early proponent of rationalism who emphasized reason over emotions and sensory experiences, or in Aristotle, who, despite acknowledging and studying human emotions, prioritized reason in his ethical and political philosophy. During the Enlightenment (16th century), René Descartes—the “father of modern philosophy” and considered the first of the modern rationalists—initiated the epistemological turn that focused philosophy on the theory of knowledge, asserting that reason is the main source of knowledge. His prominent dictum “*cogito, ergo sum*” (“I think, therefore I am”), from *Discourse on the Method*, illustrates his view that certain knowledge can be attained through reason alone [71]. Another central figure from the Enlightenment, Immanuel Kant,

argued in 1764 that principles lead to consistent and reliable actions, whereas emotions, despite their role in certain moral feelings, are triggered by specific situations and, therefore, subject to change, leading to inconsistent behavior [13]. Principles are not easily swayed by external circumstances, making them more reliable and enduring.

However, not all thinkers accepted this primacy of rationality. David Hume, for instance, famously argued that “*reason is, and ought only to be the slave to the passions*,” emphasizing the role of emotions in human behavior and decision-making [12]. Later, Friedrich Nietzsche [72] and Arthur Schopenhauer [73] critiqued pure rationalism, arguing for the importance of will and instinct over pure reason. The supremacy of reason was not universally accepted, even during the heights of rationalist thought.

Nevertheless, the primacy of rationality over other dimensions of the human mind continued to inspire progress during the Scientific Revolution (16th-18th centuries), laying the groundwork for logical positivism in the 20th century, with emphasis on reason and empirical observation—principles on which modern fields like AI were founded.

Thus, in its early days, the field focused on emulating human reasoning, namely, the desirable rational performance of our natural minds, aiming to create systems that could solve complex problems traditionally associated with human intelligence. These included board games (where chess was referred to as the “*drosophila of AI*”), medical diagnosis (with systems like MYCIN, developed in the 1970s), natural language processing (with applications to translation or summarization), logical reasoning (for theorem proving and formal logic), search and optimization (for scheduling, navigation, or logistics problems), etc. The field concentrated on the ascription of reason, or “cold logic”, to machines [62][63], and the emotional dimension of the mind was relegated to a second plane, or most frequently fully neglected.

1.2.2 Early skepticism about emotion in machines

With AI focused primarily on the cognitive dimension of the mind, many of the first academic approaches to the emotional phenomenon in machines emerged in mixed forums, encompassing both psychology and philosophy of science. These discussions revolved around several aspects: the feasibility of simulating or synthesizing emotions in machines, their potential utility in intelligent systems, and the broader philosophical and ethical implications of such endeavors.

As early as 1951, British physicist and cybernetician Donald MacKay, one the most lucid and influential figures in the philosophy of mind and early artificial intelligence, reasoned in *Mindlike Behaviour in Artefacts* that “*an artefact capable of receiving and acting on information about the state of its own body can begin to parallel many of the modes of activity we associate with self-consciousness*.” He lamented that “*present-day digital computers are deliberately designed to show as few as possible of the more human characteristics*.” Furthermore, in what might be the earliest formal attempt at designing artificial emotions, MacKay visionarily introduced “probabilistic reasoning,” suggesting that such a flexible mechanism could manifest human traits like “*continuously-variable (and irregularly excitable) prejudices, preferences, and other ‘emotional’ effects*” through appropriately linked transition probabilities [74].

Following MacKay’s early proposals, the notion of machines possessing emotions was historically met with strong criticism. Michael Scriven, in *The Mechanical Concept of Mind* (1953), argued that while machines could imitate human behavior, “*no matter how ingenious the mechanism, how complex the behaviour of a machine [...] it’s no more conscious than a clock.*” He maintained that feelings such as pleasure or self-pity were exclusive to conscious beings, dismissing the possibility of machines having true emotions—although not addressing their potential utility [75].

Paul Ziff reinforced this critique in *The Feelings of Robots* (1959), arguing that machines could perform actions resembling emotions, but these were merely “performances.” According to Ziff, “*no robot could sensibly be said to feel anything*” because robots are mechanisms, not living beings, and their behavior is entirely programmed. He emphasized that robots have no individuality or genuine psychological states: “*we could make a robot say anything we want it to say,*” but this would be no more meaningful than a “*phonograph recording*” [76].

This line of critique was extended by Keith Gunderson in *Mentality and Machines* (1971), where he questioned whether non-behavioral mental aspects, such as experiencing emotions, could be simulated by computers. He posed what he called the “simulability question,” which asks whether phenomena like feeling pain, anxiety, or boredom can be modeled using the same techniques applied to problem-solving or pattern recognition. Gunderson’s clear conclusion was negative: “*there is a variety of mental phenomena that includes the having of pains, after-images, feeling anxious, being bored, etc., that are not receptive (or primarily receptive) to being simulated in the sense that, say, theorem proving is*” [62].

But probably the best illustration of the diametrical opposition regarding the role of emotions in machines within the AI community can be seen in the contrasting views of John McCarthy and Marvin Minsky, pioneering “fathers” of the field. McCarthy argued in 1983 that while attributing mental qualities to machines could enhance our understanding of their behavior, “*the anthropomorphism should not include emotions*” [77]. In stark contrast, as will be discussed later, Minsky viewed emotions as essential to human intelligence, proposing that they could be understood as complex mechanisms, potentially replicable in machines [78]. Beyond a technical debate within AI, this clash underscores a deeper philosophical divide about the nature of mind and emotion—one that persists to this day.

1.2.3 Increasing interest but little material progress

Despite ongoing criticism and skepticism, the exploration of emotion in machines continued, with various scholars proposing significant theoretical models and approaches, as well as supporting arguments. One of the most notable contributions in the early 1960s was the introduction of “hot cognition.” This concept, championed by psychologist Robert Abelson, sought to address AI’s narrow focus on “cold” cognition—logical reasoning and problem-solving devoid of emotional content. His use of the term “hot cognition” referred to cognitive processes intertwined with emotional significance, contrasting with the emotionally neutral, “cold” operations that dominated AI research. In his 1963 paper *Computer Simulation of “Hot” Cognition*, Abelson laments that “*there seems to have been no provision in the computer game for the study of cognition dealing with affect-laden objects.*” He proposed a rule-based model to

simulate emotional processes such as incredulity, denial, and rationalization when processing new beliefs or sentences, offering an early attempt to model affect-laden cognitive dynamics using computer simulations. Abelson argued that human cognition is often shaped by emotions and suggested that computers could simulate these processes as well. His ideas, presented at a psychology conference, stood apart from the prevailing engineering-driven research, which prioritized logical problem-solving, and thus remained largely in the background [79].

Another remarkable approach came from Aaron Sloman and Monica Croucher in 1981, who, drawing on psychological insights, argued that “*the need to cope with a changing and partly unpredictable world makes it very likely that any intelligent system with multiple motives and limited powers will have emotions.*” They challenged the belief that emotions and intellect are entirely separate, suggesting instead that emotions are essential for managing competing goals in intelligent systems. In their paper *Why Robots Will Have Emotions*, they proposed a computational architecture where a “central administrative process” controls behavior by managing motives, resources, goal-directed processes, and perception devices, emphasizing the importance of emotions in complex decision-making [80].

The work of Marvin Minsky, one of the founders of MIT’s artificial intelligence program, played a pivotal role in reshaping how emotions are viewed in the context of artificial intelligence, challenging the traditional separation between emotion and intelligence, and arguing instead for their interconnectedness. In *The Society of Mind* (1986), Minsky emphasized the essential role of emotions in cognitive processes, stating that “*the question is not whether intelligent machines can have any emotions, but whether machines can be intelligent without any emotions*” [78]. He described emotions as a set of learning and coping mechanisms, required in any comprehensive theory of intelligence. Later, in *The Emotion Machine* (2006), Minsky expanded on this idea, arguing that separating intellectual and emotional processes is futile, as both are “ways of thinking” [81]. For him, emotions are integral to intelligence, arising from neural network processes in the brain.

Antonio Damásio, one of the most influential figures in neuroscience and philosophy, has significantly shaped the discourse on the role of emotions in decision-making and intelligence. On top of his above-mentioned relevant contributions to the field, his research demonstrated that individuals with damaged emotional centers in the brain struggle with decision-making, illustrating that emotions are integral to rational thought [3]. Damásio’s work has inspired key voices in AI and computational intelligence, such as Fulcher, to argue that “*instead of computers becoming more unpredictable and irrational through having emotional intelligence, they would instead act and behave more rationally and predictably*” [82].

In his later work *The Strange Order of Things* (2018), Damásio describes how our minds operate in two registers. The first governs perception, movement, memory, reasoning, language and calculations—a domain of “synaptic signals” that is effectively captured by current AI and robotics. Then, he insightfully addresses the critical limitations of AI, emphasizing that “*there is a second register that pertains to emotions and feelings that describes the state of life in our living body and that does not lend itself easily to a computational account. Current AI and robotics do not address this second register.*” [83]

In close parallel with earlier concerns about AI’s overemphasis on “cold logic,” Melanie

Mitchell argued in her 2021 paper *Why AI is Harder Than We Think* that the idea of “pure rationality” in intelligence is fundamentally flawed [84]. She contended that emotions and so-called “irrational” biases are not barriers to rationality but are deeply intertwined with human intelligence. Mitchell stressed that no evidence in psychology or neuroscience supports the notion that rational cognition can be separated from emotions. Instead, she emphasized that human intelligence is an integrated system, combining emotions, desires, and cultural biases with cognitive functions, and warned that aiming for emotionless “superintelligent” machines could lead to unrealistic and dangerous predictions.

1.2.4 The birth of Affective Computing and related fields

The echoes of these early insights led to the first serious attempts to integrate emotions in AI, which initially focused on simulating their expression during human-computer interactions for more believable agents [85][86]. For example, Bates (1994) argued that “*appropriately timed and clearly expressed emotion is a central requirement for believable characters*,” which led to his development of “believable agents”—virtual characters capable of human-like emotional expression. One major breakthrough in his research was the *Oz Project*, where real-time interactive creatures were designed using principles from traditional character animation, inspired by figures like Walt Disney and Chuck Jones (Fig. 1.5). This project, showcased in *Edge of Intention* at the AAAI-92 AI-based Arts Exhibition, implemented a behavior-based architecture for action, coupled with an emotional framework based on the Ortony, Collins, and Clore (OCC) model—which is discussed further below—to create more authentic and engaging virtual agents, expressing emotions by changing shape.

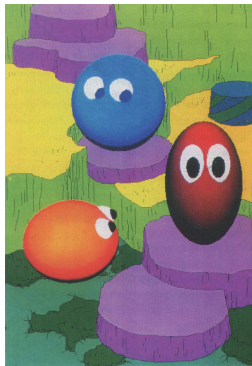


Figure 1.5: The Woggles in *Edge of Intention*. Some of the first “believable agents” were these self-animating creatures displaying contextually coherent emotions, showed at the AAAI-92.

In parallel with these developments, another critical dimension of affective computing emerged: the detection and recognition of emotions from human inputs. This evolution, initiated by Rosalind Picard, the pioneering figure of the field, led to the establishment of affective computing as a distinct interdisciplinary domain in 1995 [87]. Picard’s work framed the field as concerned with machines that can detect, interpret, and respond to human emotions through diverse signals such as text, speech, facial expressions, and physiological data. The field now spans psychology, cognitive science, and computer science, with continuous and substantial cross-disciplinary contributions.

In their review in 2005, for example, Tao and Tan [88] highlighted its reliance on data from multimodal sensors to model and recognize emotions in real-time, especially in applications like intelligent interaction systems. They examined the current methods for detecting and processing emotional states, the interactive feedback loop in human-computer interaction, and the challenges in making machines that can both interpret and express emotions. By 2024, as reflected by Khare [89], significant advancements had been made, focusing on using machine learning (deep learning and multimodal AI) to enhance the precision of emotion recognition across subjects, particularly through physical and physiological signals such as EEG (electroencephalogram), allowing for accurate cross-subject emotion recognition [90][91][92] (see Fig. 1.6). Additionally, the use of fuzzy logic or emotion-adaptive control systems in human-machine interfaces has expanded the capabilities of these systems, allowing for more nuanced interactions that account for emotional states [93][94].

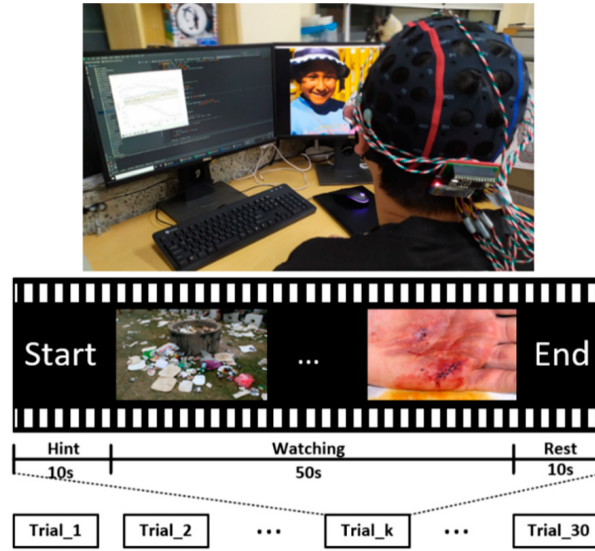


Figure 1.6: Electroencephalogram (EEG) in Affective Computing. Protocol of an EEG experiment conducted to elicit and detect emotions in human subjects, by Mai et al. (2021).

Within the field of robotics, an early influential application of emotional processing is exemplified by *Kismet*, a robot developed in the late 1990s by Cynthia Breazeal at MIT’s Artificial Intelligence Lab [95]. *Kismet* was designed to interact socially with humans by recognizing and expressing emotions through a combination of facial expressions, vocalizations, and head movements (Fig. 1.7). A motivation system regulates and maintains “*the robot’s state of ‘well being’ in the form of homeostatic regulation processes and emotive responses,*” eliciting six basic modeled emotions (anger, disgust, fear, joy, sorrow, and surprise, after Ekman), as well as arousal-based responses (interest, calm, and boredom). The robot’s multimodal interaction system used facial recognition and vocal intonation to engage in human-like social behaviors, pioneering the field of affective robotics toward socially aware interactive machines.

Another major advancement in the application of affective computing to robotics is *Pepper*, developed by SoftBank Robotics (formerly Aldebaran Robotics) [96]. Introduced in 2014, *Pepper* is a semi-humanoid robot capable of recognizing and interpreting basic human emotions through facial expressions and voice tones (Fig. 1.8). Optimized for social interaction, it has

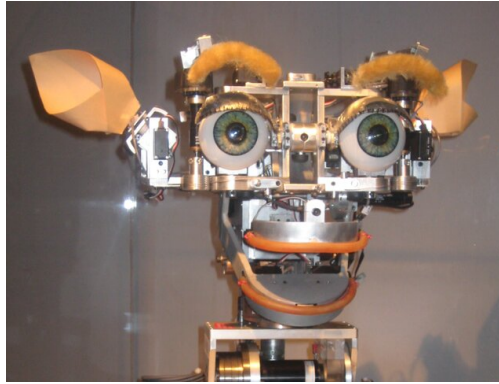


Figure 1.7: *Kismet* (1998). One of the first social robots, developed at MIT’s AI Lab, was capable of recognizing and expressing emotions.

been used in various settings, including offices, where it acts as a receptionist by identifying visitors, sending alerts, and even managing small tasks like taking drink orders or conducting health and safety briefings. Its ability to autonomously engage in conversations makes it a versatile tool in environments ranging from customer service to healthcare.



Figure 1.8: *Pepper* (2014). Capable of various social interactions, this semi-humanoid robot by SoftBank Robotics could recognize and interpret emotion in voices and faces.

Following this and similar developments, *Sophia*, an advanced female humanoid robot developed in 2016 by Hanson Robotics, was designed to engage in conversations involving emotional intelligence [97]. Equipped with cameras in its eyes, it employs computer vision to track faces, recognize individuals, and interpret emotional cues such as facial expressions and gestures (Fig. 1.9). *Sophia* can also process natural language, allowing it to respond to vocal intonations, simulating emotional intelligence during interactions. Her ability to generate over 60 lifelike facial expressions, aided by her advanced skin and AI-driven motors, helps mimic human emotions, additionally featuring a robust text-to-speech engine for fluid conversation—and even singing. Despite her sophisticated architecture, which includes symbolic AI, neural networks, and the OpenCog framework for general reasoning, much of her interaction remains scripted or semi-autonomous, blending AI with human assistance.



Figure 1.9: *Sophia* (2016). Developed by Hanson Robotics, this emotional robot could recognize faces, express facial emotions, and hold coherent conversations.

More recently, advances in humanoid robotics have focused on enhancing emotional interaction, with companies like China’s EX-Robots leading the way. At the 2024 World Robot Conference [98], these robots showcased advanced facial expressions and dexterous hands, enabling them to foster emotional connections with users in ever more accurate realistic and fashions (Fig. 1.10).



Figure 1.10: World Robot Conference (2024). A humanoid robot at the World Robot Conference 2024 in Beijing.

The applications of affective computing are extensive and diverse. In health, systems like those developed by Taylor et al. (2017) use multitask learning to predict mood, stress, and health [99], while wearable technology enables monitoring of conditions like atrial fibrillation [100]. Stress detection, as shown by Healey and Picard (2005), has practical uses in tasks like driving [101]. Educational environments benefit from emotion recognition systems, such as those created by Miranda Correa et al. (2018), to gauge engagement [102]. Affective computing is also applied in anti-bullying education, using emotionally aware agents to foster empathy and positive behavior [103], and interactive entertainment, by enabling virtual and gaming environments to adapt to users’ emotions in real-time, improving engagement and immersion in current augmented and virtual reality contexts [104].

In summary, with applications spanning brain-computer interfaces, virtual reality, robotics, and empathic dialogues [105], the field is clearly making steady progress toward agents that can both understand and display emotions. In contrast, however, from its very inception, Affective

Computing has distinctly set itself apart from the goal of creating computers that actually produce their own emotions. As Rosalind Picard explicitly stated in her foundational 1995 paper, “*I am not proposing the pursuit of computerized cingulotomies [a surgical procedure to aid severely depressed patients] or even into the business of building ‘emotional computers’*” [87]. Although this constraint has not always been strictly observed—as we will discuss in the next section—for practical, cultural and ethical reasons, the field has traditionally concentrated on the recognition and expression of human affects. As a result, significantly less progress has been registered in the complementary field of emotion synthesis, which is the focus of this research.

1.2.5 Non-learned models of emotion synthesis: State-of-the art and limitations

Therefore, the fundamental challenge remains: creating AI agents that can elicit emotions that robustly compare to natural ones, leveraging the evolutionary advantages registered and studied by the fields of biology, neuroscience and psychology. Most research in this area is predominantly theoretical [78][81][106][107][108][94], with little practical advancement in learning, synthesizing, and integrating emotions in AI models and frameworks [84]. Although a number of proposals have demonstrated some potential, specially in the field of RL, they have had limited impact, covering a very narrow scope of emotions, or failing to generalize beyond case-specific solutions and arbitrary or hard-coded emotional models [109][65], with most using predetermined categories and dynamic characteristics of emotions that are explicitly pre-programmed by researchers. Innovative exceptions are rare and typically constrained to small, predefined emotion sets [110], or based on ad-hoc engineered feature variables for specifically-designed environments or action spaces [111][112].

In this section, we review some of the most relevant non-learned proposals, highlighting their key contributions and identifying the limitations that may have hindered their broader adoption. By contrast, the following section (1.2.6) will explore the less mature field of self-learned emotion models.

Minsky and the “emotion machine”

Marvin Minsky’s influential *The Emotion Machine* (2006) [81] proposed a theory of mind that blends emotions, reasoning, and problem-solving, arguing that emotions are simply different ways the mind thinks. This builds on the concept he had earlier introduced in *The society of mind* (1986) [78], where emotions, activated in response to specific problems or situations, emerged from interactions between the simpler, mindless agents within the global model of the human mind he proposes.

However, despite the significant theoretical impact of Minsky’s work on AI and robotics, its practical applications have been relatively limited. This is largely due to the complexity of implementing these models—which involve coordinating multiple levels of emotional thinking and common sense—as well as the difficulty of translating them into algorithms—a task not explicitly addressed in his work. Additionally, the dominance of data-driven AI approaches over symbolic and cognitive models has further hindered the practical application of his ideas.

In conclusion, while Minsky’s frameworks have inspired conceptual advances, the full realization of these ideas in practical systems, beyond the realm of theory, remains a challenge.

The OCC model: A computational framework for emotion synthesis

Named after its creators Ortony, Clore, and Collins, the OCC model is one of the earliest and most commonly applied computational frameworks for emotion synthesis since its groundbreaking introduction in 1988 [107]. Introduced as a cognitive appraisal model, it defines emotions as valenced reactions to specific situations and offers a structured approach to understanding how emotions are triggered and categorized. The model outlines a fixed set of 22 distinct emotion categories based on cognitive evaluations of three key types of stimuli: goal-relevant events, actions performed by accountable agents (which may include the self), and attractive or unattractive objects. Its application to a new problem requires the ad-hoc definition and programming of these stimuli, which then interact with three predefined value fields—goals, standards, and attitudes—to determine the type and intensity of the elicited emotions.

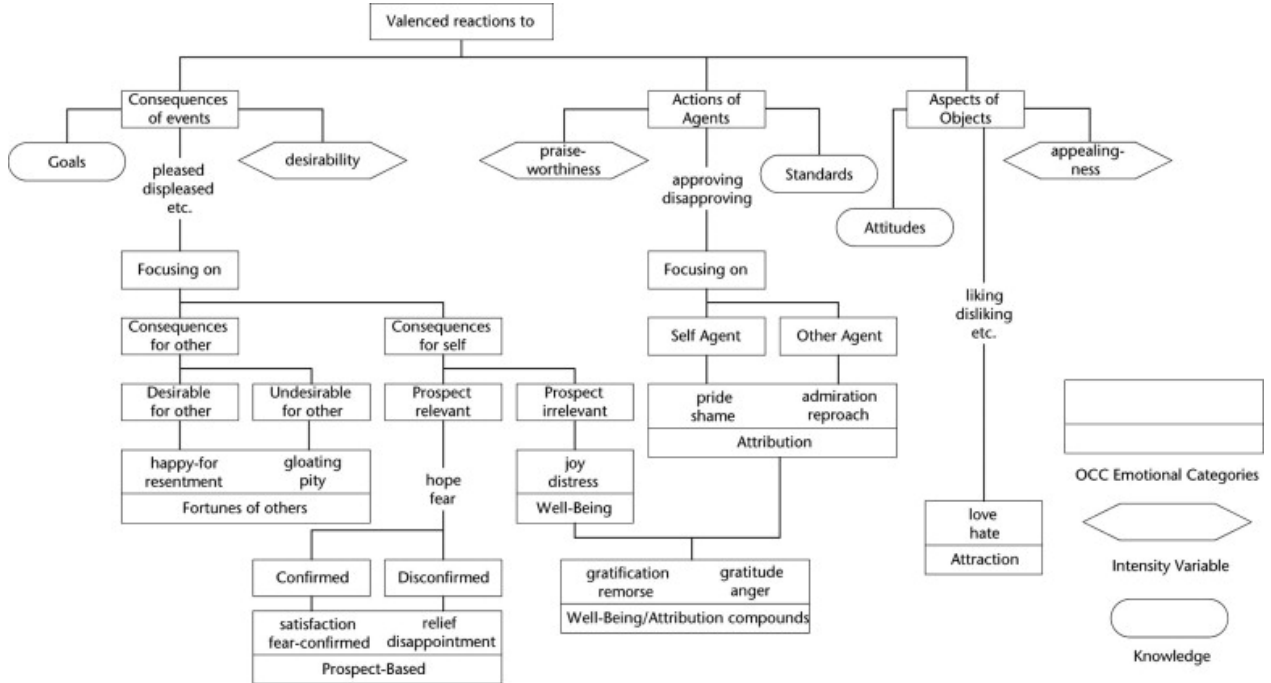


Figure 1.11: The OCC model.

At its core, the OCC model is based on three types of eliciting conditions: events, agents, and objects. Each of these corresponds to one of the value fields, creating a system in which the likelihood of an event, the accountability attributed to an agent, or the familiarity and attractiveness of an object can be assessed to generate emotional responses. The model provides a structural framework for understanding emotions, emphasizing characterization over causation. While it offers a detailed structure for emotional variables, it remains adaptable in its implementation. Primarily descriptive in nature, the model provides a framework for categorizing emotions without prescribing specific appraisal processes.

Some real implementations of the OCC model demonstrate its versatility in virtual environments. One notable example is Elliot’s *Affective Reasoner* (1992), which integrated all 22 emotional categories of the model to enable agents to communicate their emotions to each other in a multi-agent system [113]. Another significant implementation is Koda’s *Poker Playing Agent* (1996), where the OCC model was employed to express the most dominant emotional category, although its expressive capabilities were limited by its facial features [114]. These implementations show the range of applications, though they also reveal the challenges in effectively utilizing the full complexity of the model.

Further real-world uses include Bartneck’s *eMuu* (2002), an embodied emotional character developed for the ambient intelligent home, which focused only on the most dominant emotional category due to expressive limitations [115]. Additionally, Bondarev’s *Emotion Management System for a Home Robot* (2002) incorporated the OCC model to manage emotions in the context of home robotics, addressing interaction with the environment [116].

Despite these examples, and its consideration as the “standard” emotion model for many years, numerous limitations of the OCC model have been described and discussed in the field, [117], including its heavy context dependency, inability to adapt to complex real-world scenarios, reliance on predefined world models, and expressive limitations that prevent all 22 emotional categories from being conveyed convincingly. Additionally, its over-detailed categorization and requirement for an accurate user model further complicate its practical application, suggesting the need for simplification or automation to improve its usability.

The TDRL theory of emotion and related approaches

Transcending some of the limitations of models like OCC, the field of Reinforcement Learning has been the basis of several interesting proposals, associating emotions with value and reward functions [109]. The simplest ones define emotion as the expected cumulative reward from taking an action (or state-action value) [118]; as the state value, for example, to detect fear [119]; as separate positive and negative values, representing hope and fear, respectively [120]; or as the lowest registered value for a given state to indicate fear [121].

The reward signal, as a primary concept in RL, has been the basis of some other proposals: the ratio between short- and long-term average reward has been linked to valence [122][123][124]; the temporal change in its values is the basis of a four-emotion model (joy, anger, fear and relief) [125]; the average reward is used to characterize distress and comfort [126]; and the reward itself is taken as the emotional signal in several other implementations [127].

An alternative line of research draws on the well-documented parallelism between the role of dopamine in nature and the temporal difference (TD) error, a foundational concept in the Reinforcement Learning framework. Early studies, drawing on psychology insights, defined feelings of happiness, intense happiness, or disappointment as responses to actual rewards that either matched, exceeded, or fell short of an agent’s expectations, respectively [128]. This research trajectory was subsequently advanced by several scholars [129][130], ultimately leading to the most influential model based on this concept: Broekens’ *Temporal Difference Reinforcement Learning (TDRL) Theory of Emotion* (2018) [106][131], on which we will now focus more closely.

Built on the hypothesis that “*all emotions are manifestations of temporal difference error*,” the TDRL model describes five emotions: joy (positive TD), distress (negative TD), hope (anticipation of a positive TD), fear (anticipation of a negative TD), and regret (involving counterfactual thinking, identifying underestimated actions after new reward evidence).

Beyond theoretical analysis and explorations, the TDRL model has been applied in several practical tests for experimental validation, yielding mixed results. In 2019, Dai and Broekens conducted a study aimed at verifying the plausibility of simulated fear emotions as perceived by humans [132]. The experiment was carried out in a basic, fixed 21-state grid-world environment, where participants were asked to interpret the emotion of fear simulated by the TDRL model. The findings indicated that humans could indeed recognize simulated fear emotions based on the TDRL framework. However, the study’s focus was limited to a single emotion, requiring the validation of six different TDRL calculation methods specifically for fear, which points to a limitation in the model’s generalizability across other emotional states.

More recently, Lycklama et al. (2023) addressed the issue of lack of transparency in RL behavior by implementing emotional expressions grounded in the TD model [133]. In their case study, a human teacher guided a robot in understanding the meaning of three different colors. Although they successfully integrated emotional expressions into the robot’s learning process, a practical comparison between conditions with and without visible emotions revealed “minimal differences” in outcomes, suggesting that for relatively simple tasks, emotional expressions grounded in RL neither significantly enhance nor impair performance.

Despite its merits, the TDRL model presents several challenges alongside its achievements. On the positive side, the model effectively describes five crucial emotions—joy, distress, hope, fear, and regret—grounded in the fundamental values of RL. It captures the agent’s subjective appraisal of its situation by linking positive or negative temporal difference (TD) errors to correspondingly valenced emotions. This connection to TD error not only highlights the agent’s evaluative perspective but also provides a clear framework for understanding emotional responses based on performance expectations. Furthermore, the model has also demonstrated its practicality through implementation in controlled experiments, as illustrated above.

However, the TDRL model has notable limitations that constrain its broader applicability. A primary drawback is its narrow emotional spectrum, limited to just five predefined emotions. Its potential extensibility is further restricted by definition to emotions directly tied to the agent’s misjudgments: because the TD error reflects only discrepancies between expected and actual outcomes, it offers little to no information when events align with predictions. For example, receiving an expected reward can not elicit satisfaction based on TD, and continued suffering from correctly anticipated damage fails to elicit suffering or grief.

Another aspect that limits the model’s generalizability is the predefined nature of its emotional patterns, which are not learned or shaped by the environment, but pre-established through an arbitrary formulation for each emotion. As a result, their elicitation and interpretation must be customized and hardcoded for the task, rendering the process more rigid and environment-dependent. Additionally, the model does not naturally integrate the synthesized emotions with artificial neural network (ANN)-based policies (such as through vectorization), necessitating explicit, environment-specific coding of its effect on the agent’s behavior.

Moreover, the model’s dependence on model-based RL imposes additional constraints on its application. For the agent to elicit fear and hope, it requires a mental model of its interactions with the environment that can represent uncertainty and anticipate potential outcomes.

Consequently, despite its conceptual longevity, the application of the TDRL model remains limited. Its practical use has predominantly been confined to relatively simple RL environments, such as small grid-worlds with full observability, thereby limiting its applicability to more complex or dynamic real-world scenarios.

1.2.6 Efforts toward self-learned emotion

As highlighted by Moerland et al. in their comprehensive and illuminating survey about emotions in Reinforcement Learning [109], a core challenge for the field’s future is to *“integrate all aspects into one larger system, potentially taking a fully learned approach.”*

Indeed, as discussed above, one of the main shortcomings of the approaches examined thus far is their arbitrary nature. This arbitrariness affects various characteristics of their definition, including the number and types of emotions covered, the formulation or logic governing their dynamics (such as elicitation and extinction), their integration within the agent’s policy or control code, and their association with external environmental elements. Consequently, while some of these models exhibit certain practical capabilities, they often lack the flexibility and generality necessary for application to other agents and tasks. Here, we analyze a few exceptions to this trend, where certain aspects of emotional phenomena are automatically learned during the process. Similar to our approach with non-learned models, we focus on their primary contributions while also noting specific challenges that may have limited their wider application.

The IBIA architecture

In the first exception discussed, an ambitious but mostly theoretical architecture is described by Bozinovski (2001) for self-learning agents, the integrated biologically-inspired architecture (IBIA) [94], following the principle of modeling a “complete creature” and based on the theory of emotion as an internal value judgment and appraisal. The architecture’s goal is to closely mimic how real-life organisms process emotions and use them as a basis for decision-making and learning.

The architecture is based on a connectionist model that integrates three components—the genetic, neural, and hormonal systems—combining emotions as an internal value judgment system that guides learning without external reinforcement. From this schema, a specific working architecture is distilled, known as the Crossbar Adaptive Array, designed to implement both situation appraisal (evaluation of the state the agent is in) and action tendencies (evaluation of the behaviors the agent might perform) using a process called emotion backpropagation.

The five interdependent components of the system are very differently parameterized and learned. The *genetic environment* sets innate primary emotions and biases, which are predefined. The *emotional learning memory* adapts through linear emotion backpropagation updates, incrementally updating action tendencies based on internal appraisals (without

external reinforcement). *Emotional state evaluation* simulates a neurohormonal pathway to generate signals that modulate the agent’s emotional responses, shaped by feedback from its neural processes. The *personality and motivation* system acts as an internal guide, defining traits like curiosity and sensitivity; while partially predefined, it can adapt in response to learning. Finally, the *action selection* mechanism interacts with the environment, initially guided by genetic predispositions but gradually refining its choices based on learned emotional evaluations, leading to more goal-directed behaviors.

Unlike most theoretical proposals on synthetic emotions, the architecture effectively outlines bioinspired mechanisms that are amenable to actual implementation. Nevertheless, its complexity, model heterogeneity, the diversity of learning paradigms—primarily reliant on linear learning dynamics detached from the dominant RL domain—and the need for hard-coding certain components, such as genetic parameters, may have deterred its adoption or refinement by researchers in the field. Furthermore, the model’s experimental validation was restricted to a highly simplistic setup (a small graph with predefined discrete states), which may not readily translate or scale to more realistic settings, such as continuous state spaces, or larger action or state spaces.

Emotions and the Free-energy principle

A different perspective was provided by Joffily and Coricelli in 2013 where a biologically plausible computational model of emotional valence is proposed, inspired by the free-energy principle [110]. In it, the valence of a state visited by an agent at time t is associated with the negative first time derivative of free-energy at that state $-F'(t)$, or more simply, the negative rate of change of free-energy over time. The basic idea is that emotional states reflect changes in the uncertainty about negative consequences of an agent’s actions, whose dynamics are associated with six basic forms of emotion by the authors, namely, happiness, unhappiness, hope, fear, disappointment and relief. Despite emotions not being explicitly learned in the framework, the authors analytically show how they emerge naturally as a consequence of the dynamics of the free-energy minimization process.

Notably, this approach reframes the classic dichotomies of opposites, such as reward-loss or pleasure-pain, typically associated with more valuable or desirable states, as “expected states.” The authors hypothesize that the emotional valence derived from free-energy dynamics, along with the associated “emotional states,” enables an agent to adapt its learning rate, enhancing its decision-making capabilities. Furthermore, the generality of the proposed scheme is considered high, as “*it is not tied to any particular generative model of sensory inputs.*”

Unfortunately, the practical demonstrations of this elegantly formalized framework are severely limited to extremely simple settings, falling short of demonstrating its viability in real-world scenarios. In their original work, the performance of two agents is compared in a basic slot machine (one-armed bandit setting) granting \$0 or \$1 outcomes. One of them explicitly estimates environmental volatility using a hierarchical Bayesian model, while the other uses emotional valence to regulate its learning rate dynamically. Although the emotional agent successfully replicates the adaptive behavior of the Bayesian agent, the experimental setup remains simplistic, involving only binary outcomes and deterministic transitions.

More recently, elements of Joffily and Coricelli’s model were incorporated by Yanagisawa et al. (2021) [134] into their work on a free-energy model of emotion potential, seeking to incorporate the “arousal” dimension of emotion. In this theoretical extension, a mathematical framework is proposed that associates free-energy with arousal potential aiming to explain emotional valence in the context of perception and decision-making. By formulating arousal as the Kullback-Leibler divergence of the Bayesian posterior, this model sought to provide a more detailed account of how emotional states might arise from the underlying dynamics of free-energy minimization. However, no experimental validations were reported in this work.

Similarly, Pattisapu et al. (2024) extended this work by developing a more comprehensive account of emotional states within the active inference framework [135]. They also proposed mapping both valence and arousal to specific aspects of free-energy minimization, thereby creating a two-dimensional model of emotion that aligns with Russell’s Circumplex Model [7]. Despite this remarkable theoretical achievement, the practical demonstration remained similarly limited. In this case, the authors tested a simple simulated agent under their formulation in a search task within a 12-state environment. Although they reported “*commonsense variability in emotional states*,” the simplicity of the task, much like earlier work, restricts broader conclusions about the model’s applicability in more complex, real-world settings.

In conclusion, despite the theoretical soundness of the framework and its extensions, the model remains primarily theoretical, capturing only a limited set of six emotions—happiness, unhappiness, hope, fear, disappointment, and relief. The association between emotional dynamics and these emotions is largely arbitrary and has not been validated against human observers or established psychological and neuroscientific accounts. Furthermore, experimental validation is very limited, with tests conducted in overly simplified environments that fail to reflect the complexity and ambiguity of real-world conditions. Consequently, while the model still serves as a valuable theoretical foundation for understanding the relationship between emotion, cognition, and learning within the free-energy principle framework, its practical applicability to more complex scenarios remains largely untested, suggesting challenges in adapting it to real-world settings.

Learning intrinsic emotion-based rewards

In this alternative approach to integrating learning, Sequeira et al. (2014) propose a method for reward modification, where an intrinsic reward for the agent is learned as a linear combination of four predefined emotion-based rewards [111]. These four rewards, inspired by the major dimensions of emotional appraisal within the Intrinsically Motivated RL framework [136], are: *novelty* associated with an action, given the agent’s history; *goal relevance* of performing the action; *degree of control* over the outcome of executing the action; and *expected valence* of executing the action. Specific formulae are provided for each of them, analogous to other hard-coded or predefined models, but in this case their optimal weights with respect to the overall goal achievement are learned (a brute-force Monte Carlo simulation approach is used, combined with a grid search over a defined parameter space). Despite this contribution, the features themselves are not learned, but rather ad-hoc formulated and engineered by the author. Moreover, their complex nature and high dependence on the task and environment makes their transfer to other setups very arduous.

Emotion-driven robotic path planning

More advanced machine learning techniques are proposed by Williams et al. (2015) to learn and synthesize emotions in a demonstration of adaptive robotic path planning [112]. In this approach, a Learning Classifier System—specifically the accuracy-based approach XCS—is trained through reinforcement learning to select the best possible action at each state. The emotion model is structured as a basic bow-tie structure, with inputs consisting of a set of primary reinforcers that generate “emotions” as a finite number of alternative (unlabeled) internal states, which are then connected to a behavioral modifier that adjusts the robot’s behavior. The authors demonstrate how, through trial and error, the system learns to map reinforcers to emotions, and emotions to modifiers, effectively modifying its behavior to maximize the reward, outperforming a non-emotional navigation system.

Despite these promising results, the approach has several notable limitations. A significant restriction is that the number of emotional nodes must be predefined—and narrowly limited to two in their case: fear and happiness—which greatly constrains the system’s ability to adapt to the complexity of other tasks. Additionally, lacking a systematic interpretation method, the emotional characterization of the two nodes was judged post-learning, based on the learned behavior. The selection of inputs, or “reinforcers”, also raises concerns. These high-level assessments (such as novelty, sense of agency, uncertainty, pain and progress) are arbitrarily chosen from Rolls’ list of primary reinforcers [137] based on their expected utility for robotic navigation. Consequently, they must be ad hoc selected, formulated and engineered by the programmer, rather than being learned directly from the actual external sensors (LIDAR, front and rear bump sensors, and forward facing ultrasonic sensors), requiring a complete redefinition and implementation for new tasks or environments. In conclusion, while the model successfully demonstrates the potential for emotion learning and the benefits of emotion-driven behavior in the targeted task, its low generality limits its applicability to other domains, which may have severely restricted its use by other researchers in different or more complex fields.

1.3 Concluding remarks

We have demonstrated how AI’s attention toward the emotional phenomena has been historically deprioritized, with the notable exception of Affective Computing. However, even in this flourishing field, the primary focus has been on detecting and expressing emotion in machines, rather than synthesizing them. As a result, the potential for developing true emotion-driven AIs remains largely untapped.

Within the initiatives effectively tackling artificial emotions, despite the abundance of theoretical approaches—ranging from abstract dissertations or essays, to more specific frameworks and architectures—practical implementations are scarce. Among the few exceptions, even fewer involve systems that can learn emotions autonomously. Additionally, these implementations commonly suffer from arbitrary or ad-hoc choices in the selection of targeted emotions, hard-coded models, or overly simplistic environments that restrict the emotional spectrum and dynamics captured. As a result, these systems fail to generalize well to different agents or more complex tasks.

Finally, very little progress has been made to test these models in more complex environments with continuous state spaces, larger action spaces, or non-stationary conditions, all of which are crucial for real-world applicability. We have also identified a general absence of verification against independent sources, such as human observer studies or established psychological and neuroscientific accounts, further limiting the credibility of these approaches.

In conclusion, the current state of AI research lacks a comprehensive and verifiable framework for emotions that is both generic and fully learned. There remains a gap in systems capable of naturally integrating learned emotions into behavioral drives, while credibly displaying the traits and benefits of natural emotions. Validated implementations that bridge this gap are essential for advancing the field toward truly emotionally intelligent artificial agents.

1.4 Overview of this work

Motivated by the above reasoning and the possibility that there lies an unrealized potential in truly emotional AIs, particularly in autonomous agents, this proposal addresses the origination of basic emotions from primary information in a way akin to natural organisms. We introduce a generic, fully self-learning emotional framework formalizing emotions as distinct temporal patterns perceived in crucial values for living beings that artificial agents can as well spontaneously learn and elicit during environment interactions with analogous, distinguishable dynamics.

The generic methodology introduced allows any AI agent to automatically learn, elicit and utilize its own synthetic emotional spectrum, successfully mirroring natural emotions described in the literature. Based on first principles in RL, it addresses emotion learning in an unsupervised manner, devoid of any preconceptions and encompassing the entire emotional spectrum. Indeed, unlike precedent case-specific solutions and arbitrary or hard-coded models, emotional patterns are learned directly from the agent’s interactions with an environment, namely from recent rewards, expected future rewards, anticipated world states, etc.

In addition to its generality, our work deviates from the predominant focus in affective computing and related fields by directly tackling the core concept of “emotions”, their formalization, elicitation and integration within practical agents, instead of their recognition through external sensors or their communication to users.

For the human interpretation of the learned emotions, a thorough and generic methodology is also introduced and tested that can be scaled to more complex scenarios and diverse emotional spectra. This method relies on predefined emotional reference profiles, each associating an emotion term with distinct, specific trends in the experienced values. The post-hoc nature of this process readily allows for iterative refinements to account for individual or cultural biases.

Finally, we demonstrate the framework’s application in a classic RL environment, training an emotional model that spontaneously identifies and learns eight basic, recognizable emotions from past experience of the agent in an unsupervised manner. The resulting emotional spectrum is validated through an emotion attribution survey on new sequences, where the emotions elicited by the agent, unknown to the participants, convincingly align with human subjective observations and experimental psychology accounts.

Chapter 2

Objectives

2.1 Overall objectives and scope of this research

2.1.1 Hypotheses

This research posits and investigates the following interrelated hypotheses concerning artificial emotions, each building upon the findings of the preceding:

- H1: **Mathematical describability of emotions.** *All recognizable basic emotions—such as anger, satisfaction, or fear—correspond to distinct temporal patterns perceived in crucial values for a living being, including recent rewards, expected future rewards, and anticipated world states.*
- H2: **Spontaneous encoding of the emotional spectrum.** *Such patterns can spontaneously emerge and be learned in an unsupervised way from first principles in simple artificial intelligence agents, based on their current state, their recent temporal course, and their temporal projections or expectations.*
- H3: **Measurable utility of synthetic emotions.** *Incorporating learned emotions into an agent’s state can increase its expected utility, yielding benefits analogous to those observed in nature, such as improved behavioral responses, more efficient learning, and enhanced social competencies.*

Hypotheses H1 and H2 are thoroughly investigated both theoretically and experimentally in this research, while hypothesis H3, whose full investigation is beyond the scope of this work, is addressed and analyzed from a theoretical perspective.

2.1.2 General objectives

In order to validate or falsify hypotheses H1 and H2, this research primarily aims to:

- O1: Define and develop a **generic methodological framework**, both theoretical and experimental, that describes and delineates the role that synthetic emotions can play in the field of Reinforcement Learning.

- O2: Empirically validate the **applicability of the framework through a case study**, verifying the results against external references provided by independent human observers and documented experimental accounts in psychology literature.

Objectives O1 and O2, which are central to this thesis, are comprehensively addressed in the subsequent chapters.

With regards to hypothesis H3, which falls beyond the scope of this work, the proposed framework does address the theoretical integration of learned emotions within standard Reinforcement Learning (RL) architectures, as well as the methodology for their utilization and interpretation in both online and offline settings. Its actual implementation would test whether synthetic emotions can enhance purely cognitive or rational processes (such as decision making and learning) by providing measurable utility to AI agents in environments characterized by dynamics similar to natural ones, including stochasticity, partial observability, multiplicity of goals and limited resources. Additionally, the implementation would explore how synthetic emotions can promote beneficial interactions between agents in their environment, such as cooperation and competition, analogous to the ones observed in animal associations as a response to environmental pressures.

A future objective addressing hypothesis H3 could be formulated as follows:

- O3: [*Out of scope*] Empirically validate the **utility of learned emotions through a case study**, demonstrating significant improvements over a non-emotional agent in one or more of the following areas: individual performance, such as achieving higher or more stable rewards; learning efficiency, including faster learning, reduced sensitivity to hyperparameters, or better generalization; or multi-agent cooperation, such as an increase in total reward.

2.1.3 Specific objectives

The theoretical and experimental framework must define a general and coherent methodology of empirical nature for the learning of functional artificial emotions in agents with Reinforcement Learning, supported by experimental evidence of its applicability. This methodology must include the following concrete deliverables:

- A computational and mathematical **definition of artificial emotion** as an automatically synthesized product, based on the data collected during the agent’s training in an RL environment.
- An **algorithm for the training of emotional encoders**, capable of learning artificial emotions as defined above that, based on first principles of RL, is not dependant on the environment.
- An **algorithm for the elicitation of synthetic emotions**, capable of generating emotional responses in real-time as agents interact with the environment.
- An **algorithm for the interpretation of synthetic emotions**, capable of associating elicited emotions to emotion terms for human interpretation, based on predefined reference profiles of basic emotions consistent with those described in psychology.

- A **statistical study of the emotional spectra** obtained, as well as an analysis of their dynamics and congruence with on-going environmental developments.
- A **methodology for the validation of experimentally learned emotions** grounded on external references from independent human observers.

2.2 Possible future objectives

We believe that the validation of the hypotheses addressed in this research and the achievement of its central goals will establish a formal foundation on which further research might address more ambitious objectives. In addition to objectives O1, O2 and O3, centered in enhancing the performance of AI agents in RL environments, the principles and methods defined here could extend to the following related fields:

- **Affective computing:** By directly formalizing and integrating synthetic emotions within practical agents, this approach would extend the original scope of the field, with huge potential for its core goal of effectively enhancing human-machine interactions with emotional intelligence.
- **AI agents interpretability:** The spontaneous emergence of emotions in AI, evidenced in apparently unrelated tasks—like next character prediction [138]—and suggested in the field of RL as a byproduct of reward maximization [139], cannot be ignored. A generic emotional framework could provide interpretability tools for identifying and understanding such emergent emotions, which conventional “black-box” deep learning architectures cannot natively support.
- **Artificial empathy:** Looking ahead, the insights gained from this research could be extended to the development of computational models of empathy. These models would predict the emotions of other agents or humans by utilizing the perceived and estimated states and values of those entities, as interpreted by the observing agent. This application has the potential to enhance the responsiveness and sensitivity of AI systems in interactions involving emotional understanding.
- **Enhanced interactive entertainment experiences:** The definition of virtual characters endowed with realistic, fully integrated emotional reactions, grounded on the on-going progress of the game, might yield unprecedented levels of realism and engagement to the industry.
- **Other fields:** Finally, as originally suggested by the foundation works in the field of Affective Computing [87], the application of emotional AI could benefit various sectors, including computer-assisted learning, perceptual information retrieval, arts and entertainment, as well as human health and interaction.

Further details and discussions of these applications and their implications are more thoroughly covered in chapter *Conclusions*.

Chapter 3

A generic self-learning emotional framework

3.1 Inspirational background

Drawing on insights from the fields of neuroscience, psychology, and biology, we approach the problem of formalizing and learning artificial emotions from a functional and information-theoretic perspective. Thus, we analyze the end-to-end dynamics that transform original perceptions into the cognitive products we know as “emotions”. We examine their triggers, intensity, how they sequentially occur—whether they overlap, replace each other or attenuate over time—and, critically for their synthesis, the cognitive abilities that each requires.

To address this apparent complexity, we draw guidance from the following foundations, documented in nature, that play a role in emotional phenomena, so as to identify parallelisms with existing AI frameworks and methods. Our goal is to reconcile, integrate, or extend these foundations to achieve a functional understanding and emulation of emotions, rather than their exact low-level replication:

1. **Perception:** Cognition originates in what we perceive and feel. Neuroscience describes how animal perception stems from sensory neurons capturing external and internal stimuli, like vision through retinal photoreceptors or pain through nociceptors in skin and body tissues. This sensory information travels via neural pathways to the central nervous system where, in combination with previous experiences, is processed into increasingly abstract concepts, structuring a perceived reality into an individual’s internal representation [140], crucial for emotional phenomena [141].
2. **Reward and pain signals:** Among these cognitive products, the ones we identify as “emotions” are linked to the limbic system—involving structures like the hypothalamus, amygdala and parts of the cortex in humans. Closely connected to it, the reward circuit is responsible for the reward responses which, along with pain signals from the somatosensory system, provide the “common neural currency” [142] guiding animal behavior. Both are central to the emotional experiences, serving as the evolutionary compass toward achievement, survival and reproduction [143]. These subjective signals,

our referential scale for what feels “good” or “bad”, are shaped by individual homeostatic dynamics, maintaining the organism’s internal equilibrium [144][145].

3. **Retrospection:** Past experiences heavily influence an individual’s emotional state, strongly correlated with recent positive or negative outcomes [146][29], as well as with the perceived sign of trend changes, which are equally relevant [147]. More subtle processes involve “habituation” over time via repeated exposure to a reinforcing stimulus [148], as well as “extinction” via neutral exposures to previously valenced stimuli [149].
4. **Anticipation:** Critically for their survival, animals can predict the probability and variance of future rewards based on their perception and memories—which has been shown to be coded in neurons in the basal ganglia, midbrain, parietal, and parts of the cortex in humans [150]—and is associated with dopamine-producing neurons [151]. It is revealing that the dopamine neurotransmitter, a mediator of reward prediction learning [152], is prevalent in all animal phyla, (except for its equivalent octopamine in phylum Arthropoda [153]).
5. **Knowledgeability:** Some emotions may be associated with cognitive representations of states of the environment [154] (such as surprise, curiosity or confusion) or with elements judged as beneficial or harmful for an animal—objects, places or other living beings—based on its own subjective reality [155][156].
6. **Feedback mechanism:** Lastly, emotions can act as signals, communicating internal states to the external environment and influencing interactions—such as fear or anger, signalling threats to others and prompting specific responses [1][6][157]. Internally, emotions can also guide behavior adjustments, maintaining balance and achieving goals [3][5].
7. **Cognitive gradation:** All of these factors, integral to the known emotional psychodynamics, engage specific brain regions, and are therefore conditioned by each species’ unique cognitive abilities. This has suggested a gradation or genealogy of emotions corresponding to the advancement of these abilities throughout evolution alongside with the development of increasingly complex brains, as documented in biology [1][44].

3.2 Introduction to the proposed framework

In alignment with this foundational background, we propose here a generic, fully self-learning emotional framework for AIs. Grounded in first principles from Reinforcement Learning (RL), this framework enables any agent interacting with an environment to automatically learn, elicit, and utilize its own synthetic emotional spectrum, convincingly resembling natural emotions described in the literature. An overview of the framework is explained here, while its formal description and detailed methodology can be found in *Theoretical framework*.

The framework is based on the following fundamental hypothesis:

Hypothesis: *All emotions correspond to distinct temporal patterns perceived in crucial values for a living being, such as recent rewards, expected future rewards or anticipated world states.*

Furthermore, given that said crucial values, generated as cognitive variables, are determined by the individual’s specific cognitive abilities, they condition the complexity of the emotions experienced. The most basic emotions reflect trends or patterns in reward / punishment signals, while increasingly sophisticated emotions integrate their subjectively anticipated values, anticipated world states, associations with other individuals or objects, etc. (Unlike in psychology or neuroscience, the term “reward” encompasses positive and negative outcomes in RL, rarely using “punishment” for the latter, a convention that we also follow henceforth.)

This assumption suggests the viability of AI agents in the field of RL automatically learning such patterns too, based on historical information—or subjective experiences—which is the principle underpinning the self-learning framework introduced here.

Example:

To understand the correspondence between said temporal patterns and natural emotions, we introduce the main components of the proposed framework with an illustrative example of a simplified bioinspired RL setup (see Fig. 3.1). An AI agent’s goal is to survive in an environment where energy sources—the reward signal—are scarce and disputed with other agents. Its energy slowly diminishes over time (for example, average reward ≈ -0.1), but it can perceive its nearby environment as a *state*, and is endowed with simple *actuators* for displacement, feeding and combat. Its cognitive abilities include a short-term—or *replay-memory*, a state-based prediction of future rewards—or *state-value function*—and a *policy* defining its behavior. As shown in the figure, the agent observes sequences of recent and predicted rewards during the simulation, producing differentiated temporal patterns as sequences of multivariate time series (MTS). We now describe how such patterns can be associated with emotions.

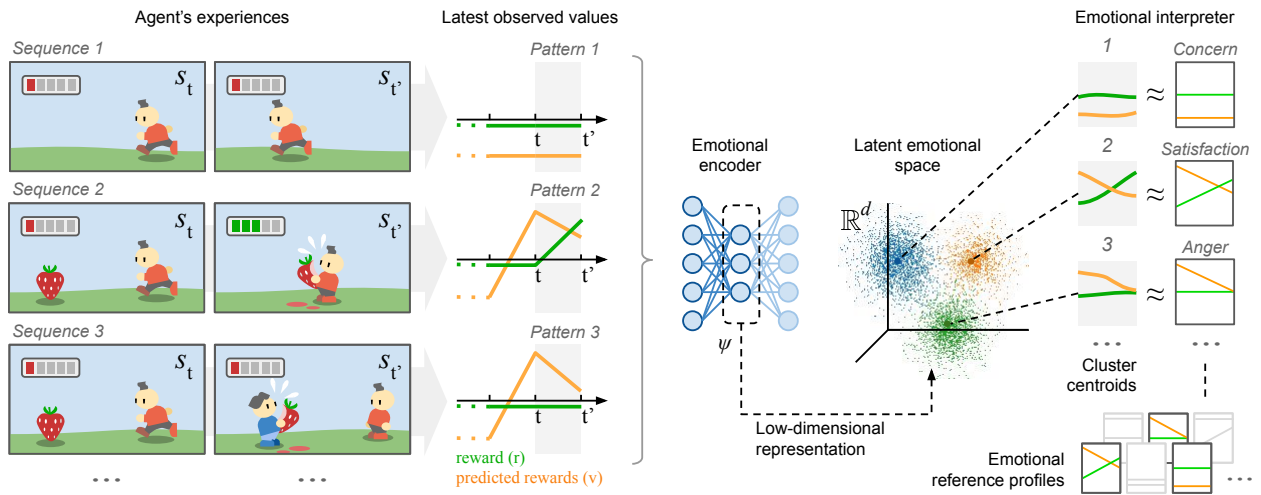


Figure 3.1: Overview of the framework: How emotions can be learned from experiences, then elicited and interpreted. A reinforcement learning agent’s interactions are registered as sequences of multivariate time series (instantaneous *reward* and future predicted rewards—or *state-value*—in the example). An emotional encoder (such as a deep autoencoder) is trained unsupervised on the sequences, encoding them into a low-dimensional latent space. Distinct dynamic patterns in the sequences emerge as clusters whose centroids can be mapped to known emotional reference profiles (showing three here for simplicity).

During a continuing successful execution without foreseeable hurdles, the agent would observe a series of positive rewards along with equally positive predictions, which might correspond to happiness. Contrarily, a disastrous and irreparable situation—such as a damaged actuator preventing progress—would yield a series of negative rewards and equally negative predictions, which might correspond to distress. Some other basic emotions might be equally described from similar stable patterns, or *emotional reference profiles*, like concern or optimism (top rows of Table 3.1).

Table 3.1: Association of some emotions to latest observed values.

Latest rewards	Latest predictions	Emotion	Rationale
positive	positive	<i>happiness</i>	Successful execution, no foreseeable hurdles
negative	negative	<i>distress</i>	Poor execution, no foreseeable improvements
average	negative	<i>concern</i>	Regular execution, expected to worsen
average	positive	<i>optimism</i>	Regular execution, expected to improve
...
average	increased	<i>excitement</i>	An opportunity emerged during regular execution
decreased	average	<i>frustration</i>	A reward is no longer being received
...
average	decreased-to-average	<i>anger</i>	A threat arose, seen as potentially addressable
average	decreased-to-negative	<i>fear</i>	A threat arose, seen as hardly addressable

Crucially, within the framework, values are deemed *positive*, *average* or *negative* based on their comparison with historical observations, thus capturing the pivotal role of homeostasis in emotions (for instance, a positive reward that is substantially lower than the average reward would be considered as negative).

Other emotions can be associated with the dynamics of change, corresponding to recent trends of increases or decreases in these values, for example excitement and frustration (middle rows of Table 3.1).

Finer interpretation of temporal trends allows the differentiation between closely linked emotions like anger and fear (bottom rows of Table 3.1), where cognitive appraisal of the agent’s control on its chances to overcome the challenge reflects a slight or steep drop respectively (Yang et al, 2018) [158]:

We introduce the following additions to the classic RL setup (see Fig. 3.1):

- an *emotional encoder* (or emotional model) which, trained on various MTSs experienced, learns its latent features as a low-dimensional representation $\Psi_t \in \mathbb{R}^d$ representing the emotional state at time-step t ;
- an *emotional interpreter* which, trained on the value distribution of ψ over the latent space \mathbb{R}^d , maps its values to emotion terms for human interpretation, based on known reference profiles.

The use of the emotional encoder within an extended RL architecture allows the emotional agent to dynamically elicit instantaneous emotions, enriching the state used by an emotionally-enabled policy with a subjective, emotional state (Fig. 3.2).

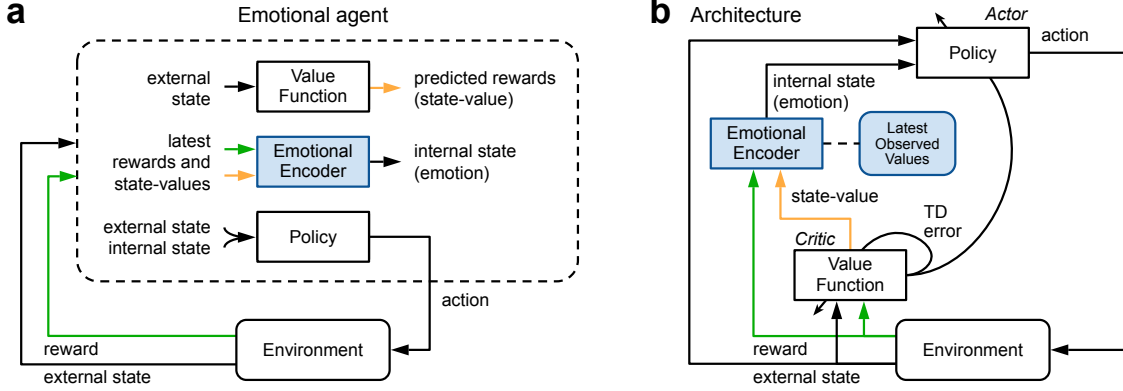


Figure 3.2: Elicitation of learned emotions. **a, Elicitation.** Once trained, the agent’s emotional encoder dynamically encodes ongoing sequences of observed values (such as *rewards* and *state-values*) into instantaneous emotions. This extends the perceived state, enriching the policy input with an internal emotional state. **b, Architecture.** Extension of the actor-critic architecture with an emotional encoder that stores the latest observed values (rewards and state-values). Unlike the actor and critic, its training is not necessarily driven by Temporal Difference (TD) errors.

For interpretability, the emotional interpreter can be used during the execution of the task to map instantaneous emotional states to known terms. Fig. 3.3 illustrates how a longer succession of events and their temporal patterns is associated with a series of coherent consecutive emotions. Richer emotional sequences in an actual RL environment, covering entire episodes, are discussed in chapter *Results*.

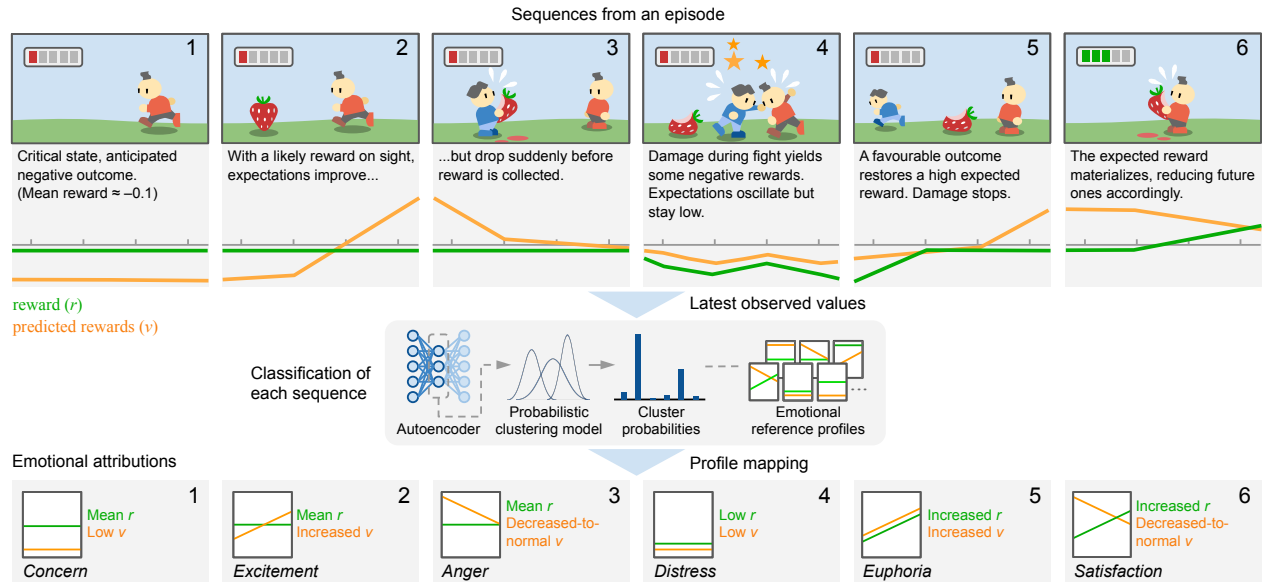


Figure 3.3: Interpretation of the learned emotions. Instantaneous emotions can be mapped to known referential profiles for external interpretation or communication to other agents. A probabilistic clustering model (such as Gaussian mixture) can predict their distribution probabilities across clusters, associating them to the clusters’ preassigned profiles. The example interprets six consecutive profiles out of the 30 combinations in a *LOVE 2:5x6* interpretability mapping (*LOVE: Latest Observed Values Encoding*, with two observed values that have 5 and 6 possible patterns respectively; see 3.4.3 and Fig. 3.4 and Fig. 4.6 for details).

This illustrative example outlines the core ideas of the framework based on a limited set of cognitive values, namely, reward and predicted rewards (or state-value). Agents endowed with higher cognitive capabilities, such as a world model anticipating future states or external associations, are likely to develop more complex emotions, which belong to higher emotional orders. These concepts are discussed in section *Theoretical framework*, where we provide detailed definitions formalizing the concepts of emotion, emotional encoder, emotional spectrum, and emotional orders (for example, the agent in Fig. 3.1 is categorized as Order III). Additionally, we describe the detailed methodology subsequently applied to a case study.

In conclusion, we have introduced how the proposed framework can capture the natural foundations described above, extending the RL framework to allow the learning, elicitation and utilization of synthetic emotions, as well as their external interpretation. The framework integrates objective perception (external state and rewards) with internal, subjective appraisals and the homeostatic definitions of *average*, based on past experiences. The case study included in *Results* illustrates how this synthetic emotional system naturally reproduces other well-known dynamics, such as elicitation / decay, coexistence, alternation, subjectivity, environment-dependency, and confusion.

3.3 Theoretical framework

We cover here the theoretical formalization and primary definitions of the introduced framework as well as its integration within the RL framework, before addressing a detailed description of its application within a generic methodology.

3.3.1 Foundational concepts from Reinforcement Learning

In the classic Reinforcement Learning setup, an agent learns how to maximize the amount of reward received through interactions with an environment over a sequence of discrete time-steps. The actions influence not just the immediate rewards, but also the subsequent states of the system, with an impact on future rewards. The environment is formulated as a Markov Decision Process, satisfying the *Markov property*, that is to say, its response at time $t + 1$ depends only on the state and action taken at t [61].

The dynamics of this setup yield a series over time with the form: $S_0, A_0, R_1, S_1, A_1, R_2, A_2 \dots$ where S_t is the state observed by the agent at time-step t , A_t is the action then taken, and R_{t+1} is the reward observed after the action. We'll follow the classic notation in which capital letters are used for random variables, whereas lower case letters are used for the values of random variables (such as s, a, r , etc.).

The action A_t is chosen by a policy π based on state S_t . The literature offers a broad variety of methods to train π , but here we'll focus on the *actor-critic* methods [159], a subset of policy-gradient methods. These methods simultaneously obtain a policy π (the actor) and a value function v (the critic), whose learning is based on the temporal difference error, an error mechanism that has been compared by neuroscience to dopamine's in animal learning [160]. In this family of methods, widely used in state-of-the-art applications, v yields state-values, an estimate of future rewards, whose sign and magnitude are key for emotional

elicitation. Significantly, the actor-critic model has been compared with the mesolimbic brain, a system that plays a crucial role in emotional processing and regulation across a wide range of vertebrate species. Specifically, the ventral and dorsal striatum areas have been associated with the critic and actor components of reinforcement learning models, respectively [161].

These are the elements taken from the framework, whose detailed formulations are thoroughly studied in the literature [61]:

- $S_t \in \mathcal{S}$ is the state observed at time t , from the set of states \mathcal{S} ;
- $A_t \in \mathcal{A}$ is the action taken at time t , from the set of actions \mathcal{A} ;
- $R_t \in \mathcal{R}$ (a subset of \mathbb{R}) is the positive or negative reward observed at time t ;
- $\pi(a | s) \in [0, 1]$ is the probability of choosing action a in state s ;
- $v_\pi(s) \in \mathbb{R}$ is the value of state s under policy π (or expected future rewards).

3.3.2 Primary definitions for the framework

In addition to the above, the following necessary definitions and components are introduced, whose possible implementations are separately addressed and extended in *Methodology*:

Emotions. *Def.: The instantaneous latent representations, or encodings, of temporal patterns in vital values recently observed by the agent.*

Said patterns can reflect significant trends in current, past and predicted values, endowing the agent with a meaningful, on-going appraisal on their progression. For simplicity, the forthcoming exposition will focus on reward and state-value, briefly discussing promising alternatives further down.

Analogous to S_t (the external state observed at time t) we'll use the following notation:

$\Psi_t \in \mathcal{E}$ is the emotional state at time-step t , from the set of emotional states \mathcal{E} .

Emotional encoder. *Def.: A model that takes as input the multivariate time series sequence formed by the latest values observed by the agent, generating a low-dimensional encoded representation of their recent dynamics, namely, the above-mentioned emotion.*

Its input would take the form of a multivariate time-series:

$$O_{t-T+1:t} = \{O^1, O^2, \dots, O^W\}_{t-T+1:t}$$

where:

- W is the number of observed values;
- T is the length of the sequence, an arbitrary value that defines the emotional window of the agent;
- $O_{t-T+1:t}^w$ is the time series sequence (w_{t-T+1}, \dots, w_t) of the latest T values observed for value w at time-step t .

In the sample case with two values, the sequence takes this form:

$$O_{t-T+1:t} = \{(R_i), (V_i)\}_{t-T+1:t}$$

where:

R_i and V_i are, respectively, the reward and estimated state-value observed at time-step i within the latest T values at time-step t ;

and the emotional encoder is formulated like this:

$$\Psi_t = e(\{(R_{t-T+1}, \dots, R_t), (V_{t-T+1}, \dots, V_t)\})$$

where:

- e is the emotional encoder (for example, a deep autoencoder);
- $\Psi_t \in \mathcal{E}$ is a point in the latent emotional space \mathbb{R}^d of dimension d , from the set of emotional states $\mathcal{E} \subset \mathbb{R}^d$, representing the emotional state at time-step t ;
- d is the dimension of the latent space learned by e , a hyperparameter of the model.

Emotional window. *Def.: The lapse before which no observation has an emotional impact, imposed by the length of the input sequences processed by the emotional encoder, taken from the agent's experience.*

Its choice is arbitrary, but too low or too high values might render the events within the time window unstable or irrelevant respectively. As a guidance, the analysis of the most basic natural emotions typically considers seconds or even minutes, rather than hours, just enough to keep relevant events within the short-term memory [146][29].

Emotional spectrum. *Def.: The range of possible values that the emotional state Ψ_t can take within the latent space \mathbb{R}^d , denoted as $\mathcal{E} \subset \mathbb{R}^d$.*

It depends on: (1) the experiences from which it has been learned, namely, the environment and the actions taken in it; (2) the choice of values integrated in Ψ_t , limited by the agent's cognitive abilities, or emotional order. Despite taking continuous values, the distribution of the random variable over the latent space may naturally tend to follow non-uniform probabilistic distributions, amenable to interpretation.

Emotional orders. *Def.: A stratification of learnable emotional spectra of increasing cognitive complexity, determined by the set of values integrated in Ψ_t , which is limited by the agent's cognitive abilities.*

The first, more basic orders, describe the gradation from the fleeting, irreflexive experience of instantaneous reward to multifaceted blends of expectations and recent events originated in the value function, replay memory and world model. For their representation in Fig. 3.4 we introduce graphically their corresponing Latest Observed Values Encodings (or LOVE patterns), that are formally defined in *Interpretation of the learned emotions*.

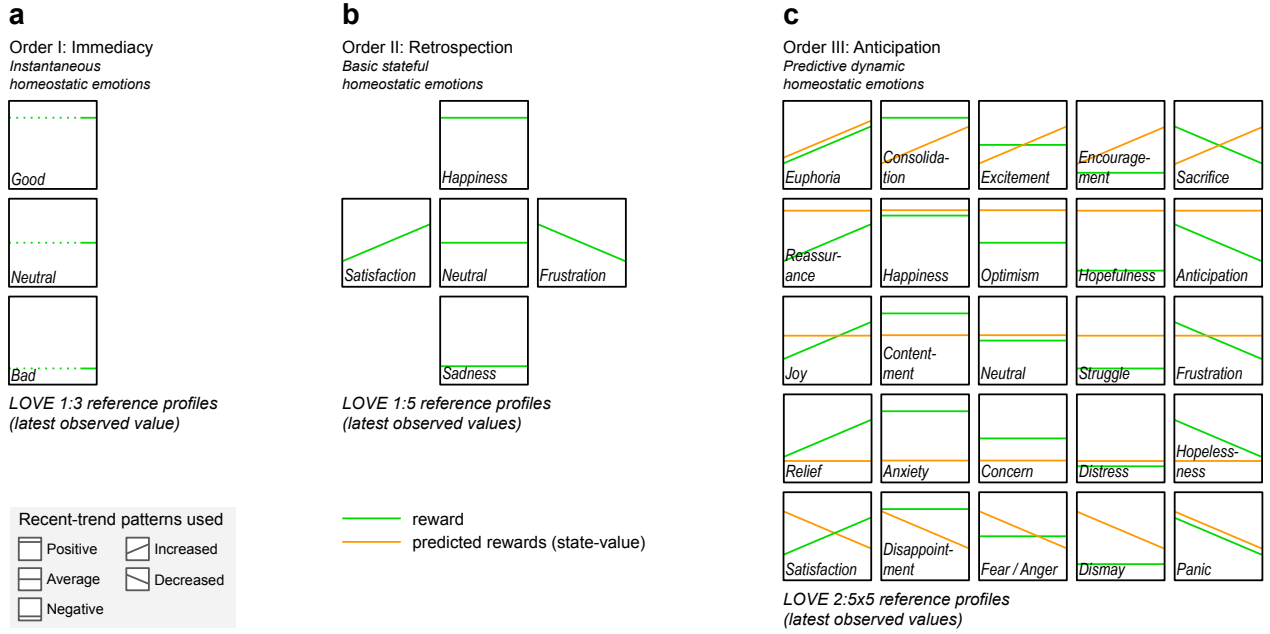


Figure 3.4: How cognitive abilities determine the emotional spectrum. A gradation of emotional spectra arises from increasingly higher cognition based on Latest Observed Values Encodings (or LOVE patterns) relative to subjective historical averages. The x-axis represents time, and the y-axis represents value magnitude. **a, Order I. Immediacy:** a single instantaneous *reward* value defines the simplest spectrum. **b, Order II. Retrospection:** a short-term memory of the latest rewards allows elementary emotional dynamics. **c, Order III. Anticipation:** the ability to predict future rewards (*state-value*) defines a much richer learnable spectrum of predictive dynamic emotions, which is used in the case study.

Order 0: A standard RL agent; cognition allows action selection (by the policy π) based on sensory information (the state s). No emotions are elicited.

Order I - Immediacy: Taking *only the current value* of the reward defines the simplest learnable spectrum, with instantaneous homeostatic emotions such as good, bad or neutral (for positive, negative and average values respectively).

Order II - Retrospection: A simple short-term memory of the latest rewards endows the agent with elementary emotional dynamics, extending the learnable spectrum with basic stateful homeostatic emotions reflecting the latest trends (for example, satisfaction, frustration, happiness, sadness, or neutral).

Order III - Anticipation: Including the recent state- or action-values predicted by a value function can signal meaningful trends in future expectations, defining a much richer learnable spectrum of predictive dynamic homeostatic emotions (including anger, fear, concern, excitement, frustration, euphoria, etc.). RL algorithms that learn a value function are suitable here, such as SARSA, Q-Learning, actor-critic, etc. [61]; we will focus on the latter in this study.

Order IV - World-knowledgeability: Agents that learn a world model (not necessarily for planning) can predict the likeliest state continuations, detecting mismatches through state-prediction errors that can elicit basic knowledge-related emotions (such as surprise, curiosity, boredom), as well as valenced emotions toward objects or places (including aversion, attraction, phobia or fondness).

Higher Orders: Agents endowed with higher cognitive abilities, like longer-term memory or more complex world- or self-models, could learn and experience higher-order emotional spectra, like remote retrospections, social, moral or self-conscious emotions, which lay beyond the scope of this work.

A detailed analysis of the specific learnable emotions for each order requires the introduction of recent-trend patterns, discussed in section *Interpretation of the learned emotions*.

Emotional interpreter. *Def.:* A system that takes as input an emotion encoded by an emotional encoder, mapping it to known emotion terms for human interpretation.

The input would be $\Psi_t \in \mathbb{R}^d$, representing the emotional state at time-step t , and the output the attributed emotion term (for instance, “joy”, “frustration”, “relief”), similar to the examples in *Introduction to the framework*.

3.4 Methodology

Here, we present how the framework operates, detailing a generic methodology for the spontaneous encoding of the emotional spectrum within an RL environment from raw agent’s experience which, based on the definitions introduced, provides the following functionality:

1. Learning emotions from experience;
2. Elicitation of emotions and integration within an RL architecture;
3. Interpretation of the learned emotions.

A fully detailed, step-by-step description of its application on a real case study is covered in chapters *Results* and *Methods and tools*.

3.4.1 Learning emotions from experience

Training of an emotional encoder

The simplest approach to training an emotional model (for example, a deep autoencoder) from direct observations of the agent is *offline learning*, following these steps:

1. Train a conventional RL agent A to the desired performance, for instance, with an actor-critic method.
2. Run the trained agent A on new episodes in the environment, saving the stepwise trajectories as MTS of the key values (for example, reward and state-value).
3. Train a deep autoencoder (unsupervised) on normalized trajectory sequences of length T , representing the emotional window.

The offline setting approaches emotional systems as “frozen” models, learned from previous agents’ experiences, comparable to “*super distilled foundation models passed onto us by our*

ancestors, through genetics and evolution,” as emotions have been described in the context of modern generative AI [162]. On top of its lower complexity, this method offers advantages such as stability, consistency and reliability, at the expense of increasingly detrimental mismatches between newly learned policies and the frozen emotional model. This might potentially impact its utility if, for instance, a new agent A' were to be defined and subsequently trained making use of the extended states (external and internal), inheriting the emotional dynamics of A , despite now learning a different policy and value function.

The alternative would be the *online learning* of the emotional model along with the original training of the agent A (for example during the training cycles over the replay buffer of previous experiences), which despite the added complexity, might yield higher-utility emotions. In the real case covered in *Results* we adopted the simplest offline learning.

Input values

The choice of input values for the encoder is arbitrary, but will determine its learnable emotional spectrum. A logical prioritization inspired by nature and psychology, and based on the principles of utility and availability, is suggested here, following the order-wise stratification introduced by the framework (see *Emotional orders* in section *Primary definitions for the framework*):

1. *reward*, as the key RL first principle. The objective value to maximize, defining *good* and *bad* for an agent, and directly received from the environment (but also applicable to less classic approaches like intrinsically-motivated RL [136]).
2. *state-value*, a prevalent cognitive product contributing contextual estimates of “how good” future expected rewards will be. Its anticipatory nature adds valuable subjective charge to emotions like optimism or fear.
3. *state-prediction error*, quantifying the discrepancies between a learned world model and the observed state transitions. It can define emotions driven by an agent’s understanding of its environment like surprise, astonishment or interest.

Beyond these values, related to orders I to IV, other possibilities might be considered:

4. *average reward*, a running estimate of $r(\pi)$, the average reward per time-step of the policy, typically expressed by \bar{R}_t at time t , could provide a good measure of life-long performance and a broader reference of *normality* for valenced emotions relative to its subjective *neutral* or homeostatic value.
5. *moving average reward*, the mean value of rewards observed only during the latest steps, provides a more recent, adaptive reference of *normality*, and could substantiate the emotional dynamics of habituation to valenced emotions relative to its dynamic value, modulated by the length and parameterization set for the averaged period.
6. Other descriptive statistics of all the values above might capture emotionally-relevant dynamics (for instance, a high fluctuation in recent rewards should probably prevail over other values, with a sense of confusion or uncertainty).

7. Finally, *temporal difference error*, or TD error, estimating instantaneous step-wise errors in the expected future rewards, is applied all over the RL literature. However, it is by definition a byproduct of reward and value estimates [61], and therefore redundant. Furthermore, as discussed in section *The TDRL theory of emotion and related approaches*, whenever predictions match observations well, TD provides little to no information regardless of the success or failures of the agent.

Model architecture

A natural choice for the unsupervised learning model that encodes recent sequences into a low-dimensional latent space is a deep autoencoder (DAE) [163]. This type of artificial neural network is extensively applied to learn compact feature spaces, which have been empirically demonstrated to capture existing similarities and relations among the original samples [164]. While the generative properties of the more sophisticated approach, variational autoencoders [165] are not required for the core task, interesting sequence-based alternatives like Recurrent Neural Networks [166][167] (dispensing with the fixed-size time window) or state-of-the-art self-attention models [55][56] might be considered in future works.

For the case of a DAE, since the purpose of the model is not input regeneration or denoising, an excessively high encoding dimension d , or too high depth and complexity of the architecture, might harm its purpose to capture high-level trends and magnitudes of the observed values.

3.4.2 Elicitation of emotions and integration within an RL architecture

Figure 3.2 in *Introduction to the proposed framework* illustrates how a trained emotional encoder can be used by an agent to dynamically enrich its policy’s input with an emotional state. The following definitions are proposed:

Extended state *Def.:* The extension of the state S_t observed from the environment at time-step t , or external state, with the emotion Ψ_t elicited by the emotional encoder, or internal state, with the form $X_t = (S_t, \Psi_t) \in \mathcal{S} \times \mathcal{E}$.

The extended state X_t can produce a richer representation that blends the objective and subjective perception of the agent at each step.

Emotional agent *Def.:* An RL agent whose policy makes use of an extended state X_t .

Its policy π would take this form:

$$\pi(a \mid x) \in [0, 1] \quad \text{is the probability of choosing action } a \text{ in extended state } x \in \mathcal{S} \times \mathcal{E}.$$

In section *Introduction to the proposed framework* we propose an architecture for such an emotional agent, based on the actor-critic method, that utilizes the extended state. While its implementation is not included in this work, we provide the pseudocode for its training in a continuing task for agents of Order III and IV in section 3.4.4, *Extensions of the actor-critic method*.

3.4.3 Interpretation of the learned emotions

Finally, we detail here the principles to create a generic emotional interpreter. While interpretability of the learned emotions is not imperative for their utilization by an emotional agent, it significantly contributes to their validation and analysis, and potentially facilitates the external communication of the instantaneous emotional state. A few new concepts and tools are required for this purpose.

Clustering of the emotional spectrum

The distinct dynamic patterns learned by the emotional encoder will translate into a non-uniform distribution of ψ over the latent space, emerging as clusters whose centroids represent their respective prototypical sequences. The number and diversity of these clusters depends on many factors, including the clustering method chosen, but once learned, each centroid's sequence can be mapped to known *emotional reference profiles* associated with prototypical emotions described in the psychology literature, allowing human interpretation, as shown in *Introduction to the proposed framework*.

Selection and validation of the interpretability mapping

The applicable interpretability mapping is determined by the emotional order of the agent, and will allow the interpretation of the instantaneous emotions. The following definitions are proposed:

Emotional reference profile Def.: *A possible combination of recent-trend patterns of the observed values in the corresponding emotional order associated to some known emotion term.*

For instance, in Order III, where values include rewards and state-values (or expectations for more clarity), the profile <average rewards, positive expectations> might correspond to optimism. Other examples might be:

Latest observed trends	Emotion term
<average rewards, negative expectations>	<i>concern</i>
<negative rewards, negative expectations>	<i>distress</i>
<increased rewards, positive expectations>	<i>reassurance</i>
etc.	

Interpretability mapping Def.: *A set of emotional reference profiles formed by the different possible combinations of their recent-trend patterns, ideally encompassing an ample range of emotion terms.*

The following naming convention is followed for mappings:

$$\text{LOVE } N_{\text{values}} : N_{\text{profiles}}$$

where *LOVE* stands for *Latest Observed Values Encoding*. For example, a LOVE 2:5x5 mapping would correspond to two values (like reward and state-value) mapped over 25 profiles (the combinations of five patterns per value).

In *Introduction to the proposed framework* we showed possible interpretability mappings for the first three emotional orders, where each possible profile is associated with an emotion based on the interpretation of its recent-trend patterns (see Fig. 3.4). The set of recent-trend patterns defined for observed values is arbitrary, but a few basic ones inspired in psychology render significant variety and representativeness to the mapping:

average: Latest observed values do not differ much from their *historical average*;
positive: Latest observed values are *higher than average*;
negative: Latest observed values are *lower than average*;
increased: Latest observed values reflect a *positive trend*;
decreased: Latest observed values reflect a *negative trend*.

The three first patterns reflect the well-known principle of subjective “normality” versus “exceptionality” associated with homeostasis [144][145]. As for the emotional effect of recent changes perceived, research shows how trends may be more relevant than actual magnitudes (for instance, a minor improvement can bring happiness to the unwell, yet a minor ailment may frustrate the healthy [147][168]), a principle that is captured in the two last patterns.

However, since the encoding of the latest values yields continuous values, specific criteria must be established for classification based on their statistical nature. For example: values can be classified as positive or negative when the mean value falls out of the range $[-k\sigma, k\sigma]$, where σ is the standard deviation of the values and k is a factor defining the “average” interval. Additionally, values can be classified as increased or decreased when the slope of the linear regression of the sequence deviates from its average value in the observed distribution by a specified amount, such as a fraction of its own standard deviation.

Profiles can nonetheless be extended for more nuanced interpretations if needed; for example, in Order III, the differentiation of *decreased* into two patterns (decreased-to-average, decreased-to-negative) for state-value results in 30 profiles (LOVE 2:5x6), and captures the relevant distinction between fear and anger respectively (see Fig. 4.6 in *Results*). (The rationale behind this distinction is based on the psychological interpretation of the subjective appraisal attributed to these two emotions [169]: while both are classified as *negative* in valence and associated with potentially adverse future outcomes, *anger* is linked to more individual control over the threat and relatively higher certainty than *fear*, which would respectively reflect into less steep declines in state-value for *anger* than for *fear*.) Further possible extensions are briefly mentioned in *Conclusions*.

Attribution of emotion terms

The resulting profiles of each mapping can then be associated with known prototypical emotion (or affect) terms through analysis. However, given the lack of a broad consensus in the number, definition and categorization of emotions or affects in the literature [170][171][16],

the following principles, partially inspired in the most influential theories, were applied to denominate the mappings of Orders I to III showed in *Introduction to the proposed framework*. Further research will possibly contribute refinements to these mappings:

- Conformity: Initial analysis of the value trends in each profile (magnitude, valence, stable / dynamic) for comparison with emotion descriptions in the relevant literature (including authoritative emotion/affect models, dimensions, lists and stability / transience).
- Meaningfulness: Prioritization of the most broadly adopted terms in the field, deprioritizing the uncommon.
- Comprehensiveness: Maximization of the overall emotional spectrum’s range of each mapping.
- Cognitive adequacy: Contextualization within the respective cognitive orders (from basic, simple concepts in emotional Order I, to finer, richer emotions in Order III). Homogenization of the emotional spectra, not mingling complex affects (such as social, moral or self-conscious) with primary, instantaneous emotions, solely elicited by the recent events.
- Theoretical sequential coherence: Offline simulations of event-guided emotional sequences and review till fully natural transitions are obtained in all cases and all profiles visited several times (see *Theoretical validation of LOVE profile terms* for details).
- Experimental sequential coherence: Finally, adaptation of the emotion terms to match the nature of the environment-agent interactions in the actual case study through the review of real emotional sequences.

Example: The pattern tagged as *satisfaction* is characterized by an increase in rewards, along with its corresponding decrease in future expected rewards or state-value, as described for example in (Schultz, 2015) [152]: “*Value informations for choices need to be updated when reward conditions change. For example, food consumption increases the specific satiety for the consumed reward and thus decreases its subjective value while it is being consumed.*”

Real-time emotional interpretation

Once the learned clusters have been associated to profiles from the LOVE mapping chosen, the instantaneous emotion encoded can be dynamically classified by the clustering model used and accordingly interpreted through its preassigned profile. If a probabilistic clustering model was used (for example, a Gaussian mixture model), a probability distribution will be obtained which, on top of the likeliest emotion, will capture richer nuances (for instance, 70% neutral, 30% fear may be interpreted as *slight fear*).

The stepwise stability of these classifications may be impacted by the different noisy values it is based on (environment rewards, model predictions, encoding model), which can be addressed by different techniques (such as smoothing moving average, reclassification thresholds, etc.). (See 4.1.3 for real-time interpretations in a concrete use-case.)

3.4.4 Extensions of the actor-critic method

We document here the pseudocode for training Order III and Order IV emotional agents, based on the architecture introduced in *Introduction to the framework* (the significance of Orders I and II is more theoretical than practical). We followed the notation and original pseudocode taken from (Sutton, 2018) [61]. For the sake of brevity, we focus on the continuing task setup—which will apply better to future biologically-inspired natural setups—but the extensions can be applied to all the variants, such as episodic tasks, eligibility traces, etc.

Training of a classic actor-critic agent (Order 0 - Non-emotional agent)

As a reference, we first include *Algorithm 1*, the original pseudocode for training the policy and the state-value function of a vanilla actor-critic agent for continuing tasks (based on average reward estimates, and without the classic *gamma* discount factor) [61]. This corresponds to a non-emotional agent (emotional Order 0). In the pseudocode provided:

$S, S' \in \mathcal{S}$	are the states observed at time-steps t and $t + 1$ from the set of external states \mathcal{S} ;
$A \in \mathcal{A}$	is the action taken at time-step t , from the set of actions \mathcal{A} ;
$R \in \mathbb{R}$	is the reward observed at time-step t ;
$\bar{R} \in \mathbb{R}$	is the estimate of average reward at time-step t ;
$\pi(a s, \boldsymbol{\theta}) \in [0, 1]$	is the probability of choosing action a in state s given policy π with parameter vector $\boldsymbol{\theta}$;
$\hat{v}(s, \mathbf{w}) \in \mathbb{R}$	is the approximate value of state s given weight vector \mathbf{w} .

(For greater clarity, the pseudocode does not optimize the number of invocations to the function \hat{v} . This also applies to forthcoming algorithm versions involving other functions.)

Algorithm 1 Training of a vanilla non-emotional actor-critic agent (continuing tasks)

- 1: **Input:** a differentiable policy parameterization $\pi(a | s, \boldsymbol{\theta})$
 - 2: **Input:** a differentiable state-value function parameterization $\hat{v}(s, \mathbf{w})$
 - 3: Initialize $\bar{R} \in \mathbb{R}$ (e.g. to 0)
 - 4: Initialize state-value weights $\mathbf{w} \in \mathbb{R}^{d'}$ and policy parameter $\boldsymbol{\theta} \in \mathbb{R}^{d''}$ (e.g., to $\mathbf{0}$)
 - 5: Algorithm parameters: $\alpha^w > 0$, $\alpha^\theta > 0$, $\alpha^{\bar{R}} > 0$
 - 6: Initialize $S \in \mathcal{S}$ (e.g. to s_0)
 - 7: **loop forever (for each time-step)**
 - 8: $A \sim \pi(\cdot | S, \boldsymbol{\theta})$ ▷ Choose an action
 - 9: Take action A , observe S', R ▷ Observe new state and reward
 - 10: $\delta \leftarrow R - \bar{R} + \hat{v}(S', \mathbf{w}) - \hat{v}(S, \mathbf{w})$ ▷ Temporal difference error
 - 11: $\bar{R} \leftarrow \bar{R} + \alpha^{\bar{R}} \delta$ ▷ Update average reward
 - 12: $\mathbf{w} \leftarrow \mathbf{w} + \alpha^w \delta \nabla \hat{v}(S, \mathbf{w})$ ▷ Adjust \hat{v}
 - 13: $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha^\theta \delta \nabla \ln \pi(A | S, \boldsymbol{\theta})$ ▷ Adjust π
 - 14: $S \leftarrow S'$ ▷ Update state
 - 15: **end loop**
-

Training of an emotional agent (Order III - Anticipation)

An Order III agent includes recent rewards and state-values, and an emotional encoder e , following the architecture introduced in *Introduction to the proposed framework*. *Algorithm 2* provides the pseudocode to train the agent based on e , previously trained offline on MTSs, where:

$R^{(T)} = [R_{t-T+1}, \dots, R_t] \in \mathbb{R}^T$	is the list of the latest T rewards R observed at time-step t ;
$V^{(T)} = [V_{t-T+1}, \dots, V_t] \in \mathbb{R}^T$	is the list of the latest T state-values V observed at time-step t ;
$\Psi = e(R^{(T)}, V^{(T)}) \in \mathbb{R}^d$	is the emotion encoded by e from $R^{(T)}$ and $V^{(T)}$ at time-step t ;
$X = (S, \Psi)$	is the extended state observed at time-step t ;
$\pi(a \mid x, \theta) \in [0, 1]$	is the probability of choosing action a in extended state x given policy π with parameter vector θ ;
T	is the emotional window of the agent.

The implementation of the short-term memory utilizes a circular buffer with the latest T observed values, as the input to e (needed for a fixed-input architecture of e , but not required in other implementations, such as a Recurrent Neural Network, whose hidden state naturally represents past inputs). For the same reason, the first $T - 1$ time-steps require special treatment until the two input lists have accumulated enough values. The simplest approach is to assign an arbitrary initial value ψ_0 , for example, corresponding to a neutral emotion.

Training of an emotional agent (Order IV - World-knowledgeability)

The most basic emotional agent of Order IV would count with an additional learned world model m , anticipating next states from the current state and the action then taken. The emotional encoder would thus observe a third value, measuring the accuracy of the latest predictions, as suggested in *Primary definitions for the framework*, *Emotional orders*.

We provide *Algorithm 3*, a basic pseudocode where:

$S' = m(s, a) \in \mathcal{S}$	is the external state predicted while being in state $s \in \mathcal{S}$ and taking action $a \in \mathcal{A}$;
$diff(s, s') \in \mathbb{R}$	is a measure of the difference between two external states s and $s' \in \mathcal{S}$ (for example, Euclidean);
$D^{(T)} = [D_{t-T+1}, \dots, D_t] \in \mathbb{R}^T$	is the list of latest T values of $diff()$ observed at time t ;
$\Psi = e(R^{(T)}, V^{(T)}, D^{(T)}) \in \mathbb{R}^d$	is the emotion encoded by e from the latest T values of R , V , and D at time t .

Algorithm 2 Training of an Order III emotional actor-critic agent (continuing tasks)

```

1: Input: a differentiable policy parameterization  $\pi(a \mid s, \boldsymbol{\theta})$ 
2: Input: a differentiable state-value function parameterization  $\hat{v}(s, \boldsymbol{w})$ 
3: Input: a trained emotional encoder  $e(R^{(0)}, V^{(0)})$ 
4: Initialize  $\bar{R} \in \mathbb{R}$  (e.g. to 0)
5: Initialize state-value weights  $\boldsymbol{w} \in \mathbb{R}^{d'}$  and policy parameter  $\boldsymbol{\theta} \in \mathbb{R}^{d''}$  (e.g., to  $\mathbf{0}$ )
6: Algorithm parameters:  $\alpha^w > 0$ ,  $\alpha^\theta > 0$ ,  $\alpha^{\bar{R}} > 0$ 
7: Initialize  $S \in \mathcal{S}$  (e.g. to  $s_0$ )
8: Initialize  $\Psi \in \mathbb{R}^d$  (e.g., to  $\psi_0 = \text{neutral emotion}$ )
9: Initialize  $R^{(T)}, V^{(T)}$  (as empty lists)
10: loop forever (for each time-step)
11:    $X = (S, \Psi)$  ▷ Extended state
12:    $A \sim \pi(\cdot \mid X, \boldsymbol{\theta})$  ▷ Choose an action
13:   Take action  $A$ , observe  $S', R$  ▷ Observe new state and reward
14:    $\delta \leftarrow R - \bar{R} + \hat{v}(S', \boldsymbol{w}) - \hat{v}(S, \boldsymbol{w})$  ▷ Temporal difference error
15:    $\bar{R} \leftarrow \bar{R} + \alpha^{\bar{R}} \delta$  ▷ Update average reward
16:    $\boldsymbol{w} \leftarrow \boldsymbol{w} + \alpha^w \delta \nabla \hat{v}(S, \boldsymbol{w})$  ▷ Adjust  $\hat{v}$ 
17:    $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha^\theta \delta \nabla \ln \pi(A \mid S, \boldsymbol{\theta})$  ▷ Adjust  $\pi$ 
18:    $S \leftarrow S'$  ▷ Update state
19:    $v \leftarrow \hat{v}(S', \boldsymbol{w})$  ▷ Evaluate new state
20:   Store  $R, V$  in lists  $R^{(T)}, V^{(T)}$ 
21:   if time-step  $\geq T$  then
22:      $\Psi = e(R^{(T)}, V^{(T)})$  ▷ Generate emotion
23:   end if
24: end loop

```

Algorithm 3 Training of an Order IV emotional actor-critic agent (continuing tasks)

```

1: Input: a differentiable policy parameterization  $\pi(a \mid s, \boldsymbol{\theta})$ 
2: Input: a differentiable state-value function parameterization  $\hat{v}(s, \boldsymbol{w})$ 
3: Input: a trained emotional encoder  $e(R^0, V^0, D^0)$ 
4: Input: a trained world model  $m(s, a)$ 
5: Input: a state-distance function  $diff(s, s')$ 
6: Initialize  $\bar{R} \in \mathbb{R}$  (e.g. to 0)
7: Initialize state-value weights  $\boldsymbol{w} \in \mathbb{R}^{d'}$  and policy parameter  $\boldsymbol{\theta} \in \mathbb{R}^{d''}$  (e.g., to  $\mathbf{0}$ )
8: Algorithm parameters:  $\alpha^w > 0, \alpha^\theta > 0, \alpha^{\bar{R}} > 0$ 
9: Initialize  $S \in S$ 
10: Initialize  $\Psi \in \mathbb{R}^d$  (e.g., to  $\psi_0 = \text{neutral emotion}$ )
11: Initialize  $R^{(T)}, V^{(T)}, D^{(T)}$  (as empty lists)
12: loop forever (for each time-step)
13:    $X = (S, \Psi)$  ▷ Extended state
14:    $A \sim \pi(\cdot \mid X, \boldsymbol{\theta})$  ▷ Choose an action
15:    $\hat{S}' \leftarrow m(S, A)$  ▷ Predict next state
16:   Take action  $A$ , observe  $S', R$  ▷ Observe new state and reward
17:    $\delta \leftarrow R - \bar{R} + \hat{v}(S', \boldsymbol{w}) - \hat{v}(S, \boldsymbol{w})$  ▷ Temporal difference error
18:    $\bar{R} \leftarrow \bar{R} + \alpha^{\bar{R}} \delta$  ▷ Update average reward
19:    $\boldsymbol{w} \leftarrow \boldsymbol{w} + \alpha^w \delta \nabla \hat{v}(S, \boldsymbol{w})$  ▷ Adjust  $\hat{v}$ 
20:    $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha^\theta \delta \nabla \ln \pi(A \mid S, \boldsymbol{\theta})$  ▷ Adjust  $\pi$ 
21:    $S \leftarrow S'$  ▷ Update state
22:    $V \leftarrow \hat{v}(S', \boldsymbol{w})$  ▷ Evaluate new state
23:    $D \leftarrow diff(\hat{S}', S')$  ▷ World-state prediction error
24:   Store  $R, V, D$  in lists,  $R^{(T)}, V^{(T)}, D^{(T)}$ 
25:   if time-step  $\geq T$  then
26:      $\Psi = e(R^{(T)}, V^{(T)}, D^{(T)})$  ▷ Generate emotion
27:   end if
28: end loop

```

Chapter 4

Results

4.1 Application of the framework on a practical case study

The methodology introduced in 3.4 was successfully applied in a Reinforcement Learning (RL) case study, and an overview of the results is presented here. The full details about each step, involving diverse numerical processing, are included in chapter *Methods and tools*.

The tests were executed on the classic RL environment *LunarLander-v2* [172]. In it, the objective is to safely land a spacecraft on a lunar surface, requiring control of the position and velocity while managing limited fuel resources (Fig. 4.1). The agent receives positive rewards for landing near the designated pad and maintaining a stable orientation, while penalties are incurred for excessive use of fuel and crashing. The task is considered solved when the average score over the last 100 episodes is greater or equal to 200.



Figure 4.1: The environment LunarLander-v2 used in the case study. A screenshot of the OpenAI Gym in which the spacecraft approaches the landing pad.

This environment was chosen for its simplicity (short episodes of 250-300 time-steps, and a maximum of 1,000 for failure), and the variety of life-or-death situations it presents for the simulated pilot, with potential for basic emotions spanning intense emotional ranges. Additionally, the lack of any emotional cues—no in-game character, face or body language is shown—guarantees an unbiased emotional attribution test.

By applying the methodology introduced in *Methodology*, the following two models were obtained (Fig. 4.2):

- Emotional encoder: A deep autoencoder (DAE) [163] was trained on 20,220 landing sequences experienced by a previously trained RL agent. The input values used were 20-step sequences of reward and state-value, in order to obtain an Order III emotional agent, generating their 5-dimensional step-wise representations.
- Emotional interpreter: A probabilistic Gaussian mixture model was trained on the 5-dimensional latent space learned by the emotional encoder, identifying eight distinct, uneven-sized clusters, whose centroids represented their respective prototypical multivariate sequences. For their interpretation, an Order III - Latest Observed Values Encoding (LOVE) 2:5x6 mapping was used (2 values, 5 reward x 6 state-value patterns), for more accurate mapping of clusters 3 and 7 than LOVE 2:5x5 in Fig. 3.4c.

A high-level diagram of the experimental pipeline is illustrated in Fig. 4.2, while additional visuals are detailed later.

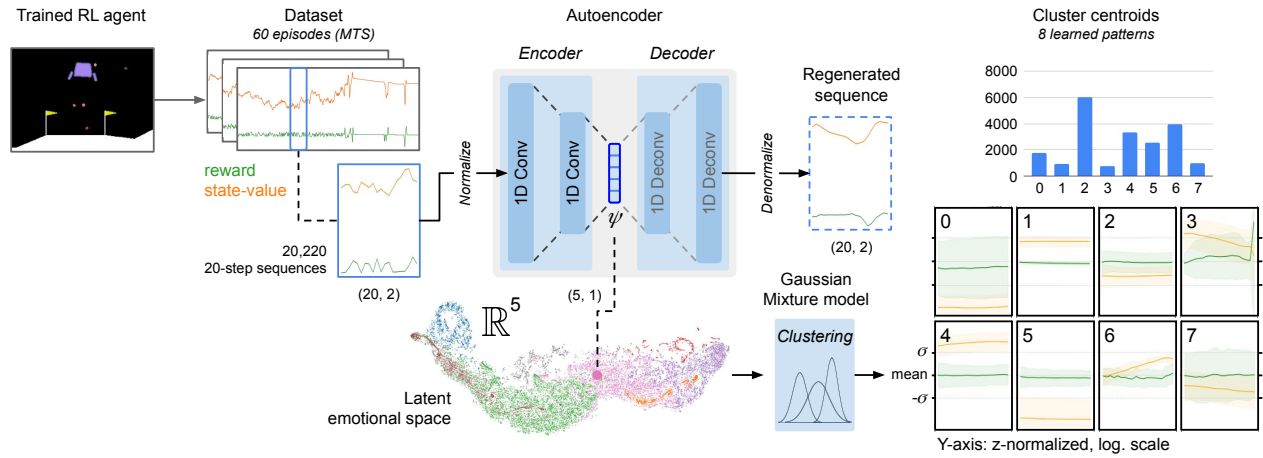


Figure 4.2: Learning emotions from experiences in a practical case study. The emotional framework was tested on the classic RL environment *LunarLander-v2*, on which an actor-critic PPO model (see 5.1.1) had been previously trained to solve the task. The trained agent was run on new scenarios to obtain a dataset with stepwise values for *reward* and *state-value*. A 1D-Convolutional autoencoder was then trained on sequences from the dataset, and their low-dimensional latent representation obtained (as shown in the 2D graph using t-distributed stochastic neighbor embedding, or t-SNE). Finally, a clustering model identified eight distinct, uneven-sized clusters, whose prototypical sequences are shown on the right as 20-step average sequences of *reward* and *state-value*, where the shaded areas indicate the standard deviation.

4.1.1 Learning emotions from experience

Pre-training of a conventional RL agent

For simplicity, the *offline* learning approach was chosen, in which the emotional model is trained with the experiences collected by an already competent non-emotional agent. In this case, an agent of the actor-critic PPO (Proximal Policy Optimization) family was previously trained on the environment to solve the task, obtaining its policy and state-value functions (see 5.1.1).

Dataset generation. Selection of input values and emotional window

The trained agent was run on unseen scenarios to obtain a representative dataset of 60 trajectories with stepwise values for a broad set of potential variables, from which *reward* and *state-value* were chosen so as to apply an Order III mapping, and an emotional window of 20 time-steps was set, obtaining 20,220 sequences. Figure 4.3 shows one example of the trajectories obtained.



Figure 4.3: Trajectory of a full episode run with the trained agent. The values registered as a multivariate time-series included a variety of step-wise variables, from which *Reward* and *(State-)value* were selected (excluding *Delta*, the temporal difference error; *Avg Rwd*, the average reward, and *Ema Rwd*, the exponential moving average of the reward).

Training of the emotional model from dataset sequences

A deep autoencoder architecture (DAE) was chosen for the task of representation learning, as explained in *Model architecture*. The 1D-Convolutional Autoencoder, suitable for time series, was trained and tested on the dataset in an unsupervised manner to reproduce 20-step x 2 values normalized multivariate time series (MTS) sequences by learning their latent representation in a low-dimensional latent space. The final architecture, with an encoding dimension of 5, was chosen for the best balance between reproduction error and compression ratio.

4.1.2 Elicitation of emotions

The trained emotional encoder was then used to encode the ongoing sequences of latest observed values from the full dataset, obtaining their 20,220 latent representations in the 5-dimensional latent emotional space. This was the emotional spectrum data used for interpretation. The integration of the emotional encoder within the extended RL actor-critic architecture introduced in *Methodology* (see 3.4.2 and Fig. 3.2) was not addressed in this experiment.

4.1.3 Interpretation of the learned emotions

Clustering of the emotional spectrum

To identify the distinct dynamic patterns present in the emotional spectrum, a probabilistic Gaussian mixture model was trained on the latent space learned. This method models data as a mixture of a number of Gaussian distributions, capturing its covariance structure in clusters of uneven spatial extents, which suited the nature of the problem.

The model identified eight distinct, uneven-sized clusters, shown in Fig. 4.4 (and Fig. A.5 in annex *3D visualization of the emotional spectrum*), whose centroids represent their respective prototypical multivariate sequences, as shown in Fig. 4.5.

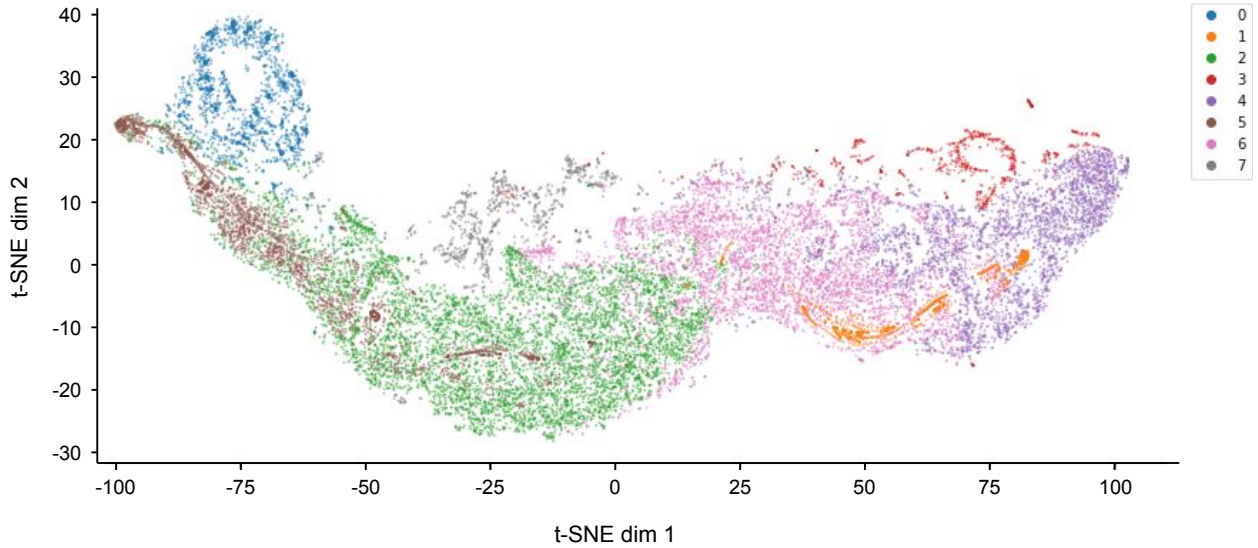


Figure 4.4: The eight classes identified by the emotional interpreter. Clustering on the learned emotional space revealed eight distinct classes, corresponding to eight hypothetical emotions. The figure shows a t-SNE representation in two dimensions of the five-dimensional emotional spectrum.

Selection and validation of the interpretability mapping

For emotional interpretation, an Order III - LOVE 2:5x6 was used (2 values, 5 reward times 6 state-value patterns, Fig. 4.6) for more accurate mapping of patterns 3 and 7 than the LOVE 2:5x5 interpretability mapping shown in Fig. 3.4. The chosen mapping was theoretically validated and refined as described in *Theoretical validation of LOVE profile terms*.

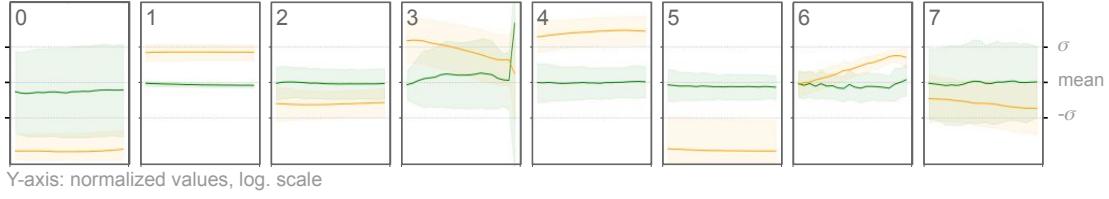


Figure 4.5: The eight emotional patterns learned. Average sequence corresponding to the centroids of each of the eight clusters identified (x-axis represents time; y-axis represents value in logarithmic scale).

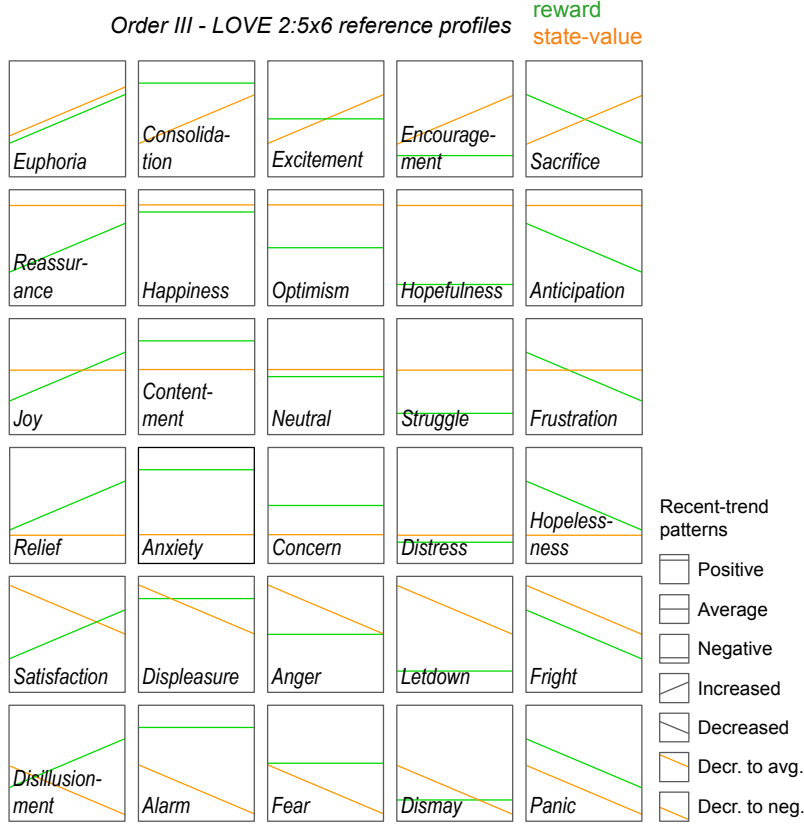


Figure 4.6: Order III - LOVE 2:5x6 Interpretability mapping. 30 reference profiles are characterized by two values (reward and state-value) for which 5 and 6 recent-trend patterns are defined respectively.

Attribution of emotion terms

Through statistical analysis of the average MTS sequence associated to each centroid, seven clearly differentiated basic emotions were identified, one of them in two degrees of intensity (*Optimism* and *high Optimism*) (Fig. 4.7a). The predominant class (cluster 2, with 29.8%), technically classified as *Neutral*, shows a distinct below-average state-value and will be treated as *Neutral / slight Concern* moving forward. As for the more clearly valenced emotions emerged from the agent’s experience, their distribution reflects its overall good competence at landing, with “positive” emotions (*Optimism*, *Satisfaction*, *high Optimism*, *Excitement*) totalling a 44.1%, while “negative” emotions (*Distress*, *Concern*, *Fear*) only add up 26.1%.

The eight patterns were associated with the best-matching emotional reference profile from the interpretability mapping by analytically comparing their features to the statistical characteristics of the registered trajectories (see *Attribution of emotion terms in Methods and tools*).

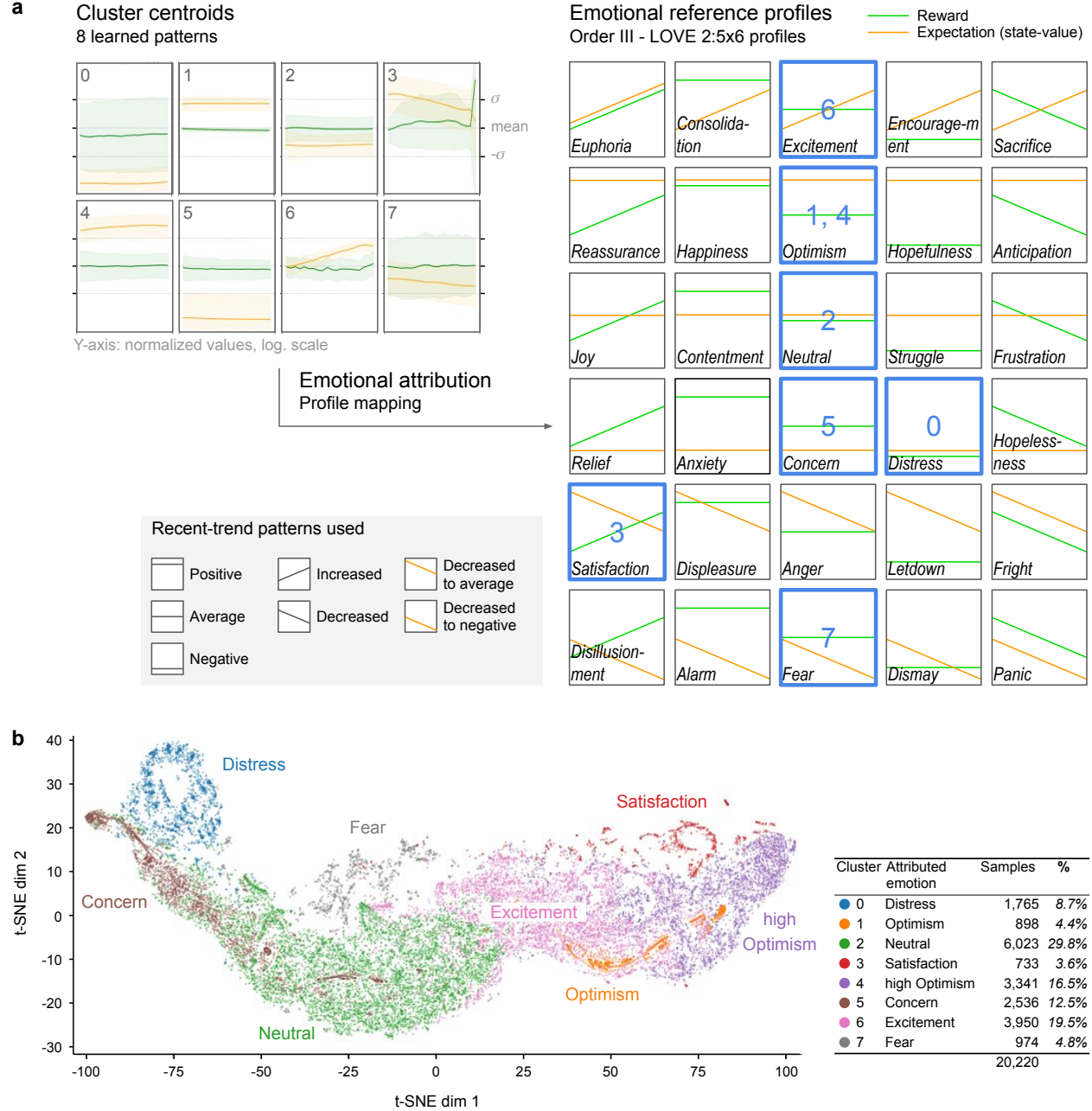


Figure 4.7: Interpretation of the learned emotions. **a, Interpreting the learned patterns.** Mapping the patterns against Order III - LOVE 2:5x6 reference profiles identified seven basic emotions, one of them (*Optimism*) in two degrees of intensity. **b, Interpretation of the learned emotional space.** The distribution of the learned emotions over the latent emotional space is shown in this 2D t-SNE graph. The overlapping between the identified classes, originating from the continuous nature of the 5-dimensional values, is accentuated in their 2D representation.

Relevantly, the fully-interpreted emotional spectrum in Fig. 4.7b suggests clearly natural transitions, critical for the utility of the model, such as:

- a) neutral \rightarrow excitement \rightarrow optimism \rightarrow high optimism \rightarrow satisfaction;
- b) neutral \rightarrow concern \rightarrow distress;
- c) neutral \rightarrow fear;
- d) etc.

Upon analyzing all the emotional transitions registered during the 60 episodes, we found that their succession flowed naturally, rather than behaving arbitrarily or randomly. A few transitions accounted for the majority of the cases, as shown in the transition matrix (Fig. 4.8a) and the transition graph (Fig. 4.8b). Episodes were initiated at “Start”, till enough steps had been registered within the emotional window, and typically transitioned to neutral or excitement, reflecting the overall high competence of the trained agent. Difficulties elicited transitions toward concern or fear, sometimes to distress, while proficient performance elicited excitement, optimism and eventually satisfaction at episode’s end.

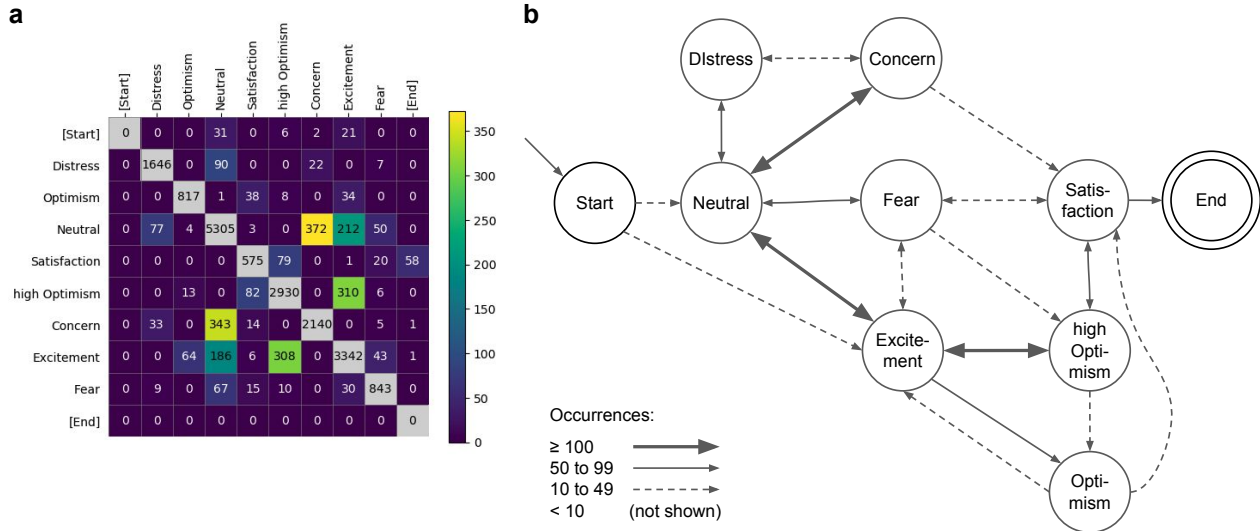


Figure 4.8: Emotional transition matrix and graph. **a**, Emotional transition matrix with the occurrence of each transtion over the 60 episodes (rows = initial emotion, columns = final emotion). **b**. Emotional transition graph illustrating the most frequent transitions.

Visualization

To further validate these results, the agent was tested on unseen scenarios, step-wise emotions were synthesized by the emotional encoder and their probability distribution over the eight clusters was predicted by the clustering model (Fig. 4.9).

Two of these scenarios are shown as Examples 1 and 2 below, vividly illustrating the stepwise elicitation and interpretation of learned emotions during two complete episodes: a successful and a failed landing (Figures 4.10 and 4.11).

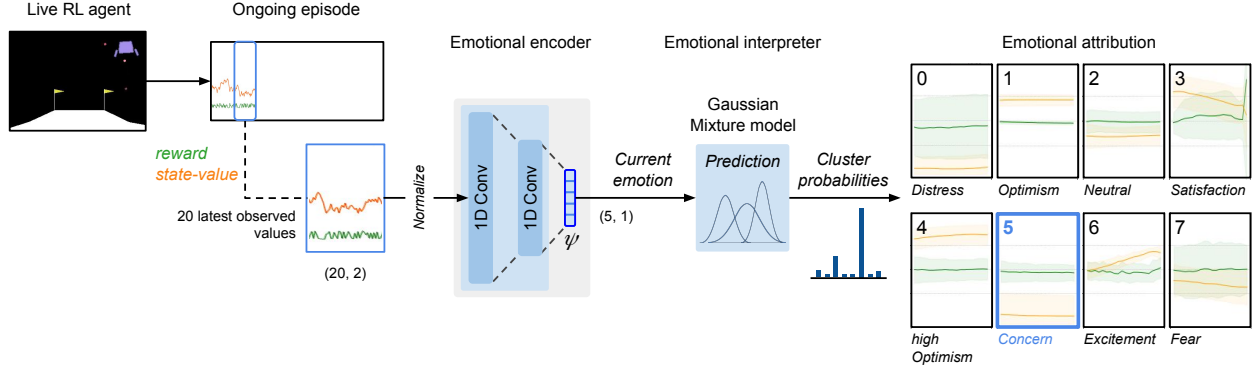


Figure 4.9: Interpretation of the instantaneous emotion. The emotional agent was tested on unseen scenarios and step-wise emotions synthesized by the learned emotional encoder from the latest observed rewards and state-values. The learned clustering model predicted the step-wise probability distribution of each encoded emotion over the eight clusters. In the image, where the agent is having difficulty aiming the spaceship toward the lunar base, a lower-than-average state-value (still unnormalized on the left) distinguishes “concern” as the predominant emotion (cluster 5).

Example 1: A successful landing (Fig. 4.10):

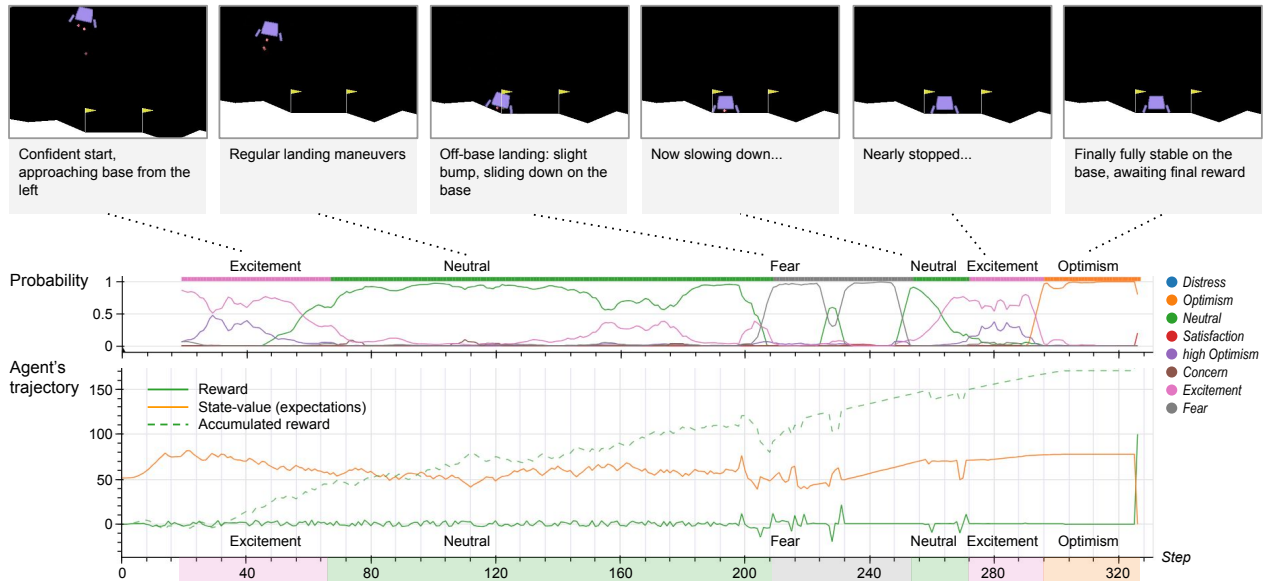


Figure 4.10: Interpreting live emotions during a successful landing.

Emotional attributions start at step 20 (the time window of the encoder) and during this episode, positive emotions convincingly match the events experienced by the agent, who only faces a minor incident around step 200 (eliciting *fear*). Emotional transitions happen naturally as reward accumulates (with some hiccups in the 200-230 interval) and expectations vary, ending in the sequence *neutral-excitement-optimism*. The stepwise probabilities frequently allow predominant and secondary emotions to blend into richer states (like *excitement* and *high optimism* at the start). Notice how the smoothing applied imposes some latency to the attributions, but reduces instability (for example, a spurious *neutral* glitch at step 230).

Example 2: A failed landing (Fig. 4.11):

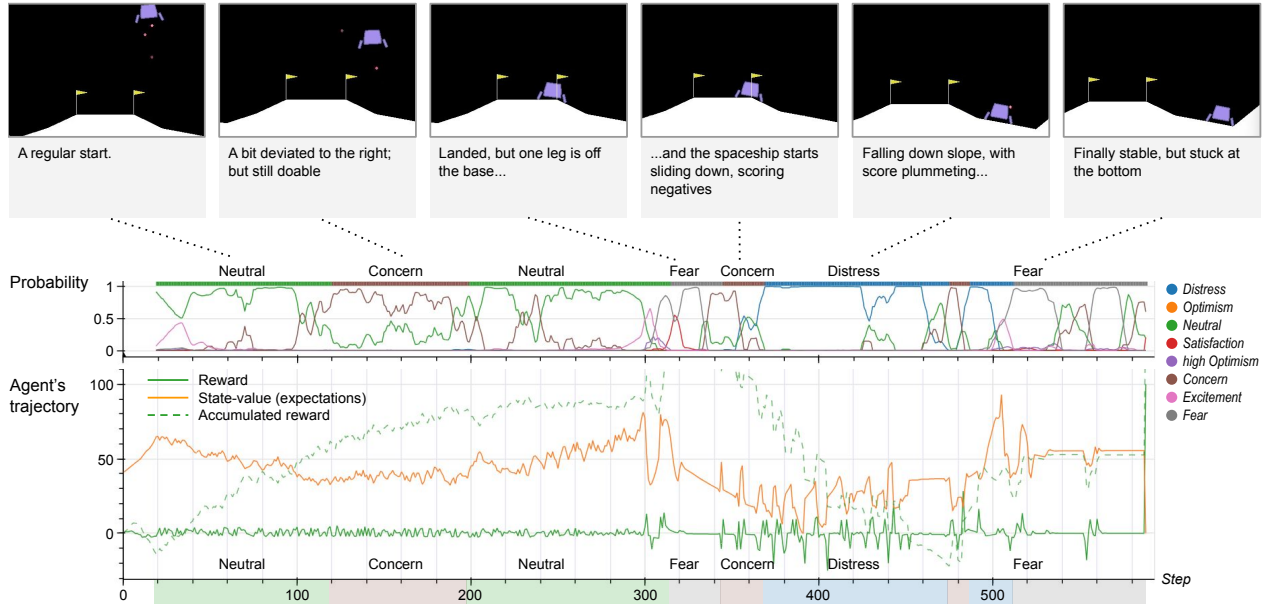


Figure 4.11: Interpreting live emotions during a failed landing.

This episode starts with lower expectations than *Example 1* (*excitement* fails to manifest itself), and *concern* is elicited at some point during the descent. Then the inaccurate landing in a precarious position produces *fear*, followed by *concern* and *distress* during the fall, as negative rewards accumulate down to a negative score. Once stabilized at the bottom, *fear* dominates the emotional state; notice though how the probabilistic attributions naturally elicit some sort of “emotional confusion” in a hectic situation, with overlapping emotions.

In summary, we found that the methodology spontaneously learned eight basic, recognizable emotions in an unsupervised manner. The synthetic emotional system naturally reproduced well-known natural emotional dynamics like:

- Elicitation and decay of step-wise emotions in synchrony with external changes observed and internal cognitive appraisals;
- Natural emotion transitions, with coexistence and progressive or sudden alternation driven by external and cognitive changes;
- Homeostasis, based on the agent’s subjective experience of average values registered;
- Subjectivity, with dependency on the individual’s appraisals (from the state-value function) and said homeostatic references.
- Environment dependency, with the emotional spectrum shaped and determined by the specific historical interactions between agent and environment;
- States of shock and confusion, with fast-overlapping negative emotions in highly unstable situations.

The results illustrate as well how, by associating instantaneous emotions with continuous values characterized by non-uniform distributions—which tend to give rise to clusters—the framework seamlessly integrates principles derived from both discrete and dimensional emotion theories, experimentally described in the literature [173].

In this experiment, however, with very short-lived episodes, the documented dynamics of “habituation” and “extinction” were not reproduced, despite their feasibility within the framework, as analyzed in *Conclusions*.

Finally, in most clustering models tried, and despite the necessary distortions produced by t-SNE, two axes consistently aligned with the two emotion dimensions most repeatedly identified by psychology: *pleasure* (or *valence*) and *arousal* (or *activation*, *dominance*) [30] [174][31][19]. As seen in Fig. 4.12, directly obtained from the use case discussed, the horizontal axis (pleasure or valence) arranges emotions from negative (concern, distress, neutral / slight concern, fear) to positive ones (excitement, optimism, satisfaction, high optimism). The vertical axis (arousal) sorts emotions from low (neutral / slight concern) to high (distress, fear, satisfaction), with the others in-between. The significance and consistence of this alignment with historically documented emotion dimensions remains unanalyzed.

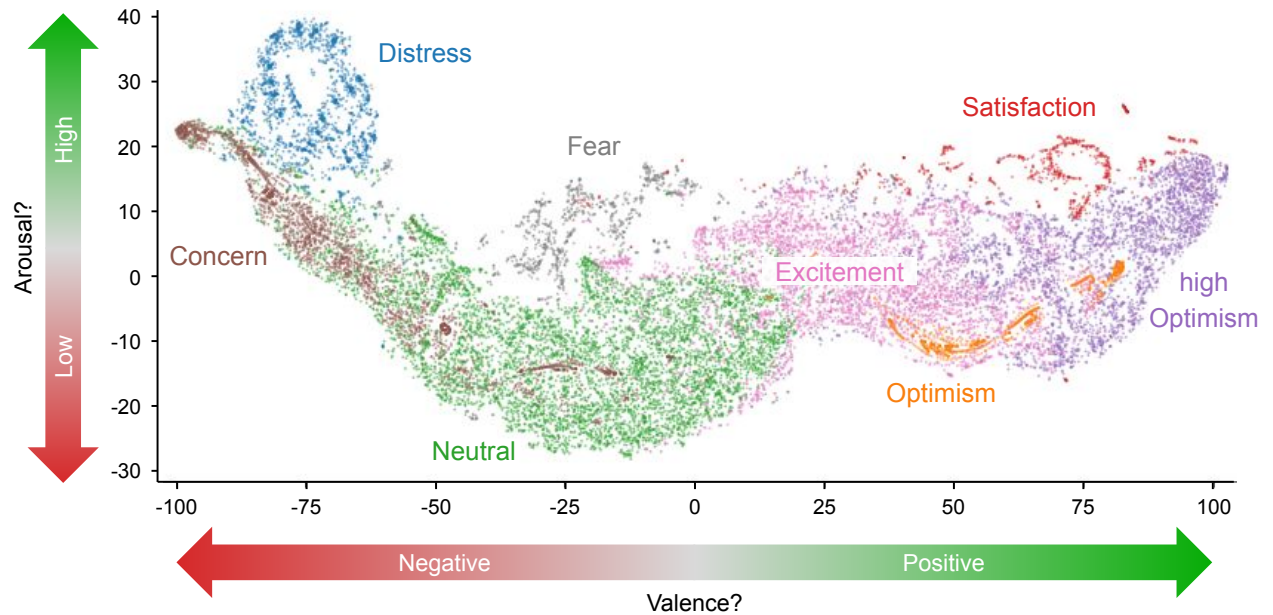


Figure 4.12: Spontaneous emergence of documented emotion dimensions. The emotional latent space obtained might naturally reproduce classic emotional dimensions.

4.2 Experimental validation of the learned emotions

4.2.1 Emotional attribution test with humans

The eight emotions learned from the agent’s experience by the emotional encoder had been convincingly mapped against predefined profiles and seemed to flow naturally in live action, but that did not prove that the synthetic emotions truly reflected natural emotions. To validate

their recognizability, an emotional attribution survey was executed, involving subjective observations from independent human participants.

Data obtained from human raters

The emotional attribution survey compared the subjective observations made by 96 independent participants during 48 different short sequences against their previously attributed emotion terms—concealed to them. The methodology used was Lang’s Self-Assessment Manikin (SAM) [175], an extensively applied evaluation technique that directly measures emotional responses on three dimensions: *pleasure*, *arousal* and *dominance* (PAD) [31], by rating each from 1 to 9 (see 5.3 in *Methods* for details). In our case, participants were asked to rate the pilot’s emotion at the end of 48 short unlabeled sequences.

Discussion of results

Upon analysis of the 2,304 PAD data points registered, we observed that their average values, both by sequence (Fig. 4.13a) and attributed emotion (Fig. 4.13b), reflected a differentiated emotional spectrum.

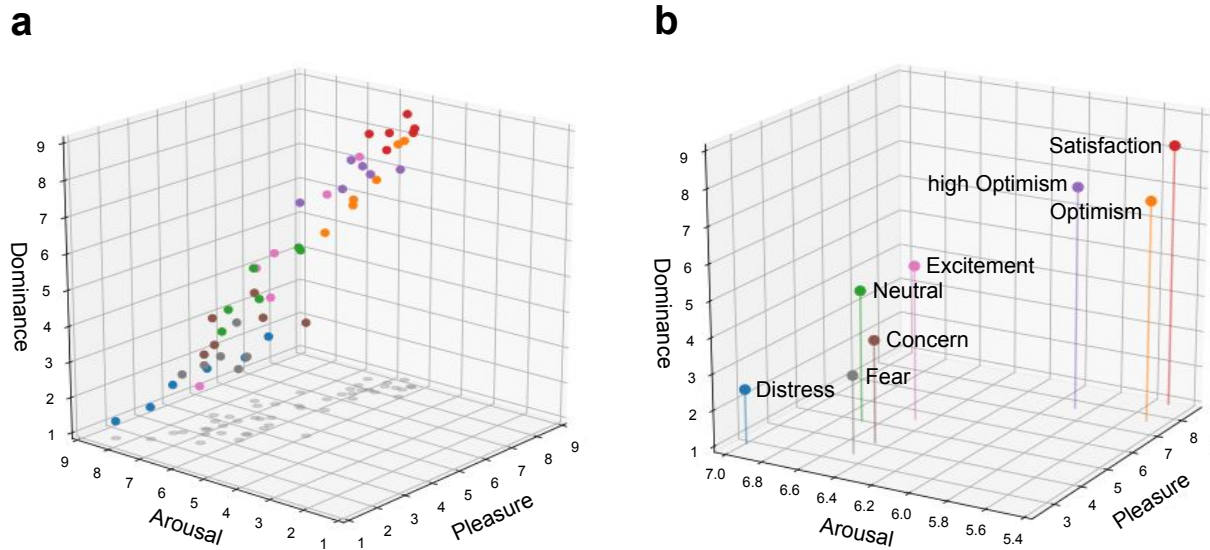


Figure 4.13: Results of the emotional attribution study. An emotional attribution survey ($n = 96$) was conducted to rate the pilot’s emotion at the end of 48 short unlabeled landing sequences. For the test, Lang’s Self-Assessment Manikin (SAM) [175] was used, capturing the dimensions of *pleasure*, *arousal* and *dominance* (PAD) with Likert scales from 1 (lowest) to 9 (highest). **a, PAD attributed to each sequence.** Unaware of the emotions previously attributed to the sequences, the participants consistently rated samples of the same class with similar PAD values, as illustrated by the color mapping. **b, PAD attributed to each emotion.** The average PAD values of the videos associated with each emotion show a meaningful and coherent progression along the axes.

Furthermore, the PAD values corresponded well with their associated emotions. For instance, sorting them by *pleasure* reveals a nearly perfect progression from negative to positive emotions (Fig. 4.14): distress, fear, concern, neutral / slight concern, excitement, high

optimism, optimism, satisfaction (with the exception of the two optimism states, similar in pleasure, but distinguished by the higher arousal attributed to high optimism).

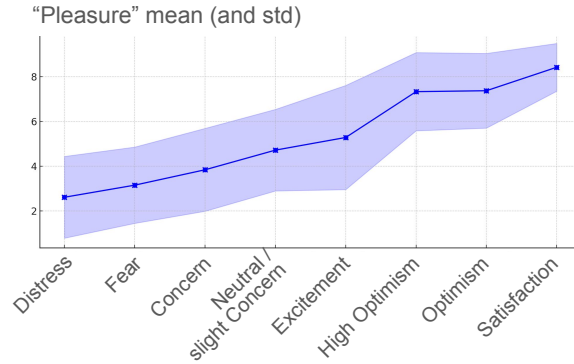


Figure 4.14: Average Pleasure value attributed to sequences of each emotion. A nearly perfect natural sorting by increasing Pleasure emerges spontaneously. The shaded area reflects the standard deviation.

Notably, low *arousal* ratings were scarce, in consonance with the dynamic, high-stakes nature of the sequences, with most ratings falling between 5 and 7.

The values for *dominance*, spanning from 2.5 to 7.2, were highly correlated with pleasure in this environment (see Table 4.2 and Table 4.3), and the ordering of emotions from “low” to “high” is almost the same: distress, fear, concern, neutral / slight concern, excitement, optimism, high optimism, satisfaction. The less clearly identified class was excitement, showing the highest dispersion in pleasure and dominance.

Table 4.1 shows the PAD rates obtained for each targeted synthetic emotion, aggregated over all raters from their corresponding six videos, clearly reflecting the differentiated emotional characterization obtained. (The PAD rates attributed to each of the 48 videos can be seen in Table A.1 in annex *Experimental validation of learned emotions with humans: Extended data*.)

Table 4.1: PAD rates of each emotion in range [1, 9] aggregated over all raters from their six corresponding videos.

Emotion	Pleasure		Arousal		Dominance	
	(mean)	(stdev)	(mean)	(stdev)	(mean)	(stdev)
● Distress	2.615	1.827	6.955	2.228	2.500	1.627
● Optimism	7.378	1.669	5.469	2.186	7.101	1.859
● Neutral / Slight Concern	4.722	1.823	6.649	1.838	4.642	1.865
● Satisfaction	8.427	1.067	5.486	2.572	8.219	1.360
● High Optimism	7.337	1.749	5.851	2.245	7.184	1.969
● Concern	3.844	1.852	6.437	2.108	3.833	1.991
● Excitement	5.288	2.329	6.437	2.059	5.292	2.405
● Fear	3.156	1.705	6.448	2.091	3.149	1.722

The correlation rates among the three dimensions across all videos (Table 4.2) indicate a strong positive correlation between pleasure and dominance, highlighting a significant relationship between these two dimensions. In contrast, pleasure and arousal, as well as arousal and dominance, show a moderate to strong negative correlation, reflecting the complex interplay

between these emotional dimensions. These correlations manifested themselves even more visibly on the PAD values, once aggregated over the eight emotions (Table 4.3).

Table 4.2: Correlation among Pleasure / Arousal / Dominance across all the videos (Pearson Two-sided).

X	Y	r	CI95%	p-unc	BF10	Power
Pleasure	Arousal	-0.69329	[-0.82, -0.51]	4.72E-08	3.42E+05	0.999933
Pleasure	Dominance	0.99436	[0.99, 1.00]	1.76E-46	1.49E+42	1
Arousal	Dominance	-0.688291	[-0.81, -0.50]	6.43E-08	2.56E+05	0.999913

Table 4.3: Correlation among Pleasure / Arousal / Dominance across the eight emotions (Pearson Two-sided).

X	Y	r	CI95%	p-unc	BF10	Power
Pleasure	Arousal	-0.922272	[-0.99, -0.62]	1.11E-03	3.37E+01	0.961207
Pleasure	Dominance	0.999416	[1.00, 1.00]	4.99E-10	6.82E+05	1
Arousal	Dominance	-0.917867	[-0.99, -0.60]	1.30E-03	3.01E+01	0.955431

The test obtained a high reliability, with a high degree of agreement on PAD values for each video sequence across raters, according to the ICC2k statistical tests run (Intraclass Correlation Coefficient, two-way random effects model, absolute agreement) (see Table 4.4). The pleasure and dominance dimensions obtained “excellent” correlation rates according to the orientative criteria by (Koo, 2016) [176] (greater than 0.90) and (Cicchetti, 1994) [177] (greater than 0.75), while arousal achieved “good” per one guideline (between 0.75 and 0.90) and “excellent” per another (greater than 0.75) (see Table 4.4).

Table 4.4: ICC2k Intraclass Correlation Coefficients obtained from the study.

Dimension	Test	ICC	F	df1	df2	p-value	CI95%
Pleasure	A	0.991303	143.006427	23	1173	< 0.001	[0.99 1.00]
Pleasure	B	0.987127	95.329519	23	989	< 0.001	[0.98 0.99]
Arousal	A	0.869813	10.420163	23	1173	< 0.001	[0.79 0.93]
Arousal	B	0.809354	6.115296	23	989	< 0.001	[0.69 0.90]
Dominance	A	0.987786	108.758808	23	1173	< 0.001	[0.98 0.99]
Dominance	B	0.984978	83.461934	23	989	< 0.001	[0.97 0.99]

Equally remarkable, the underlying hypothetical emotions, despite being unknown to the raters, showed a high degree of statistical distinguishability according to the Hotelling’s T-squared pairwise tests run: optimism and high optimism were the least distinguishable emotions (with $p = 0.116$), followed by excitement and neutral ($p = 0.002$), while all other pairwise comparisons showed p -values well below 0.001 (see Table 4.5).

4.2.2 Mapping versus documented experimental accounts

Finally, to further validate the significance of the learned emotions, the PAD rates obtained for each emotion were mapped versus select pivotal experimental accounts from human subjects, numerically documented in psychology literature, obtaining significant agreement with some

Table 4.5: p-value of the Hotelling’s T-squared statistical test for all emotion pairs.

	D	O	N	S	H-O	C	E	F
D	-	-	-	-	-	-	-	-
O	3.66E-143	-	-	-	-	-	-	-
N	8.03E-44	4.32E-65	-	-	-	-	-	-
S	9.06E-211	1.04E-17	1.91E-123	-	-	-	-	-
H-O	3.34E-135	1.16E-01	6.79E-58	8.63E-17	-	-	-	-
C	4.71E-18	1.83E-92	4.69E-08	2.45E-155	8.91E-86	-	-	-
E	1.14E-49	1.36E-33	2.13E-03	1.62E-71	1.23E-28	3.48E-15	-	-
F	4.46E-06	2.37E-127	1.30E-25	1.57E-197	1.29E-118	1.39E-05	4.34E-33	-

Abbrev.: (D)istress, (O)ptimism, (N)eutral / slight concern, (S)atisfaction, (H)igh (O)ptimism, (C)oncern, (E)xcitement, (F)ear.

of the most broadly referenced PAD lists, selected for their mathematical qualities and impact in their field [31][178][179][180][181].

To compare the PAD multivariate distribution obtained for each emotion with the referential PAD values, we applied statistical tests (Hotelling’s T-squared) and plain euclidean distance among means, focusing on the emotion terms of relevance for our context (short life-or-death landing maneuvers), excluding complex social, moral, self-conscious affects (such as kind, guilty, repentant), or the frequent non-emotion terms (like butter, cemetery, chair).

We firstly identified the three top matches in each account for each emotion, shown in Table 4.6 (the detailed methodology is described in *Methods and tools*, section *Mapping versus documented experimental accounts*).

Based on this mapping, we then produced a *semantic collage* for each learned emotion with the five top matches across authors, obtaining the final mapping shown in Table 4.7.

Remarkably, despite the disparity of terms, PAD values and applied mapping variants, significant agreement emerged between the originally attributed terms (from the LOVE 2:5x6 mapping) and their respective top matches across authors.

In summary, the videos previously associated with each emotion, invisible to the external raters, were described by them with consistent and differentiated values for their pleasure, arousal and dominance dimensions, validating their distinguishability and recognizability. The PAD values obtained could then be successfully associated with remarkably similar emotions described by their own PAD values in psychology literature.

Table 4.6: Results from mapping PAD values from the survey to five documented experimental accounts.

Learned Emotion	Mapping to Documented Emotion-PAD Pairs (Top 3 Matches)				
	Russell-Mehrabian (1977)	Bradley-Lang (1999)	Redondo (2007)	Landowska (2018)	Scott (2019)
Distress	Helpless Fearful Insecure	Scared Panic Embarrassed	Nervous Lost Insecure	Fearful Pain Terrified	Fearful Frightened Panic
Optimism	Capable Concentrating Proud	Optimism Masterful Inspired	Capable Easy Confident	Masterful Strong Powerful	Skilled Mighty Pride (feeling)
Neutral / slight Concern	Anxious Tense Startled	Startled Overwhelmed Anxious	Troubled Moody Ecstasy	Startled Anxious Suspicious	Intense Impulse Urgent
Satisfaction	Proud Capable Self-satisfied	Confident Triumph Victory	Safe Capable Satisfied	Proud Joyful Masterful	Achievement Courage Triumphant
high Optimism	Capable Concentrating Strong	Brave Pride Bold	Confident Capable Interest	Strong Powerful Inspired	Pride (feeling) Might (strength) Drive (motivation)
Concern	Confused Tense Pain	Nervous Overwhelmed Suspicious	Moody Thrill Troubled	Suspicious Confused Startled	Startle Risky Shock
Excitement	Aroused Concentrating Anxious	Startled Alert Anxious	Power Pride Activate	Anxious Aggressive Curious	Impulse Intense Alert
Fear	Insecure Confused Pain	Nervous Panic Scared	Thrill Fearful Suspicious	Fearful Despairing Frustrated	Fright Scary Startle

Table 4.7: Top pleasure-arousal-dominance (PAD) matches across authors for each learned emotion.

Learned Emotion	Top PAD Matches Across Authors
Distress	Helpless, Scared, Nervous, Fearful(x2)
Optimism	Capable(x2), Optimism, Masterful, Skilled
Neutral / Slight Concern	Anxious, Startled(x2), Troubled, Intense
Satisfaction	Proud(x2), Confident, Safe, Achievement
High Optimism	Capable, Brave, Confident, Strong, Pride (feeling)
Concern	Confused, Nervous, Moody, Suspicious, Startle
Excitement	Aroused, Startled, Power, Anxious, Impulse
Fear	Insecure, Nervous, Thrill, Fearful, Fright

Chapter 5

Methods and tools

In this chapter we detail the exact step-by-step procedures followed to obtain the results described in *Results* in the classic Reinforcement Learning (RL) environment chosen, which can be applied to other RL setups with minimal modification. The purpose of this chapter is to enable the replication of the obtained results and their reuse on different contexts. While this information is crucial for replicating our findings, it may not be essential for the comprehension of the final *Conclusions*.

5.1 Application of the framework on a practical case study

We specify here technical and procedural details that clarify and justify low-level aspects of the processes followed to learn, elicit and interpret emotions in the case study presented in *Results*.

5.1.1 Learning emotions from experience

Pre-training of a conventional RL agent

For simplicity, the *offline learning* approach was chosen, in which the emotional model is trained with the experiences collected by an already competent non-emotional agent. The open-source library *OpenAI's Spinning Up* [182], compatible with *OpenAI's Gym* [172], was chosen because its modular and well-documented implementation of RL algorithms facilitated the extensions required for the experiments.

The chosen method, actor-critic PPO (Proximal Policy Optimization) [183], from the policy-gradient family, is broadly used for its stability during training, avoiding too large policy updates. Its training learns both a policy π (the actor) and a value function v (the critic).

The non-emotional agent was trained to solve the task (average episode reward ≥ 200 over 100 consecutive episodes), with these features:

- Agent: Actor-critic PPO model, artificial neural network architecture: (64, 64), activation function: rectified linear unit (*ReLU*), seed: 10;
- Hyperparameters: *gamma*: 0.99, *lambda*: 0.97, *policy learning rate*: 0.0003, *state-value function learning rate*: 0.001, *target Kullback-Leibler (KL)*: 0.01.

Figure 5.1 reflects the performance of the agent during training, with the vertical axis showing the total accumulated reward at episode end, and the horizontal axis showing the total amount of stepwise interactions experienced by the agent.

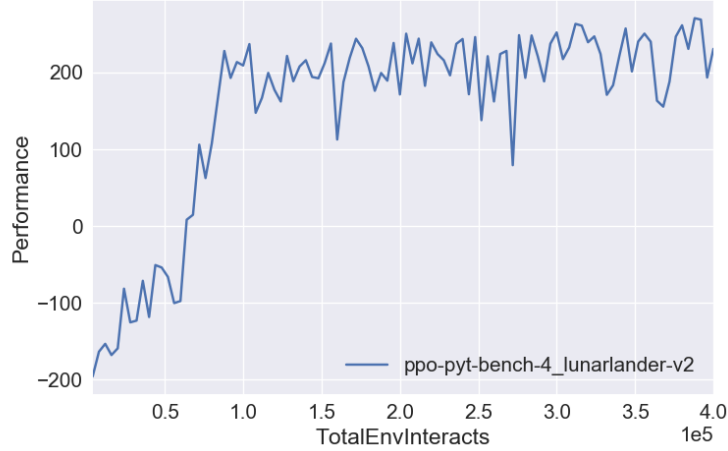


Figure 5.1: Training of the conventional RL agent chosen.

The architectures explored involved combinations of 1-2-3 hidden layers, 32-64-128 neurons per layer, *ReLU* / *tanh* activation functions, and varied seeds (see *Alternative conventional RL agents trained*).

Dataset generation. Selection of input values and emotional window

The trained agent was run on unseen scenarios to obtain a representative dataset of 60 episodes as multivariate time series (MTS) with stepwise values for a broad set of potential variables: reward, state-value, temporal difference, average reward, exponential moving average reward, and cumulative reward (see Fig. 5.2).

Upon review of the recorded episode dynamics and MTS, an Order III target mapping was chosen, for which only reward and state-value were required (see discussion of alternatives in the *Theoretical framework*, section *Emotional orders*). (We anticipate that this mapping may perform well in a large variety of setups for its potential to capture a broad range of short-term, fundamental emotions with moderate, addressable complexity.) A tentative value for the emotional window was set at 20, expected to suffice to capture instantaneous emotions, and later corroborated by results.

The resulting dataset contained 20,220 20-step long sequences from the recorded MTS (training / test split = 16,281 / 3,939). The two-variables (reward, state-value) were z-score normalized for training based on training-set statistics, thus establishing their average values as homeostatic references.

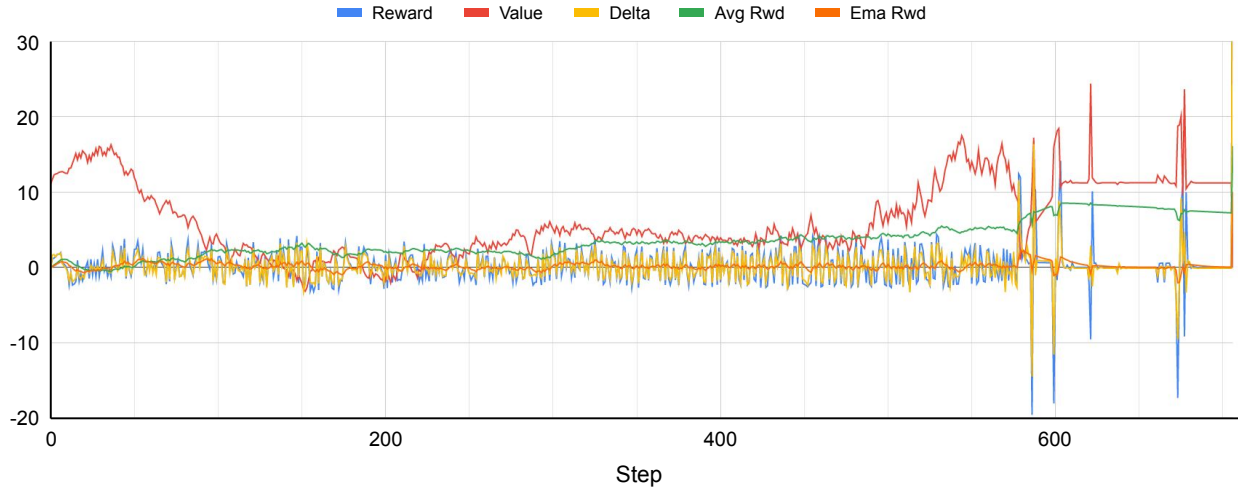


Figure 5.2: All the values registered during one trajectory. The registered multivariate time-series included a variety of step-wise variables, from which *Reward* and *(State-)value* were selected.

Training of the emotional model from dataset sequences

A 1D-Convolutional Autoencoder—a type of deep autoencoder (DAE) [163] whose architecture is suitable for time series—was chosen for the task of representation learning (as explained in *Model architecture*). The model was trained and tested on the dataset in an unsupervised manner to reproduce 20-step x 2 values normalized MTS sequences by learning their latent representation in a low-dimensional latent space [164].

We used the *Keras/TensorFlow* library [184], defining an encoder and a decoder with these features: *encoding_dim* = 5, *l1_filters* = 32, *l1_kernel_size* = 5, *l1_strides* = 2, *l2_filters* = 16, *l2_kernel_size* = 5, *l2_strides* = 2, *padding* = “same”, *activation* = “relu”. The separate encoder, used for emotion elicitation, consisted of 3,333 learned parameters. The training took 29 epochs with *batch_size*=10 and *validation_split*=0.1.

Along with different architectures and parameters, several encoding dimensions were tried, and 5 was found to produce the best balance between reproduction root mean squared error (RMSE) and compression ratio (reward: RMSE = 2.97119; state-value: RMSE = 4.32781; compression ratio = 5:40) (Figure 5.3).

For example, dim=10 yielded lower RMSE (2.48047 and 4.21228 respectively), but a poorer compression ratio of 10:40. As noted, the model was not intended for input regeneration or denoising, but to capture high-level trends and magnitudes of the observed sequences. (For context, the reward time series showed a range of [-100, 100], with a mean of 0.67 (standard deviation: 5.60), while the state-value time series spanned from [-12.37, 109.51] with a mean of 62.41 (standard deviation: 18.99).)

Figure 5.4 compares several examples of original MTS sequences with their reconstructions at the output of the autoencoder. As intended, the high-level trends and magnitudes were mostly preserved, while the details were simplified.

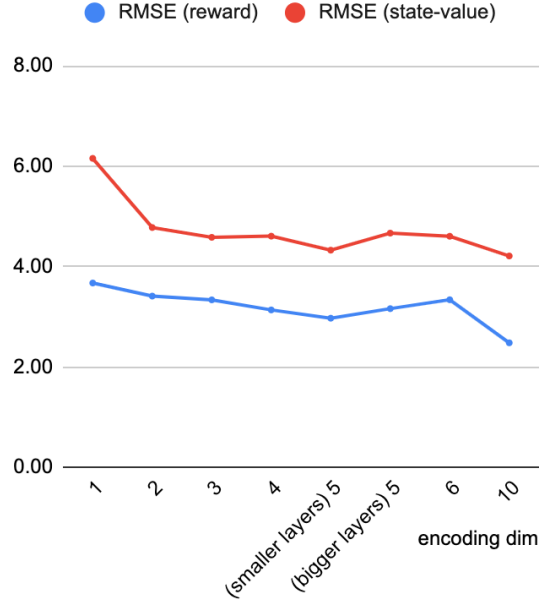


Figure 5.3: Comparison of different encoding dimensions. Root Mean Square Error obtained on the test-set by different autoencoder dimensions for the two values of the time-series. Model “(smaller layers)” with dimension 5 was chosen over best-performing dim = 10 for its better compression ratio. Model “(bigger layers)” —with 32 filters in hidden layer 2 over the 16 filters of “(smaller layers)” —showed a worse performance.

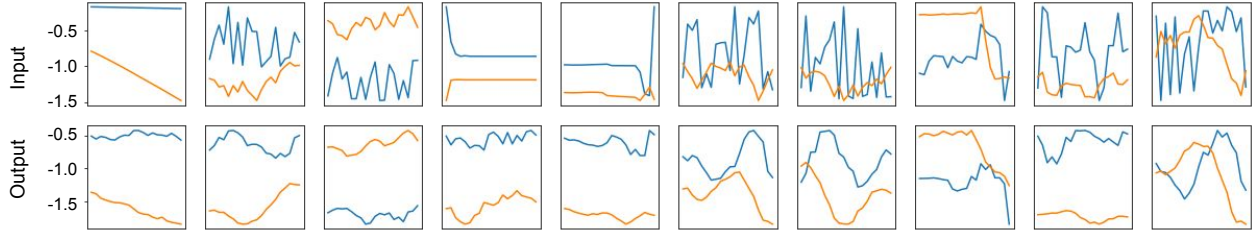


Figure 5.4: Test sequences reproduced by the autoencoder. Ten original MTS sequences (top) alongside their reproductions by the trained deep autoencoder (bottom) preserve most of the major features. Each sequence depicts the two variables in the MTS, with the x-axis representing time.

5.1.2 Elicitation of emotions

Obtaining the 5-dimensional representations of each step-wise emotion was simply a matter of using a sliding window to loop through all registered trajectories and feeding the segments into the trained autoencoder.

5.1.3 Interpretation of the learned emotions

Clustering of the emotional spectrum

To identify the distinct dynamic patterns in the emotional spectrum captured, a probabilistic Gaussian mixture model (GMM) was trained on the latent space learned. For the training of the GMM, the *Scikit-learn* library [185] was used for a number of clusters between 1 and 16, and up to ten different random seeds for each. We found the most promising clustering

distributions to consist of 7 or 8 clusters, with minimal BIC scores (Bayes Information Criterion) and sufficient differentiation, although somewhat dependent on the initial seed.

The final choice was arbitrary, following practical experimentation on real sequences, and settled on 8 components with covariance type = “full” (assigning to each component its own general covariance matrix). For this purpose, the custom tool “RL Emotion Lab” was built, reproducing full episode trajectories on the environment along with the stepwise probabilistic classifications of the interpreter of the synthetic emotions (Fig. 5.5). The *Bokeh* library [186] was used, and episodes and analysis reproduced on a regular web browser.

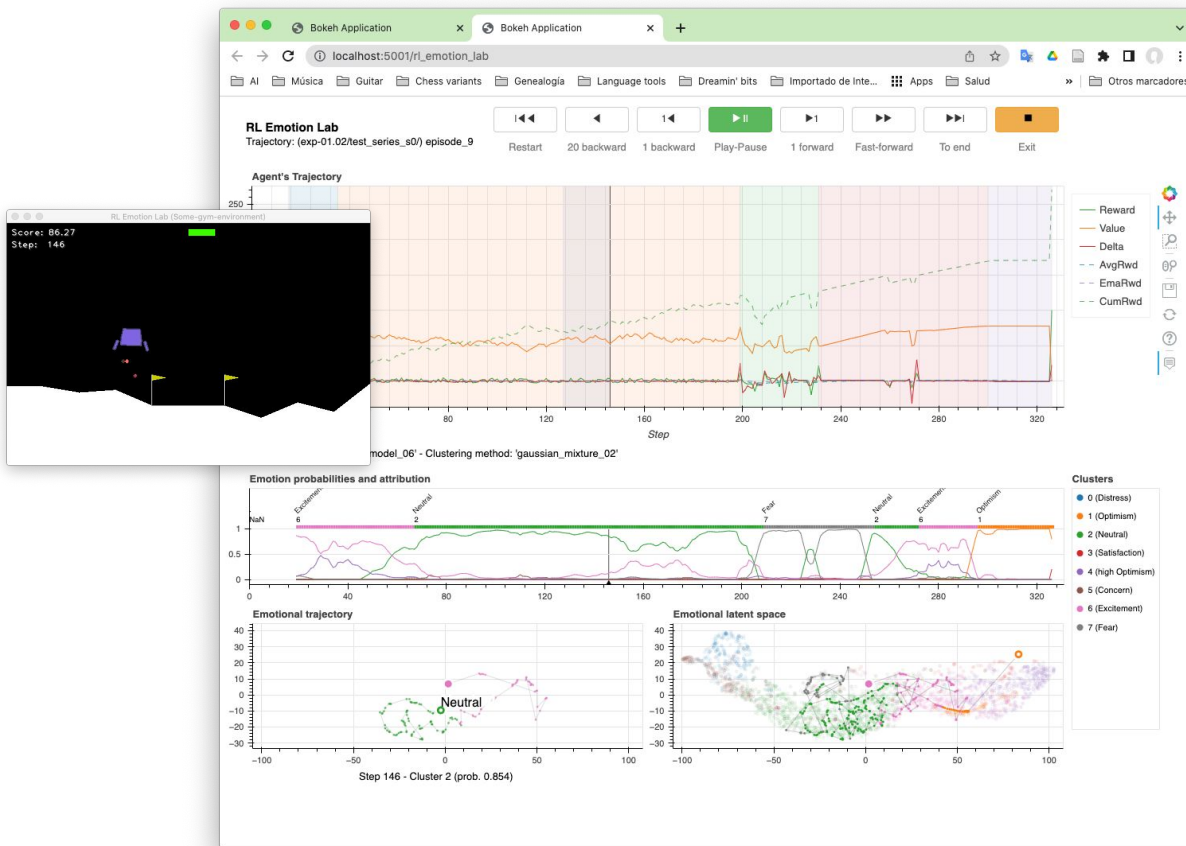


Figure 5.5: The custom-made tool “RL Emotion Lab”. The tool allowed the step-by-step reproduction and analysis of fully documented trajectories. **Floating window:** State of the agent and the environment at time-step 146. **Top:** The agent’s trajectory as an MTS, including Reward, (State-)value, Delta (or temporal difference), step-wise Average Reward, step-wise Exponentially Moving Average Reward and Accumulated Reward. Current step (146) is tracked by the vertical line, with the shaded area signaling the 20-step emotional window. **Middle:** Stepwise emotional probabilities of the eight classes from the emotional encoder, as well as their attributed emotion term. **Bottom right:** The full emotional trajectory within the emotional space, with initial emotion (solid circle) and final emotion (empty circle). **Bottom left:** The on-going emotional trajectory upto current time-step, classified as Neutral.

This approach exhibited satisfactory performance, thereby obviating the need for further automation (some alternative clustering results are shown and discussed in *Alternative*

clustering results). The eight resulting classes, along with the average multivariate sequence representing their corresponding cluster centroids, are shown in Fig. 5.6.

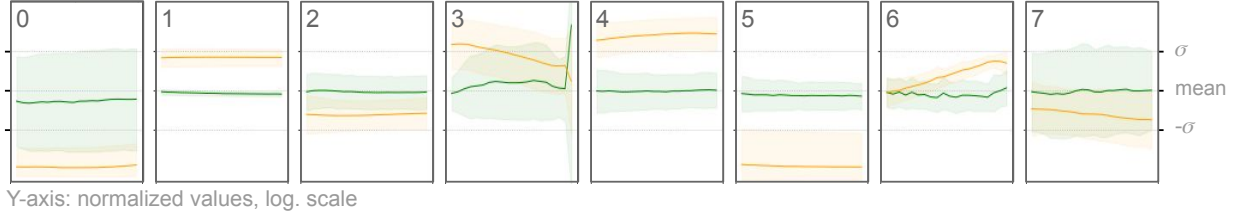


Figure 5.6: The eight emotional patterns learned. Average sequence corresponding to the centroids of each of the eight clusters identified. The line shows the mean value at each step, and the shaded area its standard deviation (x-axis represents time; y-axis is in logarithmic scale).

Selection and validation of the interpretability mapping

To map the eight resulting clusters with familiar emotion terms, the Latest Observed Values Encoding (LOVE) 2:5x5 mapping was initially tried—corresponding to the Order III emotional spectrum learned, with two values: reward and state-value (Fig. 3.4c). After a preliminary attribution (as described below), the mapping was extended to LOVE 2:5x6 for a more accurate denomination of emotions 3 and 7 (Fig. 4.7). This allowed the distinction between anger and fear based on the individual’s appraisal on its certainty and control on future outcomes (with anger associated with a decrease to average, and fear with a decrease to negative (Q. Yang et al., 2018) [158]).

The final terms for the eight emotions learned in this use case, as well as for the full set of thirty, were theoretically validated as described in *Theoretical validation of LOVE profile terms*, verified in live simulations and, finally, experimentally contrasted with external references (see *Experimental validation of learned emotions with humans*).

Attribution of emotion terms

To associate the eight patterns with the best-matching emotional reference profile from the chosen interpretability mapping, the statistical features of each pattern were analytically compared to the global statistical features of the registered trajectories.

Firstly, the mean value and slope of each variable (and their standard deviations) were calculated across all 20-step intervals of all the original MTS trajectories, as shown in Table 5.1. Note that mean values reflect the normalized distributions of the two variables. The mean slope (and its standard deviation) was computed across all 20-step sequences from their first-degree polynomial approximations.

Once these referential values were obtained, the mean value and slope were calculated for each of the eight multivariate sequences representing the learned classes. For their classification into discrete LOVE-compatible patterns, the following recent-trend classification criteria were established, based on the slopes of the sequences and their respective value ranges:

Table 5.1: Global statistics of all the original MTS trajectories by 20-step sequences.

	Reward	State-value
Mean value	0.0 (+/- 1.0)	0.0 (+/- 1.0)
Mean slope	0.0004716 (+/- 0.0259)	0.001478 (+/- 0.03235)

1. *Trend change of slopes* is the primary classification criterion, following the emotional relevance guidelines established by the theoretical framework, taking precedence over the secondary criterion below. A range around s_{mean} of length $0.5s_{std}$ was defined for “Flat” (where s_{mean} is the mean slope, and s_{std} is the standard deviation), and values above and below were respectively classified as “Increased” and “Decreased” (Table 5.2), separately for *reward* and *state-value*.

For the LOVE x6 patterns, a distinction was required for decreased trends, and patterns were classified as “Decreased-to-average” or “Decreased-to-negative” depending on the sign of the last value of the linear regression in the interval.

Table 5.2: Recent-trend classification criterion of a variable’s slope.

Slope value ranges	Classification
$\geq s_{mean} + 0.25 s_{std}$	Increased
$(s_{mean} - 0.25 s_{std}, s_{mean} + 0.25 s_{std})$	Flat
$\leq s_{mean} - 0.25 s_{std}$	Decreased

Note: s_{mean} and s_{std} are the referential mean and standard deviation of the slopes across all sequences for the normalized variable.

2. *Value range classification of the mean values* followed a similar approach, as summarized in Table 5.3, in which two extra ranges of length 0.5σ were defined for finer categorization of some patterns. This secondary criterion only applies to “Flat” slope changes.

Table 5.3: Recent-trend classification criterion of a variable’s mean value.

Mean value ranges	Classification
$\geq 0.75\sigma$	high Positive
$[0.25\sigma, 0.75\sigma)$	Positive
$(-0.25\sigma, 0.25\sigma)$	Average
$(-0.75\sigma, -0.25\sigma]$	Negative
$\leq -0.75\sigma$	high Negative

Note: σ is the standard deviation of each normalized variable.

The application of these two classification criteria over the eight clusters for the two variables is shown in Table 5.4 and Table 5.5, while the final attribution of emotion terms is presented in Table 5.6.

The soundness of this analytical attribution is supported by the following brief qualitative discussion of each cluster’s features:

Table 5.4: Local statistics and classification of the eight individual average sequences for *Reward*.

Cluster	Mean	Slope	Trend change	Value range	LOVE 2:5x5 pattern
0	-0.2679	0.0048	(flat)	Negative	Negative
1	-0.0726	-0.0028	(flat)	Average	Average
2	-0.0368	-0.0025	(flat)	Average	Average
3	0.2117	0.0275	Increased	Average	Increased
4	-0.0155	0.0018	(flat)	Average	Average
5	-0.1224	-0.0021	(flat)	Average	Average
6	-0.0992	0.0010	(flat)	Average	Average
7	-0.0270	0.0046	(flat)	Average	Average

Table 5.5: Local statistics and classification of the eight individual average sequences for *State-value*.

Cluster	Mean	Slope	Slope change	Value range	LOVE 2:5x6 Pattern
0	-1.9240	0.0015	(flat)	high Negative	high Negative
1	0.8304	0.0002	(flat)	high Positive	high Positive
2	-0.6099	0.0025	(flat)	Negative	Negative
3	0.8762	-0.0399	Decreased	high Positive	Decreased to average
4	1.3809	0.0089	(flat)	high Positive	high Positive
5	-1.9056	-0.0031	(flat)	high Negative	high Negative
6	0.3636	0.0457	Increased	Positive	Increased
7	-0.5979	-0.0160	Decreased	Negative	Decreased to negative

Table 5.6: Final emotion term attribution to learned clusters.

Cluster	Reward LOVE x5 pattern	State-value LOVE x6 pattern	LOVE 2:5x6 emotion attribution
0	Negative	high Negative	Distress
1	Average	high Positive	Optimism
2	Average	Negative	Neutral / slight Concern
3	Increased	Decreased to average	Satisfaction
4	Average	high Positive	high Optimism
5	Average	high Negative	Concern
6	Average	Increased	Excitement
7	Average	Decreased to negative	Fear

- Cluster 0: *Distress* (reward: below-average values; expectation: negative, well below $-\sigma$ values). The high variance of reward reflects very uneven values, which is subjectively perceived as negative due to the well-studied *loss aversion* principle (the pain of a loss is felt by individuals twice as intensively as the pleasure of an equivalent gain [187]).
- Cluster 1: *Optimism* (reward: average values; expectation: positive values around $+\sigma$).
- Cluster 2: *Neutral / slight concern* (reward: average values; expectation: below-average values). The most frequent emotion in this always uncertain environment (29.8% of the samples) falls closer to neutral than to concern, but the low expectation pattern justifies the compound naming.
- Cluster 3: *Satisfaction* (reward: increased values; expectation: decreased-to-average

values). The least frequent emotion, triggered upon reception of a significant reward, with expectations decreasing accordingly

- Cluster 4: *High optimism* (reward: average values; expectation: well above $+\sigma$ values). Technically, both 1 and 4 match Optimism, but expectations in 4 significantly exceed $+\sigma$.
- Cluster 5: *Concern* (reward: average values; expectation: negative, well below $-\sigma$ values).
- Cluster 6: *Excitement* (reward: average values; expectation: increased values).
- Cluster 7: *Fear* (reward: average values; expectation: decreased to negative values).

Finally, for real-time interpretation, the stepwise probabilities predicted by the emotional encoder were smoothened for more stable cluster attributions and easier external tracking: a moving average of 5 steps on values and a minimal probability of 0.9 as reclassification threshold (or, alternatively, a minimal amount of 10 consecutive attributions).

Visualization

For the 2D and 3D visualization of the learned emotional space, t-SNE (T-distributed Stochastic Neighbor Embedding) was used (see Fig. 4.7b for 2D and link in A.1 for a 3D animation). For the 2D representation with t-SNE, the *Scikit-learn* library [185] was used, with these parameters: *seed* = 90, *n_components* = 2, *perplexity* = 200, *init* = 'pca', *n_iter* = 2000.

Despite not technically required by the methodology, the use of colors, differentiating the classes learned by the emotional encoder, was instrumental for the final selection of the autoencoder.

5.2 Theoretical validation of LOVE profile terms

The principles described by the theoretical framework provide an initial foundation to associate LOVE profiles (idealized patterns of the latest values) with the best possible emotion terms in human language (see *Interpretation of the learned emotions*). However, given the difficulty of the task, their coherence was validated and refined both theoretically and experimentally. For the former, a sequence coherence test was run, following this methodology:

1. Attribute an initial term to each of the 30 profiles, based on said theoretical principles.
2. Run offline simulations of event-guided, plausible emotional sequences where each profile is a state:
 - Start from a stable state (for example, neutral).
 - Try different sequences involving positive / negative evolutions of rewards and state-value (or expectations), with brief explanatory narratives.

- Discard beyond-scope transitions (namely, positive \rightarrow increased, or negative \rightarrow decreased).
 - Avoid too many abrupt transitions (positive \rightarrow negative, negative \rightarrow positive).
 - End in stable or already visited states.
3. Review the resulting term sequences and repeat step 2 till fully natural transitions are obtained in all cases, leaving no pattern unused (ideally a few times).

For step 2., the sequence coherence test was iterated and refined over thirty-eight simulated emotional sequences with full profile coverage, such as:

Neutral \rightarrow (*An opportunity arises...*) \rightarrow Excitement \rightarrow (*and seems to hold.*) \rightarrow
 Optimism \rightarrow (*Suddenly the opportunity vanishes...*) \rightarrow Anger \rightarrow (*and we are back to normal.*) \rightarrow END.

For additional clarity, all the simulated emotional sequences tested can be seen in the annex *Theoretical Validation of LOVE Profile Terms: Extended data*.

5.3 Experimental validation of learned emotions with humans

The following methodology was used for the emotional attribution survey and its ensuing mapping to psychology literature references.

5.3.1 Emotional attribution test with humans

Dataset

A representative list of 3-6 second long sequences was automatically selected from the dataset in which one specific learned emotion clearly prevailed over the others (six for each of the eight learned emotions, totalling 48 sequences at different stages of the landing maneuver). This guaranteed equal representation of all eight emotions (which presented some difficulty in the case of 3, satisfaction, the least frequent emotion, often smoothened out at sequence end by the probability smoothing applied).

To reduce the effect of fatigue on raters, the test was randomly split in two evenly-distributed lists of 24 videos (A and B), the sequence order further randomized in two versions each, and raters assigned alternating versions (A1, A2, B1, B2, A1, etc.).

Tests with Lang’s SAM manikin

All participants were adult volunteers who were native Spanish speakers, recruited from diverse academic and professional backgrounds. A subset of 26 participants received university credits as acknowledgment for their involvement. The study was conducted online, in Spanish language, and after registering and giving legal consent, an introduction was displayed

explaining the dynamics of the study, the information shown during the sequences, the scoring system and the mission of the agent trying to land on a lunar base, described as a life-or-death task (see Fig. 5.7).

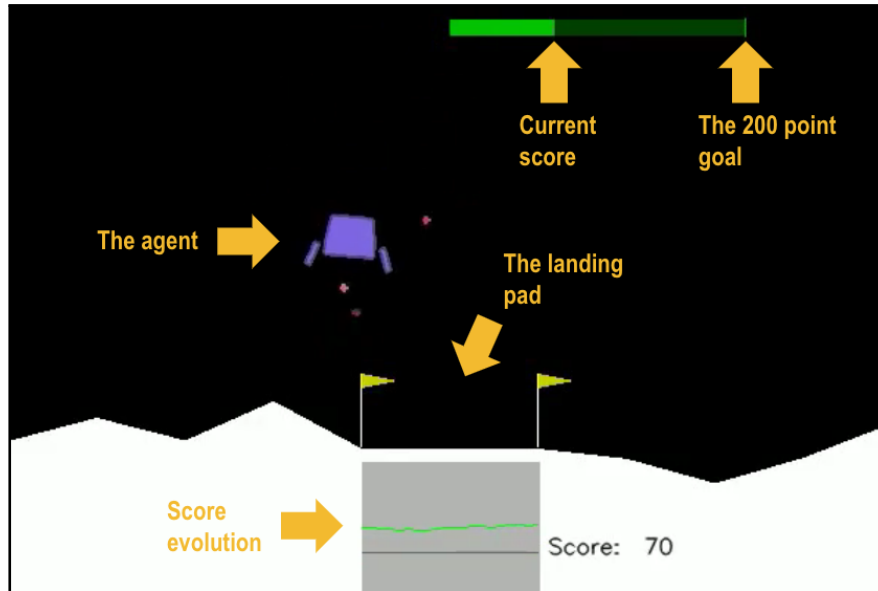


Figure 5.7: Information in a frame from one of the 48 rated sequences. The short videos showed the trained agent during landing, including its current score in relation with the 200 point target, as well as its last 20 values (the informative text and arrows in the figure were not present in the sequence).

Once familiarized with the agent’s task, Lang’s Self-Assessment Manikin (SAM) [175], a test extensively applied in psychological experiments and market research, was introduced to the participants. They were allowed to try it on two practice videos, characterizing a successful and a failed landing, whose respective results were discarded.

Finally, the subjects proceeded to the test, and for each of the 24 videos, reproduced on a separate screen, they were asked to describe the emotions they would associate to the state of the pilot at the end of each sequence using SAM. The test had no time limit, the videos could be played as many times as desired and ratings could be reviewed before the final submission.

SAM is a pictorial assessment technique that directly measures emotional responses on three main dimensions: pleasure, arousal and dominance [31], associated with a person’s affective reaction to a wide variety of stimuli, typically by rating each dimension from 1 to 9 on a Likert scale. Some SAM tests incorporate a collection of words positioned at the relevant end of each Semantic Differential scale to identify the anchors of each dimension to the subject [32]. These original terms [174] were predominantly translated into Spanish from the version by Gurbindo (1989) [188], with nuanced contributions from the French version by Detandt (2017) [189].

Fig. 5.8 shows a real screenshot with the graphic layout of the survey, and one of the videos rated during the test, while more detailed close-ups of the three Likert scales can be seen in annex *Experimental validation of learned emotions with humans: Extended data*. Videos were hosted on *YouTube*, and results collected with *Google Forms* during June 2023.

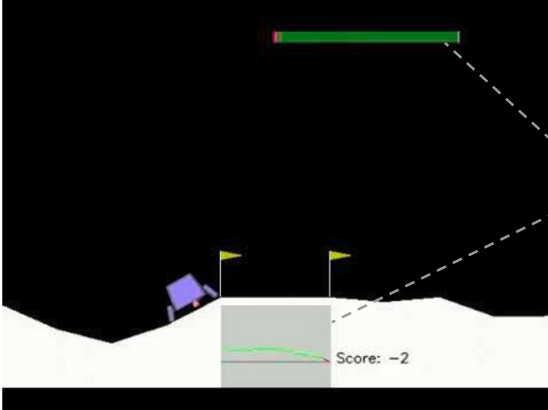
Sección 8 de 10

Video A01

Descripción (opcional)

Observa la secuencia y describe lo que sentiría el piloto al final:

(Puedes hacer "Replay" las veces que quieras.)



3-6 second long sequences, hosted in YouTube

Screen info included the latest 20 rewards and the total score

(A01) Placer

Introduce tu evaluación de esta componente.

Intelig

Contrariado

Insatisfecho

Melancólico

Desesperado

Aburrido

1

2

3

4

5

6

7

8

9

Feliz

Complacido

Satisfecho

Contento

Esperanzado

Divertido

1

2

3

4

5

6

7

8

9

(A01) Activación

Introduce tu evaluación de esta componente.

Relajado

Calmado

Lento

Apagado

Sobrio

No activado

1

2

3

4

5

6

7

8

9

Estimulado

Excitado

Frenético

Agitado

Muy despierto

Activado

1

2

3

4

5

6

7

8

9

(A01) Dominancia

Introduce tu evaluación de esta componente.

Dominado

Influenciable

Desvalido

Impresionado

Sumiso

Guiado

1

2

3

4

5

6

7

8

9

Dirigente

Influente

En control

Importante

Dominante

Autónomo

1

2

3

4

5

6

7

8

9

Figure 5.8: SAM, the classic graphic layout used for the rating of sequences.

80

Statistical significance

The study involved 96 raters aged 18 to 64 ($n = 96$), with a fairly even split of 53 males and 43 females. The majority of participants (89) held a University degree, providing a diverse pool for comprehensive statistical analysis.

Test reliability

In order to assess the reliability of the ratings, the ICC2k statistic was used (Intraclass Correlation Coefficient, two-way random effects model, absolute agreement). Results were assessed according to two commonly referred set of guidelines: (Koo, 2016) [176] and (Cicchetti, 1994) [177].

PAD values attributed to the 48 videos

Each of the 48 sequences was assigned a pleasure, arousal and dominance (PAD) triad of values, obtained as an average of the registered values, as shown in Fig. 4.13a. The values and correlations obtained are detailed in Table A.1 and Table 4.2 respectively.

PAD values attributed to the eight learned emotions

Similarly, each of the eight emotions was assigned a PAD triad of values as an aggregation over all raters from its six corresponding videos. The results can be visualized in Fig. 4.13b, and their values and correlations in Table 4.1 and Table 4.3 respectively.

Distinguishability of the learned emotions

To accurately validate the distinguishability of the eight emotions from their assigned PAD values, the robust Hotelling's T-squared statistical test was used, comparing the distribution of each pair of multivariate samples. The results are reflected in Table 4.5.

5.3.2 Mapping versus documented experimental accounts

The PAD values of each learned emotion, obtained from human ratings, were compared to select pivotal experimental findings in psychology literature. From a diverse array of studies and reports, priority was given to those offering well-documented values for the three referential dimensions, showcasing the highest significance and impact within their field. The references chosen, all of them detailing mean and standard deviation for all three dimensions, were:

1. Russell-Mehrabian (1977) [RM] [31]: 151 emotional states. The terms seem to have been selected by the authors.
2. Bradley-Lang (1999) [BL] [178]: Affective Norms for English Words (ANEW), including 1,034 terms. The list contains very heterogeneous terms (like *abduction*, *abortion*, *absurd*, *abundance*, etc.) along with actual emotions.

3. Redondo (2007) [RE] [179]: Spanish ANEW; 1,034 Spanish words corresponding to the original ANEW with newly obtained PAD values.
4. Landowska (2018) [LA] [180]: ANEW-MEHR; 112 words selected from the Russell & Mehrabian’s list with the PAD values from ANEW.
5. Scott (2019) [SC] [181]: Glasgow Norms, including 5,553 words. The list contains very heterogeneous terms (like *abattoir*, *abbey*, *abbreviate*, *abdicate*, etc.) along with actual emotions.

The task required overcoming a number of difficulties; firstly, we found a high degree of discrepancy in the terms included, heterogeneity of the emotional scopes, abundance of non-emotion terms, and other arbitrary peculiarities.

- Arbitrariness: the terminologies chosen by the authors, far from conforming a shared, standard set of emotions, often seemed inconsistently artificial (such as *weary with responsibility*, *quietly indignant*, *proud and lonely*, *snobbish and lonely*, in RM), or overly nuanced (for instance, *angry* and *angry but detached*; *hostile* and *hostile but controlled* in RM). All these terms were kept, despite producing somewhat idiosyncratic top matches.
- Heterogeneity of the emotional scopes: All authors seamlessly mingled complex affects (like social, moral, self-conscious) with primary, instantaneous or more basic emotions (like *fright*, *anxious*, *euphoria*). For the purpose of tagging this agent’s task (short life-or-death landing maneuvers), we did not map the emotions associated with social relationships, moral judgements or self-conscious reflections, along with a few overly redundant or vague ones and one case of bodily needs (including *guilty*, *kind*, *repentant*, *lonely*, *hungry* in RM and LA; *unfaithful*, *loyal*, *insolent*, *admired* in BL and RE; *dignity*, *paranoid*, *emotional* or *achievement* and *achieved*, *frightened* and *fright* in SC).
- Non-emotions: The scope of some references was not limited to emotion terms, including all sorts of concepts (such as *butter*, *cemetery*, *chair* in BL and RE; *abdominal*, *apple* (*fruit*) or *musketees* in SC), which were not used for emotional interpretation.

The terminology for the three dimensions has also historically differed among authors (namely, pleasure/valence; arousal/activation; dominance/control); however, we found that the more traditional (pleasure-arousal-dominance) model was easier to articulate and comprehend for non-expert participants.

As for the methodologies followed by the authors to obtain the PAD values, they also differed in format and profile of the participants, which probably contributed to the variance found in the reported values (see below the different values reported for “angry” in the rank of [-1, 1]):

Author	PAD values reported for “angry”
Russell-Mehrabian (1977)	(-0.510, 0.590, 0.250)
Bradley-Lang (1999), Landowska (2018)	(-0.538, 0.543, 0.138)
Redondo (2007)	(-0.700, 0.403, -0.290)
Scott (2019)	(-0.652, 0.227, 0.105)

Finally, despite all authors reporting standard deviations and number of samples, the lack of covariance matrices limited the applicability of standard statistical tests to compare two distributions, like Hotelling T-squared. To address this, we applied three different methods to map the PAD distributions obtained from our test for each emotion (sample 1) against reported PAD mean and standard deviation values (sample 2), often with unequal results in each table:

- Method 1: Hotelling T-squared test, assuming three independent variables in sample 2 (diagonal covariance matrix).
- Method 2: Hotelling T-squared test, assuming sample 2 had the same covariance matrix as sample 1.
- Method 3: Euclidean distance, comparing only the means.

To obtain the final mapping of each learned emotion, drawing inspiration from the ensemble of models concept, we independently mapped it against each reference table, identifying its three top matches, and then we merged the five top matches across authors into a semantic collage.

Chapter 6

Conclusions

6.1 Discussion

6.1.1 Key achievements

A cross-disciplinary mathematical formalization of emotions

Emotions are arguably among the most elusive concepts encountered across various disciplines, including psychology, neuroscience, biology, and AI [190][191][192][105]. The cross-disciplinary significance of the framework proposed here lies in its capacity to effectively associate them with quantifiable, mathematically describable temporal patterns in critical values for a living being, akin to a “Periodic Table” for emotions. This systematic categorization allows for an organized comprehension, analysis and synthesis of their complex spectrum as observed in nature, as well as their integration into machine learning frameworks—a crucial step toward the goal of enabling artificial agents to generate and process their own emotions [193][194].

A generic self-learning emotional framework

This work diverges from, and complements, the predominant focus in affective computing and related areas—a rapidly evolving field thanks to the convergence of multiple technologies [87][88][89][105]. Rather than the recognition or interpretation of human affects through external sensors (such as a user’s voice, text, or image) or their verbal/visual communication to users, this approach directly addresses the core concept of “emotion” itself, its formalization, elicitation, and integration within functional agents.

Furthermore, standing apart from previous predominantly theoretical, case-specific, arbitrary or hard-coded emotional models [78][81][106][107][108][109][65][94][110][111][112] this generic methodology addresses emotion learning in an unsupervised manner from first principles within the RL framework, devoid of any preconceptions, and encompassing the entire emotional spectrum. This spectrum is spontaneously encoded as temporal patterns, learned in an unsupervised fashion during agent’s interactions with computational simulations of an environment. The relevance of achieving such a robust, transferable, fully learned approach has been recognized in the field of AI as a core challenge for the future [78][109][195][196][65][84].

The framework effectively integrates both discrete and continuous aspects of emotions by representing them as low-dimensional latent points, that are subsequently interpreted through probabilistic clustering. Remarkably, the learned emotional spectrum is shaped in congruence with the environment’s dynamics and the agent’s cognitive capabilities—rather than by pre-defined categories or ad-hoc specifications—enabling it to scale naturally with the complexity of the agent-environment interaction. Future testing in new environments, producing diverse emotional spectra, will further validate or challenge these findings.

A novel mapping of RL values to human affective dimensions

To the best of our knowledge, this work is the first to analytically establish a concrete relationship between RL-derived values, such as reward and state-value, and PAD (Pleasure-Arousal-Dominance) ratings from psychological models during an agent’s task execution. By mapping these observed values onto established affective dimensions, the interpretability of agent behaviors is enhanced, providing a psychologically grounded framework that connects machine emotions to classic human affective theory. This could pave the way for richer integration of psychological insights into AI emotional frameworks, opening new avenues for intuitive AI-human emotional interactions.

However, a limitation of the method applied is its reliance on human attribution, which may introduce subjective biases or cultural variability in the emotional interpretations. The relationship was established through an experimental survey in which human participants attributed PAD values to emotional states inferred from the agent’s behavior. Future work could focus on refining or fully automating this process to reduce reliance on human input, for instance, by leveraging dynamic sequence analysis, or documented PAD values in psychology literature.

A verifiable emotional framework

Crucially, the proposed framework is both analytically and experimentally verifiable, as demonstrated in the *Results*. The recognizability of the learned emotions is remarkable: the synthetic emotions elicited during the agent’s environmental interactions are situationally coherent, distinguishable to independent observers, and directly relatable to empirically documented characteristics of natural emotions.

To elaborate, the emotional spectrum learned during unlabelled agent interactions was automatically categorized into eight basic emotions, such as distress, optimism, and satisfaction, obtaining strong empirical confirmation from subjective human observers. Participants, who were unaware of the emotions attributed by the system, rated pleasure-arousal-dominance (PAD) values during 48 sequences. The ratings exhibited high ICC2k agreement rates and, most relevantly, showed a strong alignment with the eight learned emotions (high statistical distinguishability in Hotelling’s T-squared tests), reflecting situational coherence as well as a natural progression from negative to positive emotions. Additionally, the PAD rates demonstrated significant semantic agreement with five key psychological studies in the literature.

Indeed, the emotions we have experimentally simulated show notable similarities to their natural counterparts. The analysis of fully stepwise-interpreted episodes also demonstrated

that the learned emotions are relatable to documented characteristics of natural emotions, such as the relevance of recent events in instantaneous emotional responses and their homeostasis-driven elicitation and decay, with average reward and expectations serving as referential values. Additionally, we observed natural emotional progressions that align well with external events and subjective appraisals, as well as temporal coexistence, resembling states of shock and confusion during turbulent scenarios.

An extension of the Reinforcement Learning framework

Lastly, and though not experimentally covered in our first experiments, we foresee great promise in the integration of emotional agents in the field of Reinforcement Learning. The enrichment of the agent’s perceived state with the learned synthetic emotions has the potential to increase its expected utility. Namely, the new policy could yield advantages analogous to those observed in nature, such as behavioral benefits (rapid, valuable appraisals to adjust conduct, attention and goals), efficient learning (focusing on emotionally significant sequences), and social competences (enabling agent-to-agent communication, with better collective performance).

Furthermore, the integration of the synthetic emotion—a rich vectorial representation—into the agent’s behavior is straightforward, as it merely involves a simple state extension. This process does not require external interpretation or adaptation to the task (see *Extensions of the actor-critic method*).

6.1.2 Limitations and future work

Richer emotional ranges and dynamics

Currently though, the framework has only been jointly formalized and tested on short-term “basic” emotions, spanning over a maximum of thirty potential emotional reference profiles, and has not been proved to capture richer retrospective dynamics like habituation (depending on more remote values) and extinction (requiring continuous learning of the agent’s functions), as well as emotional associations to external objects or subjects (like facilitators or blockers), knowledge-related emotions (like surprise) and higher-order emotions (such as social, moral or self-conscious).

“Habituation” may simply require the inclusion of moving averages to the observed values—with continuous homeostatic renormalization—possibly combined with an expanded, smooth emotional window.

“Extinction”, though, depends on adjustments in the agent’s appraisal, such as continuous training of the agent’s subjective state-value function on newly-neutral stimuli.

More sophisticated cognitive abilities are required for the rest, as well as higher-order interpretability mappings and the coexistence of multiple-range windows (giving rise to concurrent, higher-order emotions from life-long temporal patterns), that are not discussed here

Refinement of the interpretability mappings

In relation with the introduced mappings, the tentative emotional attributions assigned might benefit from further analysis and testing of the numerous profiles not experimentally

demonstrated in this first case study. Future research might also explore extended patterns such as high/low decreases, high-/low-variance policy outputs, and their potentially uneven prevalence.

More biologically-inspired environments

More ambitious research might explore the extension of the framework to multiple reward setups, inspired by how living beings experience and balance very heterogeneous and often contradictory signals. The scope should encompass physical sensations like pain, hunger or sensual pleasure, as well as higher-order feelings like calmness, love, and self-realization.

These tests could be expanded to include partially observable environments, continuous learning scenarios (beyond episodic tasks) and multi-agent setups (exploiting the social role that elicited / displayed emotions play). Exploring these lines could further enrich our understanding and lead to more comprehensive models, setting the stage for the most innovative contributions outlined in this work.

6.1.3 Ethical considerations

Finally, in addressing ethical considerations, particularly whether machines should have emotions, it is important to clarify that the emotions elicited in our study are synthetic, mathematically derived from learned patterns, and no AI agent experienced suffering or enjoyment during simulations.

However, the possibility of their spontaneous emergence in AI should not be overlooked. Early evidence indicates spontaneous sentiment representation in apparently unrelated tasks—like next character prediction [138]. Furthermore, within the RL field, it has been suggested that emotions might emerge in deep neural network layers as a byproduct of the generic objective of reward maximization [139].

In such scenarios, our framework offers invaluable interpretability tools to identify and understand these emergent emotions. Looking ahead, this clarity could be extended to computational models of empathy, enabling predictions of other agents' or humans' emotions based on their approximate states and values.

6.2 Final thoughts

We have introduced here a universal emotional language learned from first principles in Reinforcement Learning and inspired by primary cognitive variables like reward/punishment signals and their predicted future values. The supporting theoretical framework and demonstrated methodology are, to the best of our knowledge, the first successful attempt to formally describe and synthesize a comprehensive range of recognizable, functional emotions of comparable characteristics to those of living beings. This pioneering work not only bridges the gap between artificial and natural emotional processes, but also opens new avenues for exploring the intersection of cognition, emotion, and machine learning. We believe that the generality of our framework could lay a foundation upon which further research and applications will lead us toward emotional machines that think and act more like us.

Appendix A

Extended data

A.1 Code and data

Access

The datasets, step-by-step code and some additional information may be found in the public GitHub repository at <https://github.com/Alberto-Hache/love-emotional-framework>.

Licensing

The code and data repository associated with this PhD thesis are licensed under the MIT License. This permissive open-source license allows anyone to freely use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the software and associated documentation, provided that the original copyright notice and this permission notice are included in all copies or substantial portions of the software.

In summary, this means that:

- You can use the software for any purpose.
- You can distribute the software.
- You can modify the software and distribute your modifications.
- You can sublicense the software.

However, the software is provided “as is”, without any warranty of any kind. This means that the author is not liable for any damages or issues that arise from using it.

For the full text of the MIT License, please refer to the repository where the code and data are hosted.

A.2 Application of the framework on a practical case study: Extended data

A.2.1 Alternative conventional RL agents trained

We show here the performance registered by different Reinforcement Learning (RL) agents trained for the case study with alternative choices of some key hyperparameters (while keeping fixed the following ones: γ : 0.99, λ : 0.97, policy learning rate: 0.0003, state-value function learning rate: 0.001, target KL: 0.01).

The initial tests compared architectures with two hidden layers (of 32 or 64 neurons) and two types of activation functions ($ReLU$ and \tanh). Figure A.1 shows their comparative performance, with the vertical axis showing the total accumulated reward at episode end, and the horizontal axis showing the total amount of stepwise interactions experienced by the agent. Each combination was tested with three different random seeds to initialize the environments, and allowed to run for 40,000 time-steps (with typical episodes lasting 250-200 time-steps, and a maximum of 1,000 for failure). The best performance was obtained by the architecture: (64, 64), with activation function: $ReLU$.

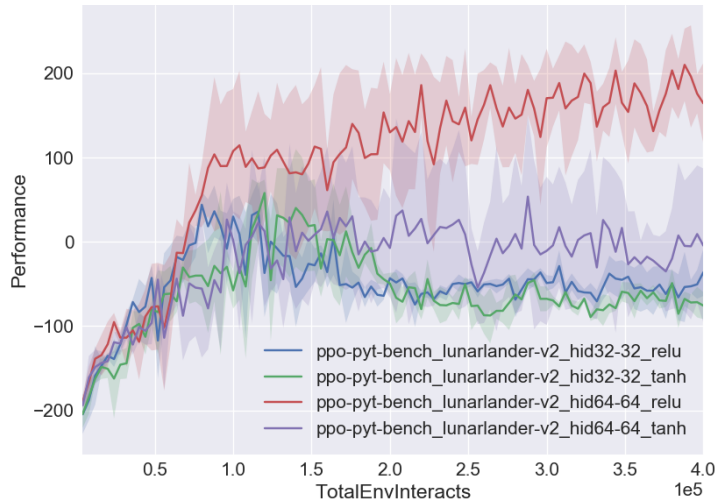


Figure A.1: Quick initial exploration of hyperparameters.

Figure A.2a shows the tests with denser hidden layers, of (128, 64), (64, 32) and (64, 64) each, using the $ReLU$ activation function, tested with three different random seeds. The best performance was obtained by the architecture (64, 64).

Architectures with more hidden layers were tested as well, like the one shown in Figure A.2b, not registering better results than the shallower (64, 64) one.

Finally, a quick search for an optimal random seed over ten alternatives resulted in seed = 10 as the best choice, reaching total rewards around 200 earlier, as shown in Figure A.3.

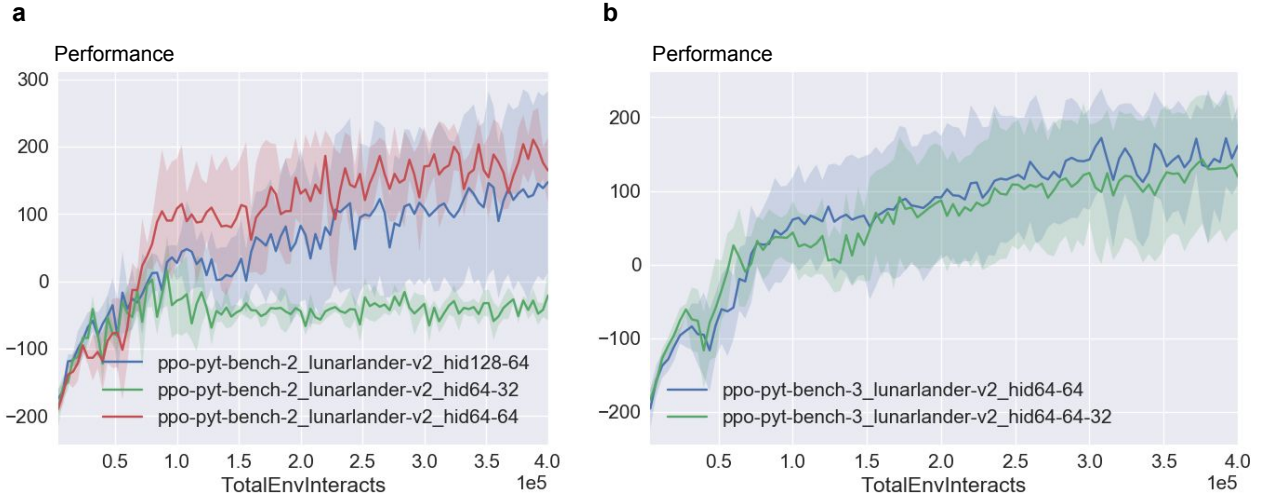


Figure A.2: Alternative architectures tested. **a**, Tests with denser hidden layers. **b**, Tests with more than two hidden layers.

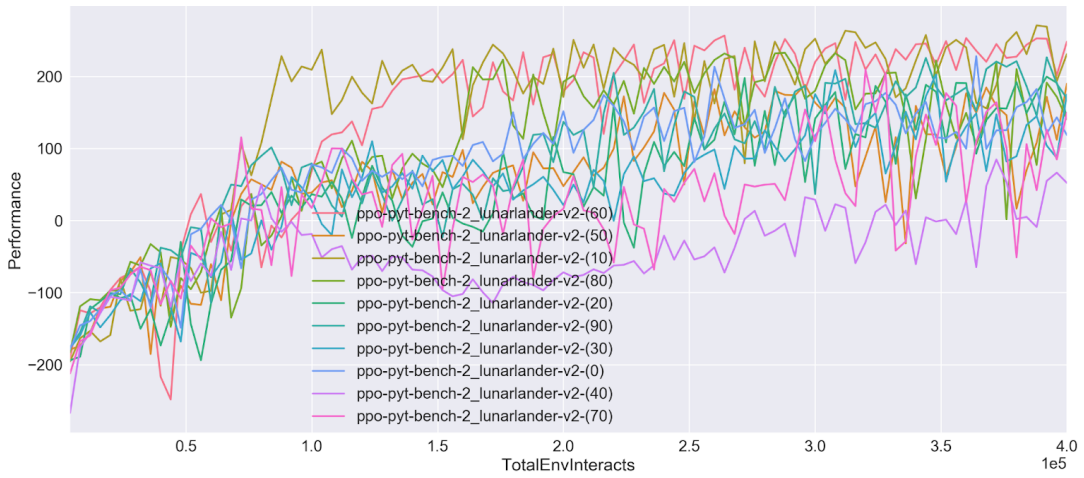


Figure A.3: Tests with several random seeds on the final architecture.

A.2.2 Alternative clustering results

Figure A.4 shows some of the different clustering models obtained during the interpretation phase. We found a high overall consistency of the resulting clusters across most alternatives, but live testing of the corresponding classifiers revealed that two of the minority classes, highlighted in the figures, were frequently merged into a single class despite happening in visibly differentiated situations.

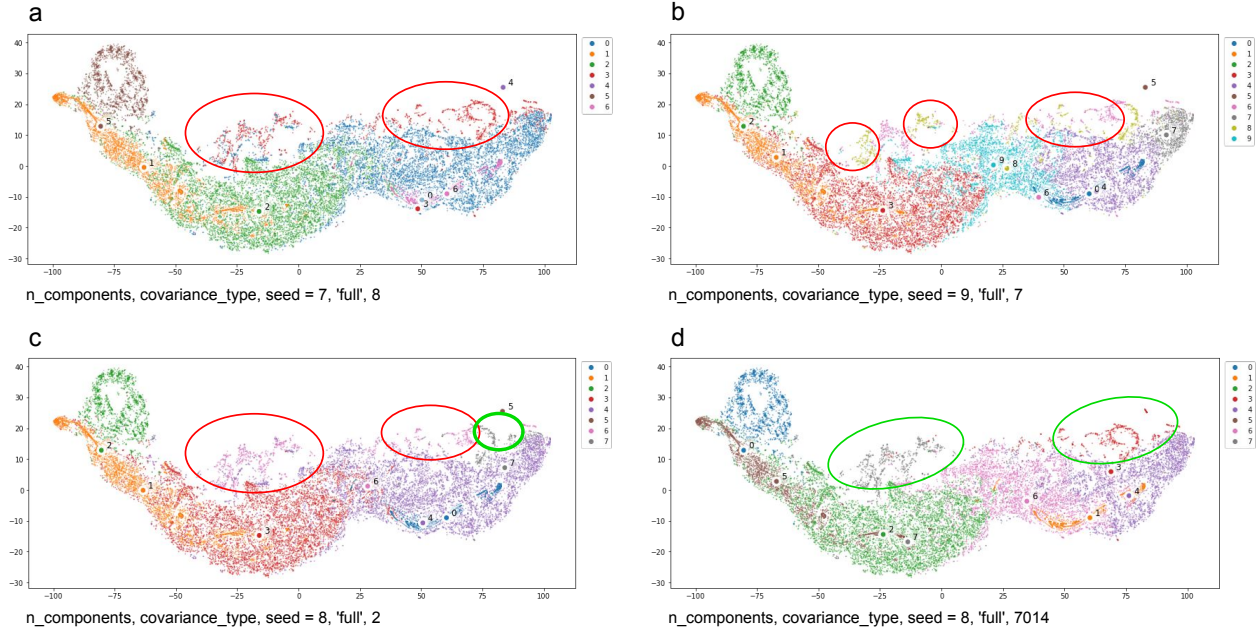


Figure A.4: Alternative clustering options evaluated. Four clustering results obtained during the interpretation of the learned emotional space. **a**, Seven classes, resulting in a too generic classification that missed the differentiation between the two highlighted regions. **b**, This option, with ten classes, only obtained a partial split between the highlighted areas. **c**, In this case, with eight classes, only a partial split is obtained on the top right region. **d**, The final clustering, chosen for its balanced number of classes and the nearly perfect differentiation of the two highlighted regions.

A.2.3 3D visualization of the emotional spectrum

Figure A.5 shows a 360° rotation of a tridimensional t-SNE (T-distributed Stochastic Neighbor Embedding) representation of the emotional spectrum learned during the case study.

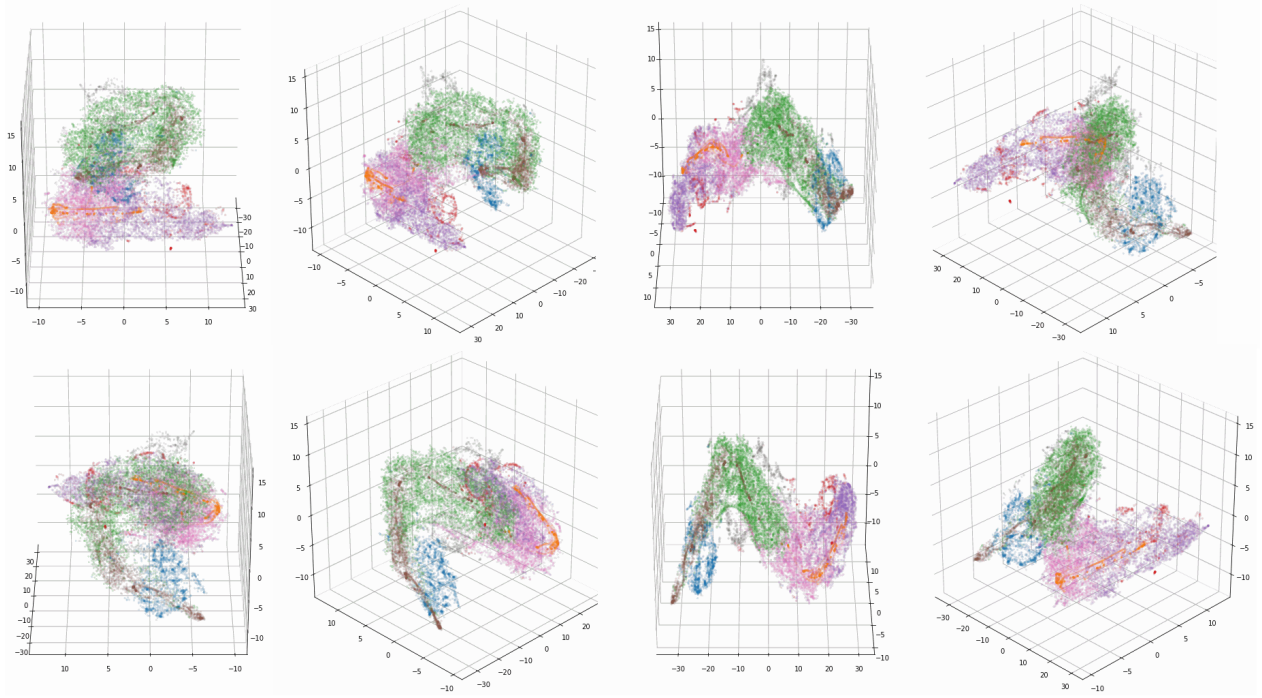


Figure A.5: 3D representation of the eight identified classes.

A.3 Theoretical Validation of LOVE Profile Terms: Extended data

This section illustrates all the sequence coherence tests conducted to refine and theoretically validate the profile terms of the LOVE 2:5x6 interpretability mapping.

Series 1.1 and 1.2

(Fig. A.6 and A.7)

- Description: Variations starting with an opportunity.
- Sequences: Neutral \rightarrow Excitement \rightarrow Optimism \rightarrow ...
- Coverage: 18 fully coherent paths; 63 emotional states.

Series 2.1

(Fig. A.8)

- Description: Variations starting with a threat.
- Sequences: Neutral \rightarrow Fear \rightarrow ...
- Coverage: 11 fully coherent paths; 34 emotional states.

Series 3.1, 3.2 and 3.3

(Fig. A.9, A.10, and A.11)

- Description: Focus Studies
- Coverage: 9 fully coherent paths; 34 emotional states.
- Sub-Series:
 - Series 3.1: Study of the symmetry between anticipated quick positive/negative rewards.
 - Series 3.2: Evolution of a supposedly temporary negative reward.
 - Series 3.3: Paths including Contentment.

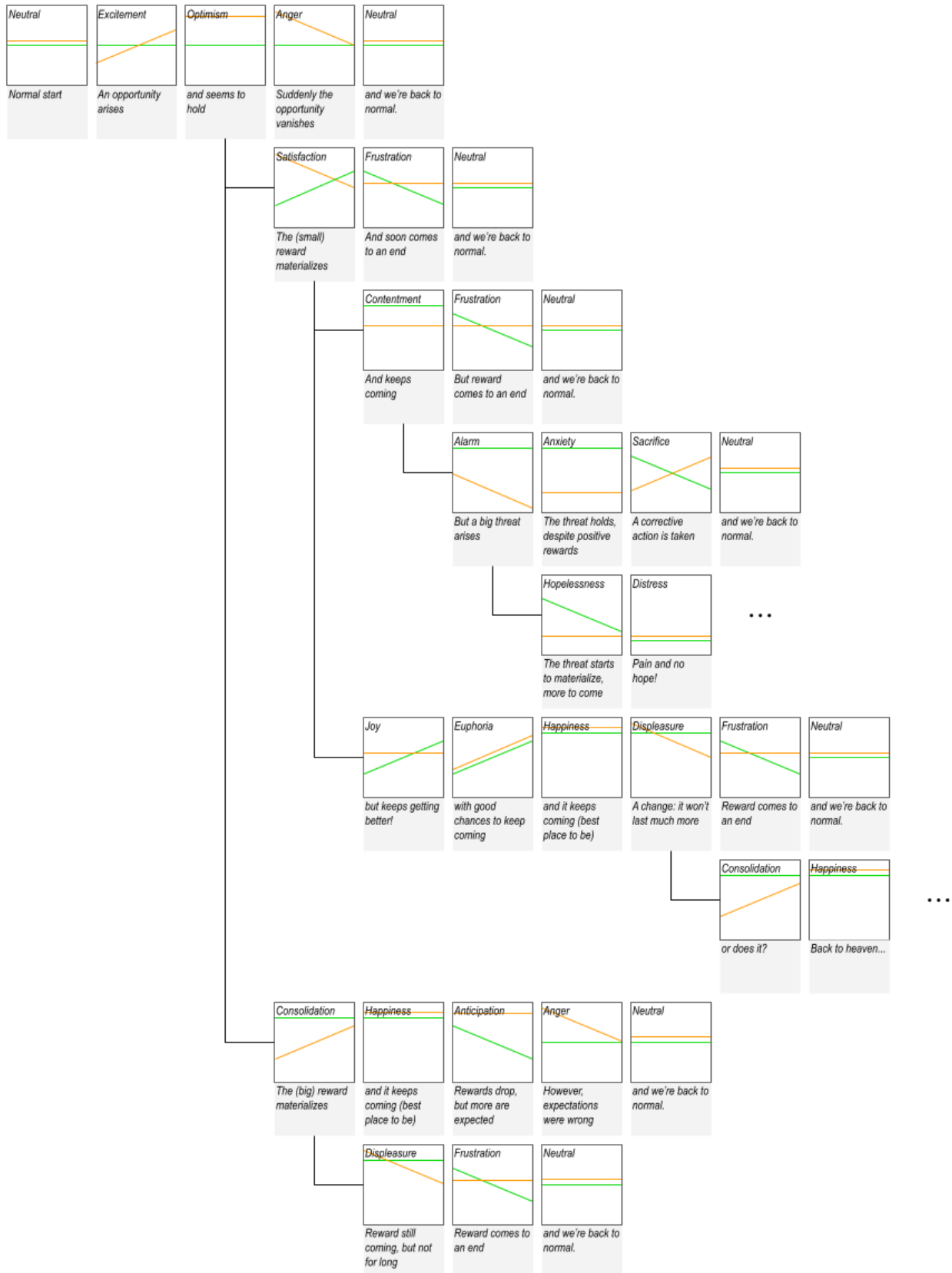


Figure A.6: Sequence coherence test: Series 1.1.

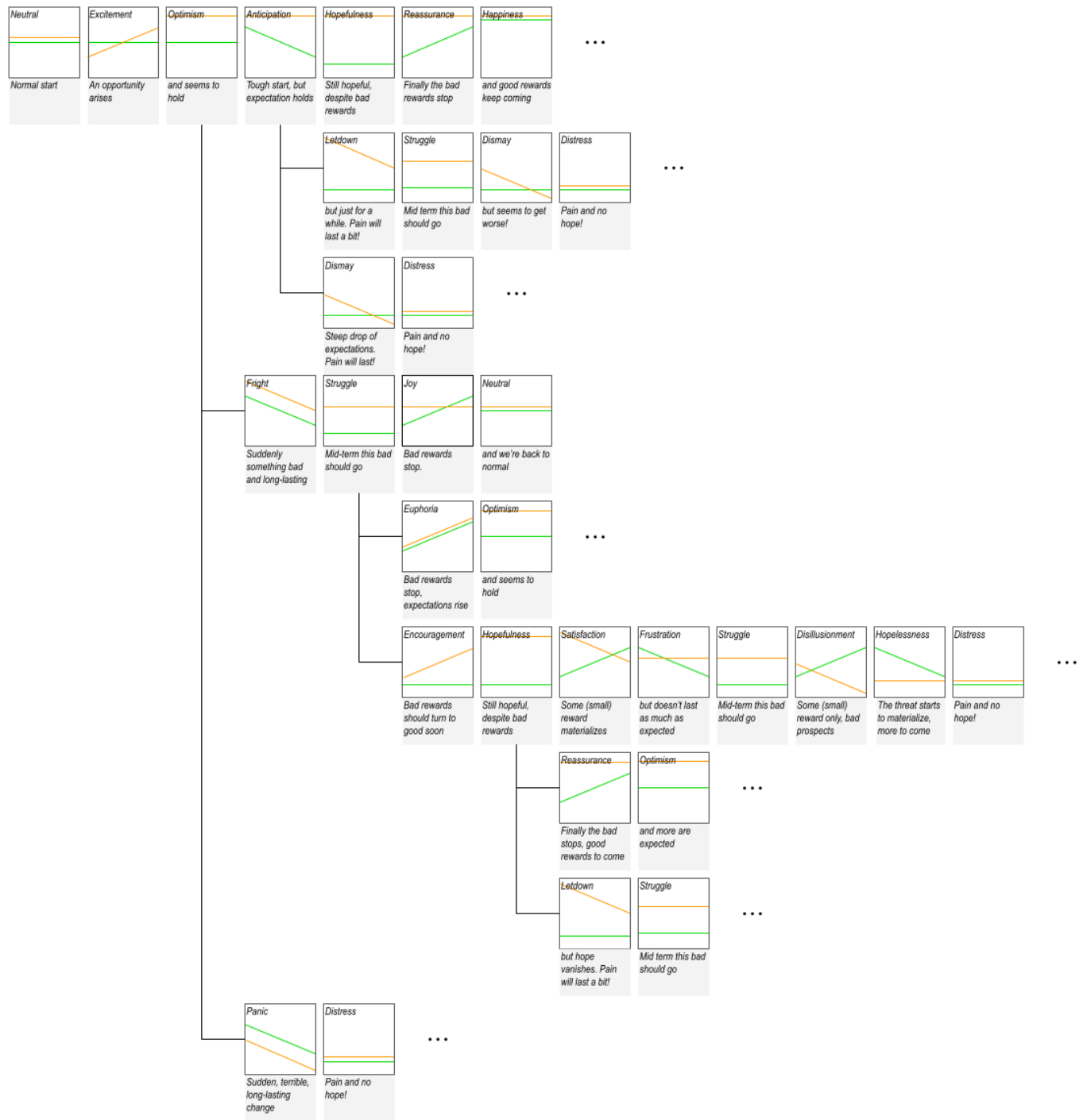


Figure A.7: Sequence coherence test: Series 1.2.

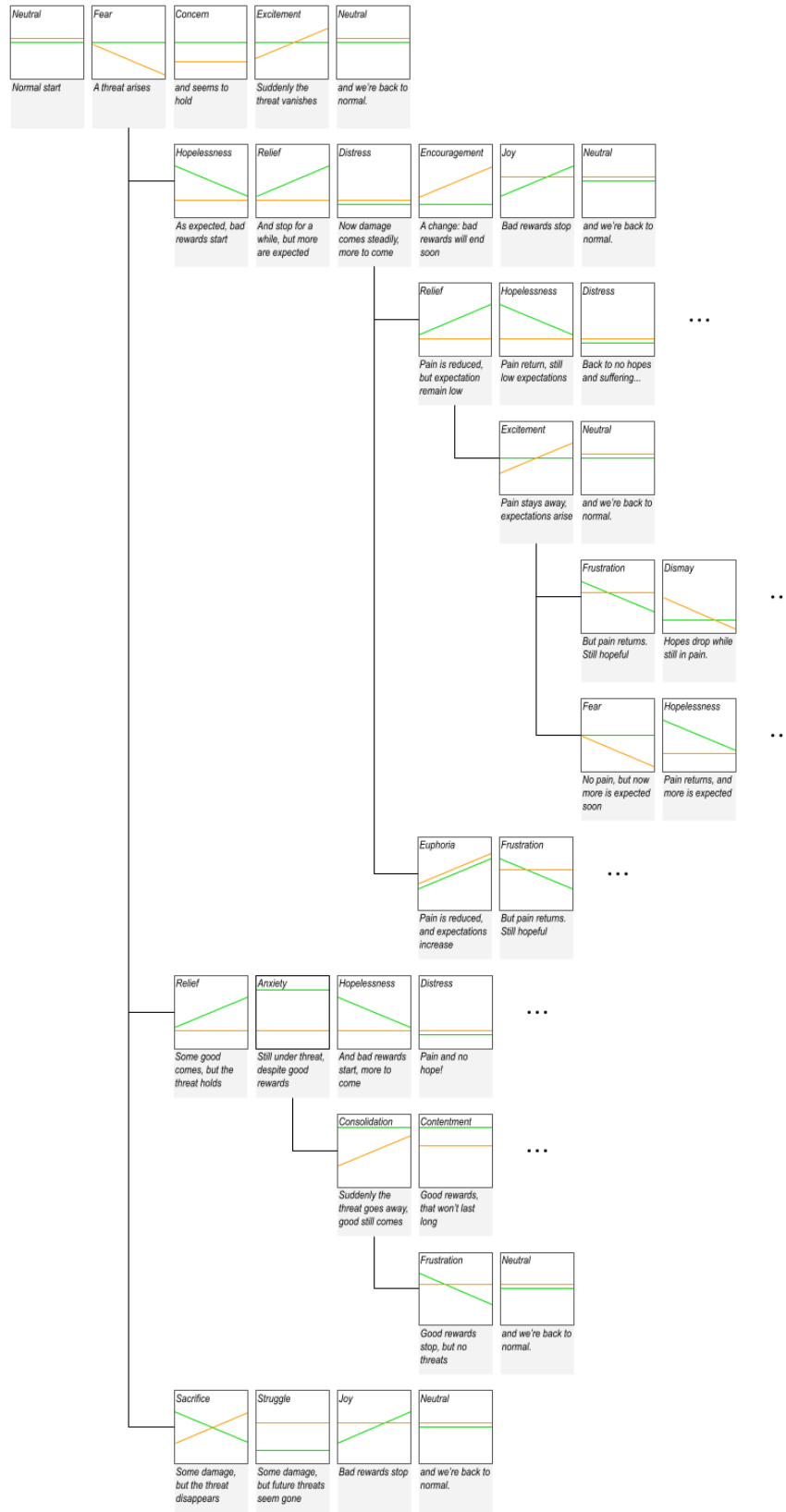


Figure A.8: Sequence coherence test: Series 2.1.

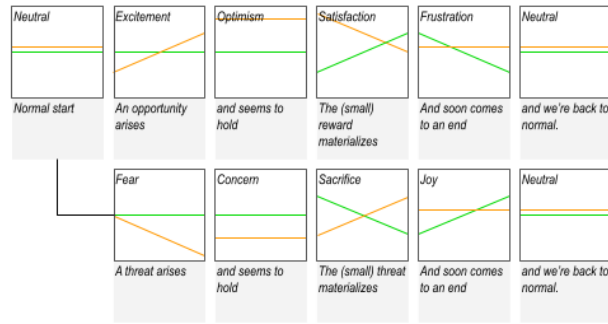


Figure A.9: Sequence coherence test: Series 3.1.

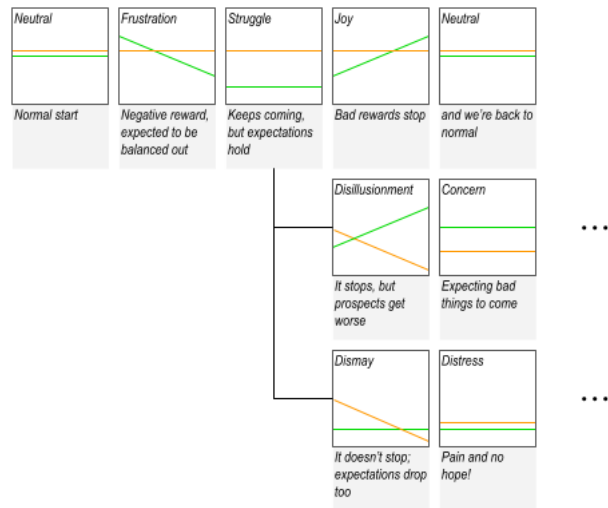


Figure A.10: Sequence coherence test: Series 3.2.

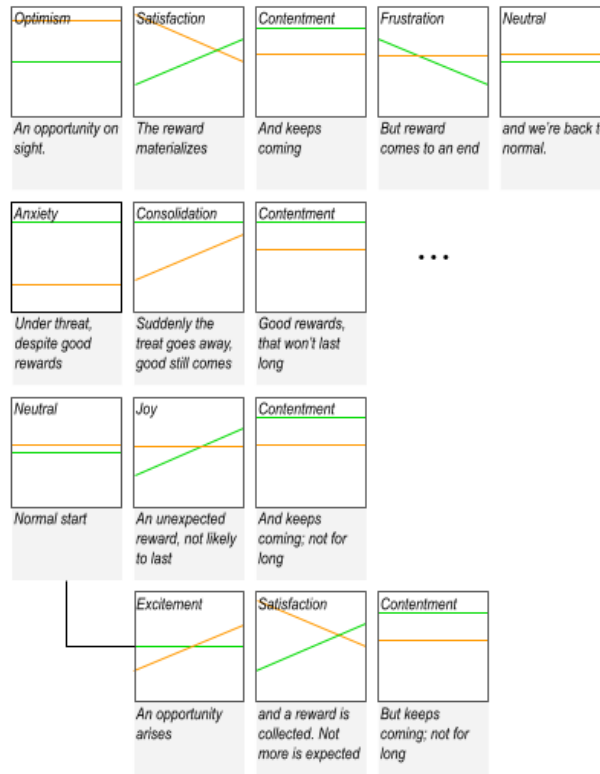


Figure A.11: Sequence coherence test: Series 3.3.

A.4 Experimental validation of learned emotions with humans: Extended data

A.4.1 Likert scales used for emotion attribution in Spanish

Likert scales for Pleasure, Arousal and Dominance in Lang’s Self-Assessment Manikin (SAM) included a collection of words at the ends (anchors of each dimension), that were adapted to Spanish.

Placer: Introduce tu evaluación de esta componente. *

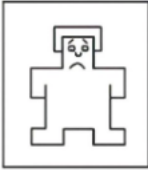
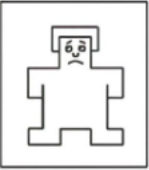
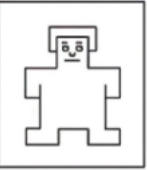
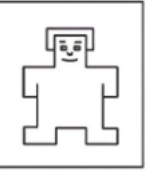
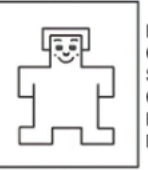
Infeliz Contrariado Insatisfecho Melancólico Desesperado Aburrido						Feliz Complacido Satisfecho Contento Esperanzado Divertido		
1	2	3	4	5	6	7	8	9
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Figure A.12: Likert scale for the “Pleasure” dimension.

Activación: Introduce tu evaluación de esta componente. *

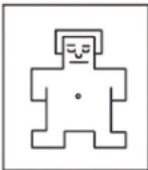
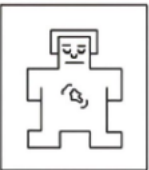
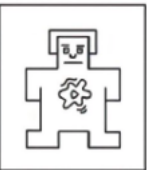
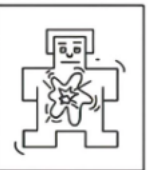
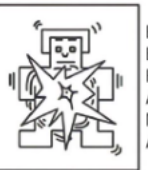
Relajado Calmado Lento Apagado Soñoliento No activado						Estimulado Excitado Frenético Agitado Muy despierto Activado		
1	2	3	4	5	6	7	8	9
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Figure A.13: Likert scale for the “Activation” dimension.

Dominancia: Introduce tu evaluación de esta componente. *




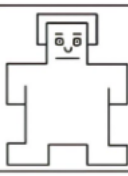
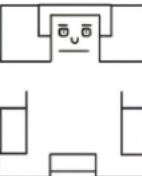
Dominado Influyente Desvalido Impresionado Sumiso Guiado						Dirigente Influyente En control Importante Dominante Autónomo		
1	2	3	4	5	6	7	8	9
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Figure A.14: Likert scale for the “Dominance” dimension.

A.4.2 PAD values attributed to the 48 videos

Table A.1 shows the pleasure-arousal-dominance (PAD) rates attributed to each of the 48 videos aggregated over all raters. Each learned emotion (from clusters 0 to 7), concealed to the participants, was represented by six short videos.

Table A.1: PAD rates for each video in range [1, 9] aggregated over all raters.

Video	Emotion	Pleasure		Arousal		Dominance	
		(mean)	(stdev)	(mean)	(stdev)	(mean)	(stdev)
A01	0	1.711538	0.976921	7.442308	2.199942	1.903846	1.332249
A02	0	1.230769	0.703364	8.115385	2.210969	1.480769	1.350242
A03	0	3.384615	1.561219	5.942308	2.118207	3.173077	1.617565
A04	1	8.211538	1.242018	5.288462	2.483961	7.980769	1.378979
A05	1	6.923077	1.569890	5.596154	1.763015	6.692308	1.698194
A06	1	8.211538	1.242018	5.096154	2.491237	8.115385	1.338038
A07	2	3.480769	1.552864	6.750000	2.085195	3.673077	1.700302
A08	2	5.615385	1.457285	6.134615	1.940576	5.500000	1.650906
A09	2	4.038462	1.825329	7.038462	1.596093	4.057692	1.808623
A10	3	8.500000	1.019419	5.019231	2.532072	8.403846	1.256506
A11	3	8.519231	1.244445	5.076923	2.936225	8.269231	1.509647
A12	3	8.115385	1.423250	5.557692	2.476663	7.788462	1.718829
A13	4	7.538462	1.954999	5.826923	2.332390	7.423077	2.295316
A14	4	7.673077	1.279111	5.692308	2.305479	7.192308	1.633455
A15	4	7.750000	1.631492	4.846154	2.452567	7.480769	2.033932
A16	5	4.384615	1.816632	6.519231	2.033932	4.557692	1.984385
A17	5	3.961538	1.385662	5.884615	1.986759	4.134615	2.067762
A18	5	3.038462	1.759696	6.923077	2.094307	3.115385	1.652732
A19	6	4.634615	1.680673	6.230769	1.722007	4.423077	1.933666
A20	6	6.500000	1.995092	6.057692	2.127444	6.846154	1.944168
A21	6	5.057692	1.719707	7.038462	1.888682	4.942308	2.090252
A22	7	3.423077	2.190408	5.923077	2.416010	3.192308	1.950751
A23	7	3.942308	1.764725	6.692308	1.754976	3.807692	1.749380
A24	7	3.076923	1.412079	6.442308	2.145798	3.173077	1.801102
B01	0	2.363636	1.699397	7.318182	2.066035	2.363636	1.526371
B02	0	4.204545	2.063859	5.931818	1.945749	3.522727	1.591966
B03	0	3.068182	1.909556	6.863636	1.959959	2.727273	1.436463
B04	1	7.636364	1.348400	5.477273	2.337646	7.090909	1.877878
B05	1	5.863636	1.936082	5.590909	1.821298	6.045455	1.790862
B06	1	7.204545	1.439954	5.840909	2.090323	6.409091	2.127540
B07	2	5.272727	1.436463	7.318182	1.307807	4.818182	1.742395
B08	2	4.409091	1.980347	6.386364	2.037051	4.409091	1.920734
B09	2	5.704545	1.439954	6.272727	1.689415	5.522727	1.620919
B10	3	8.818182	0.581607	5.500000	2.782838	8.636364	0.942311
B11	3	8.363636	0.685087	5.681818	2.310469	8.181818	0.994701
B12	3	8.272727	0.996824	6.227273	2.208479	8.068182	1.387615
B13	4	7.000000	1.524986	6.000000	2.090955	6.886364	1.781245
B14	4	7.431818	1.969509	6.090909	2.055262	7.568182	1.822313
B15	4	6.454545	1.810236	6.863636	1.664833	6.454545	2.028340
B16	5	3.590909	1.834022	7.090909	1.582141	3.181818	1.768888
B17	5	3.886364	1.768141	7.409091	1.574775	3.772727	1.903041
B18	5	4.227273	2.270781	4.795455	2.247268	4.181818	2.191662
B19	6	5.431818	1.619614	6.795455	1.862476	5.318182	1.681261
B20	6	7.772727	1.411969	6.113636	2.071017	7.568182	1.453108
B21	6	2.272727	1.420181	6.386364	2.553756	2.590909	1.674962
B22	7	2.545455	1.454017	7.181818	1.781986	2.636364	1.495589
B23	7	2.772727	1.178561	6.681818	1.877315	2.954545	1.524292
B24	7	3.000000	1.656558	5.818182	2.244090	3.000000	1.540160

Appendix B

Research publications

B.1 Journal papers

Publication

Hernández-Marcos, A., Ros, E. A generic self-learning emotional framework for machines. *Scientific Reports* **14**, 25858 (2024). <https://doi.org/10.1038/s41598-024-72817-x>

Received: 03 Apr 2024

Accepted: 10 Sep 2024

Published: 28 Oct 2024

Journal Information

- **Journal:** *Scientific Reports*
- **Publisher:** Nature Publishing Group
- **ISSN:** 2045-2322
- **Impact Factor (JIF):** 3.8 (2023)
- **5-year Impact Factor:** 4.3
- **Best Ranking:** Multidisciplinary Sciences (Q1), Percentage rank: 81.3%
- **H-Index:** 315
- **Country:** United Kingdom

Scientific Reports has a 2-year impact factor of 3.8 (2023), and is the 5th most-cited journal in the world, with more than 734,000 citations in 2023*.

*2023 Journal Citation Reports® Science Edition (Clarivate Analytics, 2024).

B.2 Other dissemination activities

Keynote Speaker at “The Discussion Club” Session

- **Title:** “A generic self-learning emotional framework for Reinforcement Learning agents”
- **Organizer:** BBVA - AI Factory, Av. Manoteras 44, 28050 Madrid, Spain
- **Date:** October 1st, 2024
- **Duration:** 1 hour
- **Attendees:** 42

Keynote Speaker at the Department of Information and Communication Technologies

- **Title:** “A generic self-learning emotional framework for Reinforcement Learning agents”
- **Organizer:** Eduardo Ros Vidal - Department of Information and Communication Technologies - Neural Engineering and Bio-inspired Integrated Systems
- **Date:** June 4th, 2024
- **Duration:** 1 hour
- **Attendees:** 17

Keynote Speaker at “Computation-Neuroscience-ML-Robotics” Meeting

- **Title:** “Un marco genérico para el aprendizaje automático de emociones funcionales en agentes con Aprendizaje por Refuerzo - State review”
- **Organizer:** Eduardo Ros Vidal - Department of Information and Communication Technologies - Neural Engineering and Bio-inspired Integrated Systems
- **Date:** April 10th, 2023
- **Duration:** 1 hour
- **Attendees:** 22

Appendix C

Resumen de la tesis

Introducción

Las emociones, tanto en humanos como en otros seres vivos, son una ventaja evolutiva clave para la adaptación al entorno. Lejos de ser perturbaciones irracionales o ajenas a la toma de decisiones, como históricamente se ha entendido en la tradición filosófica occidental, están profundamente integradas en los procesos cognitivos y conductuales. Teorías contemporáneas de la biología y la neurociencia, como las de Charles Darwin y Antonio Damásio, subrayan la importancia de las emociones en la supervivencia y la toma de decisiones, demostrando cómo daños en los sistemas emocionales pueden perjudicar gravemente las capacidades cognitivas.

Desde el punto de vista biológico, Darwin (1872) fue pionero al argumentar que las emociones humanas y animales tienen una base evolutiva común, siendo esenciales para la supervivencia. En la neurociencia, Damásio (1994) mostró que las emociones son cruciales para la toma de decisiones racionales, mientras que estudios de Jaak Panksepp han identificado sistemas emocionales primarios en el cerebro de los mamíferos, como el miedo, la ira o el placer, basados en circuitos neuronales específicos.

En psicología, diversas teorías han intentado clasificar y describir las emociones. De entre las más influyentes, las teorías discretas, como las de Paul Ekman, identifican un número limitado de emociones básicas universales (alegría, tristeza, ira, sorpresa, miedo, asco). Por otro lado, teorías dimensionales, como el modelo de afecto circunflejo de James Russell (1980), proponen que las emociones se distribuyen en un espacio bidimensional definido por la valencia (placer-displacer) y la activación (alta-baja).

A pesar de estos avances en biología y psicología, la inteligencia artificial (IA) ha ignorado en gran medida las emociones, centrándose en replicar únicamente los procesos racionales de la mente humana. Desde la fundación de la IA por John McCarthy en 1955, la disciplina se ha focalizado en la imitación de la lógica y el razonamiento, desvinculándose de las emociones por la influencia de una tradición racionalista que, desde Descartes, las consideraba perturbadoras para la razón. Marvin Minsky fue una de las voces críticas frente a esta visión reduccionista, argumentando que las emociones no son un obstáculo, sino una forma de pensamiento que toda inteligencia avanzada debería poseer.

En años más recientes, el campo de la informática afectiva, liderado por Rosalind Picard (1997), ha tratado de integrar emociones en sistemas artificiales, pero los avances han sido limitados. Los modelos actuales de emociones en IA suelen ser soluciones específicas y no generalizables. Esto resalta la necesidad de un marco teórico más sólido que permita a los sistemas de IA aprender y generar emociones de manera autónoma, un vacío que este trabajo intenta abordar proponiendo un marco autoaprendido basado en principios de aprendizaje por refuerzo.

Objetivos

El objetivo principal de esta investigación es demostrar que las emociones artificiales pueden ser aprendidas de forma autónoma por agentes de inteligencia artificial a partir de principios fundamentales, sin necesidad de supervisión externa o modelos predefinidos. Para ello, se proponen tres hipótesis clave:

- H1: **Descriptibilidad matemática de las emociones.** Las emociones básicas, como la ira, satisfacción o miedo, corresponden a patrones temporales diferenciados percibidos en valores clave para un ser vivo, tales como recompensas recientes, recompensas futuras esperadas y estados anticipados del entorno.
- H2: **Codificación espontánea del espectro emocional.** Estos patrones pueden emerger espontáneamente y ser aprendidos de forma no supervisada por agentes de inteligencia artificial a partir de primeros principios, utilizando únicamente su estado actual, su experiencia reciente y sus expectativas temporales.
- H3: **Utilidad medible de las emociones sintéticas** La integración de estos patrones emocionales en la arquitectura de un agente puede mejorar su capacidad para tomar decisiones, su eficacia en el aprendizaje y su interacción social, de modo análogo a los beneficios que las emociones proporcionan a los seres vivos.

Las hipótesis H1 y H2 se investigan a fondo tanto teórica como experimentalmente en esta investigación, mientras que la hipótesis H3, cuya investigación completa excede el alcance de este trabajo, se aborda y analiza desde una perspectiva teórica.

Los objetivos generales del estudio son desarrollar un marco teórico y experimental que permita definir, aprender e integrar emociones en sistemas de aprendizaje por refuerzo (o *Reinforcement Learning*), y validar empíricamente su aplicabilidad mediante un caso de estudio. Se busca proporcionar una metodología para el aprendizaje empírico de emociones artificiales, utilizando algoritmos que permitan la codificación y la interpretación de respuestas emocionales en tiempo real.

Un marco emocional autoaprendido genérico

El marco emocional autoaprendido propuesto en este trabajo, basado en principios de Aprendizaje por Refuerzo (RL), permite que un agente de IA, interactuando en un entorno, aprenda y utilice un espectro emocional sintético, imitando emociones naturales documentadas. Este

marco concibe las emociones como patrones temporales derivados de valores críticos, como recompensas recientes y recompensas futuras esperadas, postulando que estos patrones son la base de las emociones tanto en sistemas naturales como artificiales.

Hipótesis y supuestos clave

Se establece que las emociones pueden entenderse como patrones temporales observados en variables cognitivas. La complejidad de las emociones depende de las capacidades cognitivas del agente, donde emociones más básicas reflejan tendencias en recompensas inmediatas, y emociones más sofisticadas integran valores anticipados y estados del mundo anticipados. Esto sugiere que los agentes de IA, mediante el aprendizaje de sus experiencias previas, pueden desarrollar emociones autónomamente.

Metodología y arquitectura del marco emocional

El marco incluye un *codificador emocional*, usualmente una red neuronal autocodificadora (*autoencoder*), que aprende y codifica patrones emocionales de secuencias temporales multivariadas observadas por el agente. Éste, basado en una arquitectura de RL extendida a partir de los métodos *actor-crítico*, emplea un estado “extendido” que combina su estado objetivo (el entorno) con un estado emocional subjetivo codificado, enriqueciendo así las entradas de su política de decisión (*policy*) con un componente emocional. Se presentan definiciones clave para los estados emocionales y el proceso de aprendizaje de emociones:

1. **Entrenamiento del codificador emocional:** Utilizando secuencias de recompensas y valores de estado, el codificador aprende a identificar patrones temporales en estos valores. Se exploran tanto enfoques *offline*, que congelan el modelo emocional una vez entrenado, como enfoques *online*, donde el aprendizaje del modelo emocional se actualiza junto con la política del agente.
2. **Elicitación e integración emocional:** El codificador emocional permite al agente enriquecer su representación del estado, creando estados extendidos que abarcan percepciones objetivas y apreciaciones subjetivas, proporcionando una respuesta emocional inmediata ante cambios en recompensas y valores de estado.
3. **Interpretación y mapeo emocional:** Los patrones emocionales aprendidos se agrupan en *clusters* en un espacio latente, cuyos centroides se asocian a perfiles emocionales referenciales. Estos grupos (o *clusters*) facilitan la interpretación humana y se mapean a términos emocionales conocidos (como “satisfacción” o “preocupación”), basándose en modelos probabilísticos como los modelos de mezclas gaussianas (*Gaussian mixture models*). Este proceso permite la atribución e interpretación de emociones en tiempo real, aportando transparencia a los estados emocionales del agente.

Órdenes emocionales y escalabilidad

El marco establece diferentes niveles de complejidad emocional, denominados **órdenes emocionales**:

- **Orden 0:** Agente estándar de RL, sin emociones.
- **Orden I:** Emociones inmediatas basadas en recompensas actuales, como “bueno” o “malo”.
- **Orden II:** Emociones retrospectivas basadas en tendencias de recompensas recientes, permitiendo emociones como “satisfacción” o “frustración”.
- **Orden III:** Emociones anticipatorias mediante el uso de predicciones de valores de estado, permitiendo emociones dinámicas como “temor”, “excitación” o “euforia”.
- **Orden IV:** Emociones basadas en el conocimiento del mundo, donde el agente evalúa discrepancias entre sus predicciones y estados observados, generando emociones de “sorpresa” o “curiosidad”.
- **Órdenes superiores:** Emociones de mayor complejidad, como la retrospección distante, las emociones sociales o morales, y la autoconciencia, pueden surgir en agentes con habilidades cognitivas avanzadas, tales como memoria a largo plazo o modelos internos complejos del mundo y de sí mismos (no explorados en el presente trabajo).

Aplicabilidad y resultados

La integración de emociones en la arquitectura de RL tiene el potencial de mejorar la toma de decisiones y la eficiencia del aprendizaje, similarmente a los beneficios observados en sistemas biológicos. Los resultados experimentales demuestran que el espectro emocional aprendido es coherente y comprensible para observadores humanos, quienes identificaron patrones emocionales que se alinean con las emociones humanas, validando así la efectividad y naturalidad del sistema en escenarios de prueba.

Este marco establece una base sólida para avanzar hacia una IA emocional que pueda no sólo procesar, sino también expresar estados emocionales complejos, integrando tanto respuestas emocionales instintivas como evaluativas en su interacción con el entorno y otros agentes.

Resultados

La aplicación del marco emocional se evaluó en el entorno *LunarLander-v2*, en el cual el agente de Aprendizaje por Refuerzo (RL) debía alunizar una nave espacial con control de posición, velocidad y consumo de combustible. Los resultados experimentales obtuvieron el aprendizaje de ocho emociones distintas, modeladas como patrones temporales de recompensas y valores de estado en secuencias de 20 pasos, permitiendo su interpretación y la elicitación en tiempo real.

Validación y evaluación experimental

1. **Codificación emocional y agrupamiento:** A través de un autocodificador y un modelo de mezcla gaussiana, el agente categorizó sus emociones en ocho grupos representando emociones básicas como “temor”, “excitación” u “optimismo”. La disposición espacial

de estas emociones se verificó mediante análisis t-SNE, revelando una organización coherente con los ejes de placer y activación, dimensiones reconocidas en psicología emocional.

2. **Transiciones emocionales naturales:** El agente demostró transiciones naturales y progresivas en su espectro emocional. En los episodios de prueba, inició en un estado neutral o de emoción leve, y transitó a emociones de preocupación o temor ante dificultades, o a emociones positivas como la satisfacción tras un alunizaje exitoso. Estas transiciones se visualizaron mediante una matriz y un grafo de transición que destacaban las secuencias emocionales más comunes durante el testeo del agente.
3. **Estudio de atribución emocional con humanos:** Para validar la reconocibilidad de las emociones aprendidas, 96 participantes evaluaron secuencias de video del agente alunizando. Utilizando el método *Self-Assessment Manikin* (SAM), los participantes puntuaron las secuencias en las dimensiones de placer, activación y dominancia (PAD). Los resultados reflejaron alta consistencia en la evaluación de emociones, con los valores de PAD alineándose estrechamente con emociones humanas documentadas, y se observó una notable diferenciación en el espectro emocional de cada secuencia.
4. **Maapeo frente a registros experimentales documentados:** Los valores de placer, activación y dominancia (PAD) de las emociones aprendidas se compararon con registros experimentales de estudios psicológicos, mostrando una alineación significativa. Cada emoción aprendida mostró correspondencias estrechas con emociones humanas documentadas en la literatura, lo que valida la coherencia del modelo emocional respecto a estudios previos. Esta correspondencia respalda la precisión del marco en la reproducción de patrones emocionales reconocibles y naturales.

Conclusiones y observaciones clave

El agente desarrolló espontáneamente un sistema emocional con patrones y transiciones comparables a emociones naturales, exhibiendo homeostasis (equilibrio emocional) y subjetividad en respuesta a sus interacciones. Aunque las emociones de orden superior como la habituación y extinción no fueron observadas, el modelo demostró ser adaptable y coherente con la teoría emocional tanto discreta como dimensional, sentando las bases para una IA con percepción emocional compleja.

Métodos y herramientas

1. **Entrenamiento del agente de aprendizaje por refuerzo (RL):** Se preentrenó un agente clásico de RL con el método actor-crítico, específicamente el modelo PPO (*Proximal Policy Optimization*), para manejar de manera competente la tarea en el entorno *LunarLander-v2*.
2. **Generación de datos de entrenamiento:** Con el agente preentrenado, se ejecutaron escenarios adicionales para capturar series temporales multivariadas, enfocándose en secuencias de recompensa y valor del estado. Estas series se normalizaron y secuenciaron

con una longitud de “ventana emocional” de 20 pasos.

3. **Modelo de codificación emocional:** Un autocodificador convolucional de una dimensión se entrenó para capturar las representaciones latentes de los datos emocionales en un espacio de baja dimensionalidad, logrando un balance entre precisión y compresión de los datos.
4. **Agrupamiento e interpretación emocional:** Los patrones emocionales latentes fueron agrupados mediante un modelo de mezcla gaussiana, obteniendo ocho categorías emocionales, cada una correspondiente a una emoción prototípica experimentada por el agente.

Este enfoque metodológico permite la reproducción y el análisis de emociones aprendidas de manera autónoma, proporcionando las bases para la replicación y el uso en distintos entornos de RL.

Conclusiones

Principales logros

Este trabajo introduce un marco emocional autoaprendido que permite a los agentes de Aprendizaje por Refuerzo (RL) experimentar emociones de manera autónoma. A través del uso de patrones matemáticos derivados de valores críticos del entorno, como recompensas y expectativas de estado, el marco facilita la emergencia de emociones sin supervisión externa. Los agentes pueden identificar y responder a categorías emocionales básicas, organizadas en un espectro de emociones sistemático, coherente y reconocible.

El potencial de este modelo radica en su capacidad para integrar emociones como estados extendidos en la arquitectura de RL, y mejorar la capacidad del agente para evaluar y adaptarse a situaciones complejas. Además, la estructura del modelo permite el mapeo de emociones en términos humanos (por ejemplo, “temor”, “excitación”, “satisfacción”) a través de agrupamiento mediante técnicas de mezcla gaussiana. La validación mediante observadores humanos demostró que estas emociones son comprensibles y alineadas con el comportamiento emocional esperado, lo cual subraya la aplicabilidad del marco en la construcción de agentes más interpretables y adaptativos.

Este marco establece una base sólida para avanzar hacia agentes de IA con una riqueza emocional compleja, abriendo nuevas posibilidades para aplicaciones en robótica, interfaces humano-computadora y sistemas de toma de decisiones, donde las respuestas emocionales aporten beneficios adicionales. La capacidad de los agentes para desarrollar estados emocionales situacionales y transitorios sin programación explícita representa un avance significativo hacia la emulación de la inteligencia emocional en IA.

Limitaciones y futuras investigaciones

Aunque el marco se probó con éxito en emociones básicas de corto plazo, la extensión a emociones de orden superior y dinámicas de largo plazo, como la habituación y extinción,

requerirá mejoras. También se propone explorar entornos parcialmente observables y escenarios de aprendizaje continuo, lo cual enriquecerá la comprensión de emociones complejas y sus aplicaciones prácticas en IA.

Consideraciones éticas

El marco de emociones sintéticas es un sistema matemático que no implica experiencias de sufrimiento o disfrute reales para los agentes de IA.

References

- [1] Charles Darwin. *The Expression of the Emotions in Man and Animals*. London: John Murray, 1872.
- [2] Marc Bekoff. “Animal Emotions: Exploring Passionate Natures: Current interdisciplinary research provides compelling evidence that many animals experience such emotions as joy, fear, love, despair, and grief—we are not alone”. In: *BioScience* 50.10 (Oct. 2000), pp. 861–870.
- [3] Antonio Damasio. “Descartes’ error and the future of human life.” In: *Scientific American* 271.4 (1994), pp. 144–144.
- [4] Antonio Damasio. *The Feeling of What Happens: Body and Emotion in the Making of Consciousness*. New York: Harcourt Brace, 1999.
- [5] Jaak Panksepp and Lucy Biven. *The Archaeology of Mind: Neuroevolutionary Origins of Human Emotions*. New York: W. W. Norton & Company, 2012, pp. xxvii, 562.
- [6] Paul Ekman and Wallace V Friesen. “The repertoire of nonverbal behavior: Categories, origins, usage, and coding”. In: *Semiotica* 1.1 (1969), pp. 49–98.
- [7] James A. Russell. “A circumplex model of affect”. In: *Journal of Personality and Social Psychology* 39 (1980), pp. 1161–1178. DOI: [10.1037/h0077714](https://doi.org/10.1037/h0077714).
- [8] Richard S Lazarus. “Progress on a cognitive-motivational-relational theory of emotion.” In: *American psychologist* 46.8 (1991), pp. 819–834.
- [9] John Tooby and Leda Cosmides. “The past explains the present: Emotional adaptations and the structure of ancestral environments”. In: *Ethology and sociobiology* 11.4-5 (1990), pp. 375–424.
- [10] Charles Lindholm. “An Anthropology of Emotion”. In: *A Companion to Psychological Anthropology: Modernity and Psychocultural Change*. Ed. by Conerly Casey and Robert B. Edgerton. Wiley, 2005. Chap. 2, pp. 30–47. DOI: [10.1002/9780470996409.ch3](https://doi.org/10.1002/9780470996409.ch3).
- [11] René Descartes. *The Passions of the Soul*. Translated in various editions, commonly included in *The Philosophical Writings of Descartes*, Volume 1, Cambridge University Press, 1985. Amsterdam and Paris: Henri Le Gras, 1649.
- [12] David Hume. *A Treatise of Human Nature: Being an Attempt to Introduce the Experimental Method of Reasoning into Moral Subjects*. London, UK: John Noon, 1739–1740.
- [13] Immanuel Kant. *Observations on the Feeling of the Beautiful and the Sublime*. Translated by J. T. Goldthwait, Original work published 1764. Berkeley: University of California Press, 1960.
- [14] William James. *The Principles of Psychology*. New York: Henry Holt and Company, 1890.

- [15] Fabio Massimo Zennaro. “Theories of Emotion”. In: *Chapters in PhD dissertation*. 2013. Chap. 1.
- [16] Dacher Keltner and Jennifer S Lerner. “Emotion”. In: *The Handbook of Social Psychology*. Ed. by Daniel T Gilbert, Susan T Fiske, and Gardner Lindzey. New York: John Wiley & Sons, Ltd, 2010, pp. 317–352.
- [17] Carroll E Izard. *Human emotions*. New York: Springer, 1977.
- [18] Robert Plutchik. “A general psychoevolutionary theory of emotion”. In: *Theories of Emotion*. Ed. by Robert Plutchik and Henry Kellerman. Vol. 1. New York: Academic Press, 1980, pp. 3–33.
- [19] Lisa Feldman Barrett. *How Emotions Are Made: The Secret Life of the Brain*. Boston New York: Houghton Mifflin Harcourt, 2017.
- [20] William James. “What is an Emotion?” In: *Mind* 9.34 (1884), pp. 188–205.
- [21] Klaus R. Scherer. “What are emotions? And how can they be measured?” In: *Social Science Information* 44.4 (2005), pp. 695–729. DOI: [10.1177/0539018405058216](https://doi.org/10.1177/0539018405058216).
- [22] Ross Buck. “Mood and Emotion: A Comparison of Five Contemporary Views”. In: *Psychological Inquiry* 1.4 (1990), pp. 330–336.
- [23] Craig DeLancey. *Passionate Engines: What Emotions Reveal about Mind and Artificial Intelligence*. Oxford: Oxford University Press, 2002, p. 272.
- [24] Alan Fridlund. *Hard Feelings: Science’s Struggle to Define Emotions*. *The Atlantic*, email to Julie Beck. 2015.
- [25] Paul Ekman and Wallace V. Friesen. “Constants across cultures in the face and emotion”. In: *Journal of Personality and Social Psychology* 17.2 (1971), pp. 124–129. DOI: [10.1037/h0030377](https://doi.org/10.1037/h0030377).
- [26] Silvan S. Tomkins. *Affect Imagery Consciousness: Volume I, The Positive Affects*. New York: Springer Publishing Company, 1962, p. 536.
- [27] Silvan S. Tomkins. *Affect Imagery Consciousness: Volume II, The Negative Affects*. New York: Springer Publishing Company, 1963, p. 580.
- [28] Paul Ekman. “Basic Emotions”. In: *Handbook of Cognition and Emotion*. Ed. by Tim Dalgleish and Mick J. Power. Sussex, UK: John Wiley & Sons, 1999, pp. 45–60.
- [29] Paul Ekman. *Emotions Revealed: Recognizing Faces and Feelings to Improve Communication and Emotional Life*. New York, NY: Times Books, 2003.
- [30] Wilhelm Max Wundt. *Outlines of Psychology*. Leipzig: Wilhelm Engelmann, 1897.
- [31] J. A. Russell and A. Mehrabian. “Evidence for a Three-Factor Theory of Emotions”. In: *J. Res. Personal.* 11 (1977), pp. 273–294. DOI: [10.1016/0092-6566\(77\)90037-X](https://doi.org/10.1016/0092-6566(77)90037-X).
- [32] Margaret M. Bradley and Peter J. Lang. “Measuring emotion: The self-assessment manikin and the semantic differential”. In: *Journal of Behavior Therapy and Experimental Psychiatry* 25.1 (1994), pp. 49–59. DOI: [10.1016/0005-7916\(94\)90063-9](https://doi.org/10.1016/0005-7916(94)90063-9).
- [33] Peter J. Lang, Margaret M. Bradley, and Bruce N. Cuthbert. *International Affective Picture System (IAPS): Affective ratings of pictures and instruction manual*. Technical Report A-8. Gainesville: University of Florida, Center for Research in Psychophysiology, 2008.
- [34] Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. “Norms of valence, arousal, and dominance for 13,915 English lemmas”. In: *Behavior Research Methods* 45.4 (2013), pp. 1191–1207. DOI: [10.3758/s13428-012-0314-x](https://doi.org/10.3758/s13428-012-0314-x).

-
- [35] Elliot Aronson, Timothy D. Wilson, and Robin M. Akert. *Social Psychology*. 7th. Upper Saddle River, NJ: Pearson Education, Inc., 2005.
 - [36] Pedro Sequeira, Francisco Melo, and Ana Paiva. “Emergence of emotional appraisal signals in reinforcement learning agents”. In: *Autonomous Agents and Multi-Agent Systems* 29 (July 2014). DOI: [10.1007/s10458-014-9262-4](https://doi.org/10.1007/s10458-014-9262-4).
 - [37] Ira J. Roseman, Ann Aliki Antoniou, and Paul E. Jose. “Appraisal Determinants of Emotions: Constructing a More Accurate and Comprehensive Theory”. In: *Cognition and Emotion* 10.3 (1996), pp. 241–277. DOI: [10.1080/026999396380240](https://doi.org/10.1080/026999396380240).
 - [38] Bernard Weiner. *An Attributional Theory of Motivation and Emotion*. Springer Series in Social Psychology. New York: Springer-Verlag, 1986. DOI: [10.1007/978-1-4612-4948-1](https://doi.org/10.1007/978-1-4612-4948-1).
 - [39] Richard S. Lazarus. *Emotion and Adaptation*. New York: Oxford University Press, 1991, p. 557.
 - [40] David Matsumoto and Hyisung C. Hwang. “Culture and Emotion: Integrating Biological Universality with Cultural Specificity”. In: *The Handbook of Culture and Psychology*. Ed. by David Matsumoto and Hyisung C. Hwang. New York: Oxford University Press, 2019, pp. 567–597. DOI: [10.1093/oso/9780190679743.003.0012](https://doi.org/10.1093/oso/9780190679743.003.0012).
 - [41] Peter Salovey and John D. Mayer. “Emotional intelligence”. In: *Imagination, Cognition, and Personality* 9.3 (1990), pp. 185–211. DOI: [10.2190/DUGG-P24E-52WK-6CDG](https://doi.org/10.2190/DUGG-P24E-52WK-6CDG).
 - [42] Christian Montag and Jaak Panksepp. “Primal emotional-affective expressive foundations of human facial expression”. In: *Motivation and Emotion* 40.5 (2016), pp. 760–766. DOI: [10.1007/s11031-016-9570-x](https://doi.org/10.1007/s11031-016-9570-x).
 - [43] Joseph E. LeDoux. *The Emotional Brain: The Mysterious Underpinnings of Emotional Life*. New York: Simon & Schuster, 1996, p. 384.
 - [44] J. Panksepp. *Affective Neuroscience: The Foundation of Human and Animal Emotions*. New York: Oxford University Press, 1998, p. 480.
 - [45] Trevor W. Robbins and Barry J. Everitt. “Neurobehavioural mechanisms of reward and motivation”. In: *Current Opinion in Neurobiology* 6.2 (1996), pp. 228–236. DOI: [10.1016/S0959-4388\(96\)80077-8](https://doi.org/10.1016/S0959-4388(96)80077-8).
 - [46] Roshan Cools, Angela C. Roberts, and Trevor W. Robbins. “Serotonergic regulation of emotional and behavioural control processes”. In: *Trends in Cognitive Sciences* 12.1 (2008), pp. 31–40. DOI: [10.1016/j.tics.2007.10.011](https://doi.org/10.1016/j.tics.2007.10.011).
 - [47] Hugo Lövhelm. “A new three-dimensional model for emotions and monoamine neurotransmitters”. In: *Medical Hypotheses* 78.2 (2012), pp. 341–348. DOI: [10.1016/j.mehy.2011.11.016](https://doi.org/10.1016/j.mehy.2011.11.016).
 - [48] Alberto Prieto et al. “Neural networks: An overview of early research, current frameworks and new challenges”. In: *Neurocomputing* 214 (2016), pp. 242–268. DOI: [10.1016/j.neucom.2016.06.014](https://doi.org/10.1016/j.neucom.2016.06.014).
 - [49] Warren S McCulloch and Walter Pitts. “A logical calculus of the ideas immanent in nervous activity”. In: *The bulletin of mathematical biophysics* 5 (1943), pp. 115–133.
 - [50] Frank Rosenblatt. *The Perceptron—a perceiving and recognizing automaton*. Project PARA Technical Report 85-460-1. Ithaca, NY: Cornell Aeronautical Laboratory, 1957.
 - [51] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. “Learning representations by back-propagating errors”. In: *Nature* 323.6088 (1986), pp. 533–536. DOI: [10.1038/323533a0](https://doi.org/10.1038/323533a0).

- [52] John J Hopfield. “Neural networks and physical systems with emergent collective computational abilities.” In: *Proceedings of the national academy of sciences* 79.8 (1982), pp. 2554–2558.
- [53] Yann LeCun et al. “Gradient-based learning applied to document recognition”. In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324.
- [54] Stephen Grossberg. “Adaptive pattern classification and universal recoding: I. Parallel development and coding of neural feature detectors”. In: *Biological Cybernetics* 23.3 (1976), pp. 121–134. DOI: [10.1007/BF00344744](https://doi.org/10.1007/BF00344744).
- [55] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. *Neural Machine Translation by Jointly Learning to Align and Translate*. Preprint at <https://arxiv.org/abs/1409.0473>. 2014. DOI: [10.48550/arXiv.1409.0473](https://doi.org/10.48550/arXiv.1409.0473).
- [56] Ashish Vaswani et al. “Attention is all you need”. In: *Advances in Neural Information Processing Systems* 30 (2017), pp. 5998–6008.
- [57] Edmund Ronald and Marc Schoenauer. “Genetic lander: An experiment in accurate neuro-genetic control”. In: *Parallel Problem Solving from Nature — PPSN III*. Ed. by Yuval Davidor, Hans-Paul Schwefel, and Reinhard Männer. Berlin, Heidelberg: Springer Berlin Heidelberg, 1994, pp. 452–461.
- [58] Kenneth O Stanley and Risto Miikkulainen. “Evolving neural networks through augmenting topologies”. In: *Evolutionary computation* 10.2 (2002), pp. 99–127.
- [59] Michael Wooldridge. *An Introduction to MultiAgent Systems*. Chichester, England: John Wiley & Sons, 2002, p. 348.
- [60] G. Beni and J. Wang. “Swarm Intelligence in Cellular Robotic Systems”. In: *Robots and Biological Systems: Towards a New Bionics?* Ed. by P. Dario, G. Sandini, and P. Aebischer. Vol. 102. NATO ASI Series. Berlin, Heidelberg: Springer, 1993. DOI: [10.1007/978-3-642-58069-7_38](https://doi.org/10.1007/978-3-642-58069-7_38).
- [61] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction - Second edition*. Cambridge, MA, USA: A Bradford Book, 2018.
- [62] K. Gunderson. *Mentality and Machines*. Anchor books. Garden City, New York: Doubleday, 1971.
- [63] Allen Newell. “Intellectual issues in the history of artificial intelligence”. In: *Artificial Intelligence: Critical Concepts* (1982), pp. 25–70.
- [64] Stuart J. Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. 4th ed. Hoboken: Pearson, 2021.
- [65] Suman Ojha, Jonathan Vitale, and Mary-Anne Williams. “Computational emotion models: a thematic review”. In: *International Journal of Social Robotics* 13 (2021), pp. 1253–1279.
- [66] John McCarthy et al. “A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence”. In: *AI Magazine* 27.4 (1955), p. 12. DOI: [10.1609/aimag.v27i4.1904](https://doi.org/10.1609/aimag.v27i4.1904).
- [67] Martha Nussbaum. “Compassion: The basic social emotion”. In: *Social Philosophy and Policy* 13 (1996), pp. 27–58.
- [68] Keith Oatley. *Emotions: A brief history*. Malden, MA: Blackwell, 2004.
- [69] Martha Nussbaum. *Upheavals of Thought: The Intelligence of Emotions*. New York: Cambridge University Press, 2001.

-
- [70] Cheshire Calhoun and Robert Solomon. “What is an Emotion?” In: *Readings in Philosophical Psychology*. Ed. by Cheshire Calhoun and Robert Solomon. New York: Oxford University Press, 1984.
 - [71] René Descartes. *Discourse on the Method*. Leiden, Netherlands: Ian Maire, 1637.
 - [72] Friedrich Nietzsche. *Jenseits von Gut und Böse: Vorspiel einer Philosophie der Zukunft* [*Beyond Good and Evil: Prelude to a Philosophy of the Future*]. Leipzig: C. G. Naumann, 1886.
 - [73] Arthur Schopenhauer. *Die Welt als Wille und Vorstellung* [*The World as Will and Representation*]. Leipzig: Brockhaus, 1819.
 - [74] Donald M. MacKay. “Mindlike Behaviour in Artefacts”. In: *British Journal for the Philosophy of Science* 2.6 (1951), pp. 105–121. DOI: [10.1093/bjps/ii.6.105](https://doi.org/10.1093/bjps/ii.6.105).
 - [75] Michael Scriven. “The Mechanical Concept of Mind”. In: *Mind* 62.246 (1953), pp. 230–240. DOI: [10.1093/mind/LXII.246.230](https://doi.org/10.1093/mind/LXII.246.230).
 - [76] Paul Ziff. “The Feelings of Robots”. In: *Analysis* 19.3 (1959), pp. 64–68. DOI: [10.1093/analysis/19.3.64](https://doi.org/10.1093/analysis/19.3.64).
 - [77] John McCarthy. “The Little Thoughts of Thinking Machines”. In: *Psychology Today* 17.12 (1983), pp. 46–49.
 - [78] Marvin Lee Minsky. *The Society of Mind*. New York: Simon and Schuster, 1986, p. 339.
 - [79] Robert P. Abelson. “Computer simulation of “hot” cognition”. In: *Computer Simulation of Personality*. Ed. by S. S. Tomkins and S. Messick. New York: Wiley, 1963, pp. 277–298.
 - [80] Aaron Sloman and Monica Croucher. “Why Robots Will Have Emotions”. In: *Proceedings of the Seventh International Joint Conference on Artificial Intelligence (IJCAI)*. Also available as Cognitive Science Research Paper 176, University of Sussex. IJCAI, 1981, pp. 197–202.
 - [81] Marvin Lee Minsky. *The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind*. New York: Simon & Schuster, 2006, p. 387.
 - [82] John Fulcher, ed. *Computational Intelligence: A Compendium*. Vol. 115. Studies in Computational Intelligence. Berlin, Heidelberg: Springer, 2008. DOI: [10.1007/978-3-540-78293-3](https://doi.org/10.1007/978-3-540-78293-3).
 - [83] Nathan Gardels. “Human Intelligence Can’t Be Transferred to Machines”. In: *The WorldPost* (Mar. 2018). Published by The Washington Post and Berggruen Institute.
 - [84] Melanie Mitchell. *Why AI is Harder Than We Think*. Preprint at <https://arxiv.org/abs/2104.12871>. 2021.
 - [85] Joseph Bates. “The Role of Emotion in Believable Agents”. In: *Communications of the ACM* 37.7 (1994), pp. 122–125. DOI: [10.1145/176789.176803](https://doi.org/10.1145/176789.176803).
 - [86] Marco Paleari and Christine Lisetti. *Psychologically grounded avatars expressions*. First Workshop on Emotion and Computing at KI 2006, 29th Annual German Conference on AI, 2006. Citeseer, 2006.
 - [87] R Picard. “Affective Computing”. In: *MIT Media Laboratory Perceptual Computing Section Technical Report No. 321* (1995).
 - [88] Jianhua Tao and Tieniu Tan. “Affective computing: A review”. In: *International Conference on Affective computing and intelligent interaction* (2005), pp. 981–995.

- [89] Smith K. Khare et al. “Emotion recognition and artificial intelligence: A systematic review (2014–2023) and research recommendations”. In: *Information Fusion* 102 (2024), p. 102019. DOI: [10.1016/j.inffus.2023.102019](https://doi.org/10.1016/j.inffus.2023.102019).
- [90] Qinglin She et al. “Cross-subject EEG emotion recognition using multi-source domain manifold feature selection”. In: *Computers in Biology and Medicine* 159 (2023), p. 106860. DOI: [10.1016/j.combiomed.2023.106860](https://doi.org/10.1016/j.combiomed.2023.106860).
- [91] Shakir Yasin et al. “Machine learning based approaches for clinical and non-clinical depression recognition and depression relapse prediction using audiovisual and EEG modalities: A comprehensive review”. In: *Computers in Biology and Medicine* 159 (2023), p. 106741. DOI: [10.1016/j.combiomed.2023.106741](https://doi.org/10.1016/j.combiomed.2023.106741).
- [92] N.-D. Mai, B.-G. Lee, and W.-Y. Chung. “Affective Computing on Machine Learning-Based Emotion Recognition Using a Self-Made EEG Device”. In: *Sensors* 21.15 (2021), p. 5135. DOI: [10.3390/s21155135](https://doi.org/10.3390/s21155135).
- [93] Yasin Özlük and Esra Akman Aydin. “Fuzzy Logic Control of a Head-movement Based Semi-autonomous Human-machine Interface”. In: *Journal of Bionic Engineering* 20.2 (2023), pp. 645–655. DOI: [10.1007/s42235-022-00272-3](https://doi.org/10.1007/s42235-022-00272-3).
- [94] Stevo Bozinovski and Liljana Bozinovska. “Self-learning agents: A connectionist theory of emotion based on crossbar value judgment”. In: *Cybernetics and Systems* 32.6 (2001), pp. 637–669. DOI: [10.1080/01969720118145](https://doi.org/10.1080/01969720118145).
- [95] Cynthia L. Breazeal. “Sociable Machines: Expressive Social Exchange Between Humans and Robots”. Doctoral Thesis. PhD thesis. Massachusetts Institute of Technology, 2000.
- [96] Amit Kumar Pandey and Rodolphe Gelin. “A Mass-Produced Sociable Humanoid Robot: Pepper: The First Machine of Its Kind”. In: *IEEE Robotics & Automation Magazine* 25.3 (2018), pp. 40–48. DOI: [10.1109/MRA.2018.2833157](https://doi.org/10.1109/MRA.2018.2833157).
- [97] Sigrid Schmitz. “Sophia: Potentials and Challenges of a Modern Cyborg”. In: *Humanity In-Between and Beyond*. Ed. by Monika Michałowska. Springer Verlag, 2023, pp. 153–178. DOI: [10.1007/978-3-031-27945-4_9](https://doi.org/10.1007/978-3-031-27945-4_9).
- [98] *World Robot Conference 2024*. <https://www.worldrobotconference.com/en/>. Accessed: 2024-10-08. People’s Government of Beijing Municipality, Ministry of Industry, Information Technology of China, China Association for Science, and Technology, 2024.
- [99] Sara A. Taylor et al. “Personalized Multitask Learning for Predicting Tomorrow’s Mood, Stress, and Health”. In: *IEEE Transactions on Affective Computing* 11.2 (2017), pp. 200–213. DOI: [10.1109/TAFFC.2017.2717321](https://doi.org/10.1109/TAFFC.2017.2717321).
- [100] Supreeth Prajwal Shashikumar et al. “A Deep Learning Approach to Monitoring and Detecting Atrial Fibrillation Using Wearable Technology”. In: *IEEE International Conference on Biomedical and Health Informatics (BHI)*. 2017, pp. 141–144. DOI: [10.1109/BHI.2017.7897225](https://doi.org/10.1109/BHI.2017.7897225).
- [101] Jennifer A. Healey and Rosalind W. Picard. “Detecting Stress During Real-World Driving Tasks Using Physiological Sensors”. In: *IEEE Transactions on Intelligent Transportation Systems* 6.2 (2005), pp. 156–166. DOI: [10.1109/TITS.2005.848368](https://doi.org/10.1109/TITS.2005.848368).
- [102] Juan Abdon Miranda-Correa et al. “AMIGOS: A Dataset for Affect, Personality and Mood Research on Individuals and Groups”. In: *IEEE Transactions on Affective Computing* 12.2 (2021), pp. 479–493. DOI: [10.1109/TAFFC.2018.2884461](https://doi.org/10.1109/TAFFC.2018.2884461).

-
- [103] Ruth Aylett et al. “Affective Agents for Education Against Bullying”. In: *Affective Information Processing*. Ed. by Jianhua Tao and Tieniu Tan. Beijing: Springer, 2009, pp. 75–90. DOI: [10.1007/978-1-84800-306-4_5](https://doi.org/10.1007/978-1-84800-306-4_5).
 - [104] Qiaochu Wang. *AI Tunes into Emotions: The Rise of Affective Computing*. Neuroscience News. 2024. URL: <https://neurosciencenews.com/affective-computing-ai-emotion-25668/>.
 - [105] Guanxiong Pei et al. “Affective Computing: Recent Advances, Challenges, and Future Trends”. In: *Intelligent Computing 3* (2024), Article 0076. DOI: [10.34133/icomputing.0076](https://doi.org/10.34133/icomputing.0076).
 - [106] Joost Broekens. *A Temporal Difference Reinforcement Learning Theory of Emotion: Unifying Emotion, Cognition and Adaptive Behavior*. Preprint at <https://doi.org/10.48550/arXiv.1807.08941>. 2018.
 - [107] Andrew Ortony, Gerald L. Clore, and Allan Collins. *The Cognitive Structure of Emotions*. Cambridge: Cambridge University Press, 1988.
 - [108] Rainer Reisenzein. “Emotional experience in the computational belief–desire theory of emotion”. In: *Emotion Review* 1.3 (2009), pp. 214–222.
 - [109] Thomas M Moerland, Joost Broekens, and Catholijn M Jonker. “Emotion in reinforcement learning agents and robots: a survey”. In: *Machine Learning* 107 (2018), pp. 443–480.
 - [110] M Joffily and G Coricelli. “Emotional Valence and the Free-Energy Principle”. In: *PLoS Comput Biol* 9.6 (2013), e1003094. DOI: [10.1371/journal.pcbi.1003094](https://doi.org/10.1371/journal.pcbi.1003094).
 - [111] Pedro Sequeira, Francisco S Melo, and Ana Paiva. “Learning by appraising: an emotion-based approach to intrinsic reward design”. In: *Adaptive Behavior* 22.5 (2014), pp. 330–349.
 - [112] Henry Williams et al. “Emotion inspired adaptive robotic path planning”. In: *2015 IEEE congress on evolutionary computation (CEC)* (2015), pp. 3004–3011.
 - [113] Clark D. Elliott. “The Affective Reasoner: A Process Model of Emotions in a Multi-Agent System”. Technical Report No. 32. Evanston, Illinois: Northwestern University, The Institute for the Learning Sciences, 1992.
 - [114] Tomoko Koda. “Agents with Faces: A Study on the Effect of Personification of Software Agents”. M.S. Thesis. Cambridge, MA: Massachusetts Institute of Technology, Media Laboratory, 1996.
 - [115] Christoph Bartneck. “eMuu - An Embodied Emotional Character for the Ambient Intelligent Home”. PhD thesis. Eindhoven, Netherlands: Eindhoven University of Technology, 2002.
 - [116] Andriy Bondarev. “Design of an Emotion Management System for a Home Robot”. Pd Eng Thesis. Eindhoven, Netherlands: Eindhoven University of Technology, 2002.
 - [117] Christoph Bartneck, Michael J Lyons, and Martin Saerbeck. “The relationship between emotion models and artificial intelligence”. In: *arXiv preprint arXiv:1706.09554* (2017).
 - [118] Stevo Bozinovski. “A self-learning system using secondary reinforcement”. In: *Cybernetics and Systems Research*. Ed. by Robert Trappl. Amsterdam: North-Holland, 1982, pp. 397–402.
 - [119] Atsushi Matsuda, Hideaki Misawa, and Keiichi Horio. “Decision Making Based on Reinforcement Learning and Emotion Learning for Social Behavior”. In: *2011 IEEE*

- International Conference on Fuzzy Systems (FUZZ)*. IEEE, 2011, pp. 2714–2719. DOI: [10.1109/FUZZY.2011.6007694](https://doi.org/10.1109/FUZZY.2011.6007694).
- [120] Elmer Jacobs, Joost Broekens, and Catholijn M. Jonker. “Emergent Dynamics of Joy, Distress, Hope and Fear in Reinforcement Learning Agents”. In: *Adaptive Learning Agents Workshop at AAMAS 2014*. 2014.
 - [121] Miguel A. Salichs and María Malfaz. “A New Approach to Modeling Emotions and Their Use on a Decision-Making System for Artificial Agents”. In: *IEEE Transactions on Affective Computing* 3.1 (2012), pp. 56–68. DOI: [10.1109/T-AFFC.2011.33](https://doi.org/10.1109/T-AFFC.2011.33).
 - [122] Joost Broekens, Walter A. Kusters, and Fons J. Verbeek. “Affect, Anticipation, and Adaptation: Affect-Controlled Selection of Anticipatory Simulation in Artificial Adaptive Agents”. In: *Adaptive Behavior* 15.4 (2007), pp. 397–422. DOI: [10.1177/1059712307084686](https://doi.org/10.1177/1059712307084686).
 - [123] Nicolas Schweighofer and Kenji Doya. “Meta-learning in Reinforcement Learning”. In: *Neural Networks* 16.1 (2003), pp. 5–9. DOI: [10.1016/S0893-6080\(02\)00232-1](https://doi.org/10.1016/S0893-6080(02)00232-1).
 - [124] Eric Hogewoning et al. “Strategies for Affect-Controlled Action-Selection in Soar-RL”. In: *Nature Inspired Problem-Solving Methods in Knowledge Engineering*. Ed. by José Mira and José R. Álvarez. Vol. 4528. Lecture Notes in Computer Science. Springer, 2007, pp. 501–510. DOI: [10.1007/978-3-540-73055-2_52](https://doi.org/10.1007/978-3-540-73055-2_52).
 - [125] Xuefei Shi, Zhiliang Wang, and Qiong Zhang. “Artificial Emotion Model Based on Neuromodulators and Q-learning”. In: *Future Control and Automation: Proceedings of the 2nd International Conference on Future Control and Automation (ICFCA 2012)*. Ed. by Wei Deng. Vol. 172. Berlin, Heidelberg: Springer, 2012, pp. 293–299. DOI: [10.1007/978-3-642-31006-5_35](https://doi.org/10.1007/978-3-642-31006-5_35).
 - [126] Arnaud J. Blanchard and Lola Cañamero. “From Imprinting to Adaptation: Building a History of Affective Interaction”. In: *Proceedings of the 5th International Workshop on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems*. Vol. 123. Lund University Cognitive Studies. Lund, Sweden: Lund University, 2005, pp. 23–30.
 - [127] Hyungil Ahn and Rosalind W. Picard. “Affective Cognitive Learning and Decision Making: The Role of Emotions”. In: *Proceedings of the 18th European Meeting on Cybernetics and Systems Research (EMCSR 2006)*. Vienna: Austrian Society for Cybernetic Studies, 2006, pp. 1–6.
 - [128] M. Lahnstein. “The Emotive Episode is a Composition of Anticipatory and Reactive Evaluations”. In: *Proceedings of the AISB’05 Symposium on Agents that Want and Like: Motivational and Emotional Roots of Cognition and Action*. Hatfield, UK: SSAISB, 2005, pp. 62–69.
 - [129] Elmer Jacobs, Joost Broekens, and Catholijn M. Jonker. “Joy, Distress, Hope, and Fear in Reinforcement Learning”. In: *Proceedings of the 13th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2014)*. Vol. 2. Paris, France: International Foundation for Autonomous Agents and Multiagent Systems, 2014, pp. 1615–1616.
 - [130] Thomas M. Moerland, Joost Broekens, and Catholijn M. Jonker. “Fear and Hope Emerge from Anticipation in Model-Based Reinforcement Learning”. In: *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI)*. New York, USA: AAAI Press, 2016, pp. 848–854.

- [131] Joost Broekens and Laduona Dai. “A TDRL Model for the Emotion of Regret”. In: *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. Cambridge, UK: IEEE, 2019, pp. 150–156. DOI: [10.1109/ACII.2019.8925441](https://doi.org/10.1109/ACII.2019.8925441).
- [132] Laduona Dai. “Human perception of an adaptive agent’s fear simulated based on TDRL Theory of emotions”. Master’s thesis. University of Twente, 2019.
- [133] Floortje Lycklama ‘a Nijeholt and Joost Broekens. “The Role of Simulated Emotions in Reinforcement Learning: Insights from a Human-Robot Interaction Experiment”. In: *2023 IEEE International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE. 2023, pp. 1–8. DOI: [10.1109/ACII.2023.10388167](https://doi.org/10.1109/ACII.2023.10388167).
- [134] Hideyoshi Yanagisawa. “Free-Energy Model of Emotion Potential: Modeling Arousal Potential as Information Content Induced by Complexity and Novelty”. In: *Frontiers in Computational Neuroscience* 15 (2021), Article 698252. DOI: [10.3389/fncom.2021.698252](https://doi.org/10.3389/fncom.2021.698252).
- [135] Candice Pattisapu et al. *Free Energy in a Circumplex Model of Emotion*. 2024. arXiv: [2407.02474](https://arxiv.org/abs/2407.02474) [cs.AI].
- [136] N. Chentanez, A. Barto, and S. Singh. “Intrinsically motivated reinforcement learning”. In: *Advances in Neural Information Processing Systems*. Vol. 17. 2004, pp. 1281–1288.
- [137] Edmund T. Rolls. “What Are Emotions, Why Do We Have Emotions, and What Is Their Computational Basis in the Brain?” In: *Who Needs Emotions? The Brain Meets the Robot*. Ed. by Jean-Marc Fellous and Michael A. Arbib. New York: Oxford University Press, 2005, pp. 117–146.
- [138] Alec Radford, Rafal Jozefowicz, and Ilya Sutskever. *Learning to Generate Reviews and Discovering Sentiment*. Preprint at <https://arxiv.org/abs/1704.01444>. 2017. DOI: [10.48550/arXiv.1704.01444](https://doi.org/10.48550/arXiv.1704.01444).
- [139] David Silver et al. “Reward Is Enough”. In: *Artificial Intelligence* 299 (2021), p. 103535. DOI: [10.1016/j.artint.2021.103535](https://doi.org/10.1016/j.artint.2021.103535).
- [140] E. B. Goldstein. *Encyclopedia of Perception*. University of Pittsburgh, USA, University of Arizona, USA: SAGE Publications, 2009.
- [141] Alessia Celeghin et al. “Basic emotions in human neuroscience: neuroimaging and beyond”. In: *Frontiers in psychology* 8 (2017), p. 1432.
- [142] Morten L Kringelbach and Kent C Berridge. “Towards a functional neuroanatomy of pleasure and happiness”. In: *Trends in cognitive sciences* 13.11 (2009), pp. 479–487.
- [143] Robert W Levenson. “The intrapersonal functions of emotion”. In: *Cognition & Emotion* 13.5 (1999), pp. 481–504.
- [144] Antonio Damasio. *The Strange Order of Things: Life, Feeling, and the Making of Cultures*. New York, NY: Pantheon Books, 2018.
- [145] Antonio Damasio. *Feeling & knowing: Making minds conscious*. New York, NY: Pantheon Books, 2021.
- [146] R. W. Levenson. “Human Emotions: A Functional View”. In: *The Nature of Emotion: Fundamental Questions*. Ed. by P. Ekman and R. J. Davidson. New York, NY: Oxford University Press, 1994, pp. 123–126.
- [147] Daniel Kahneman and Alan B. Krueger. “Developments in the Measurement of Subjective Well-Being”. In: *Journal of Economic Perspectives* 20.1 (2006), pp. 3–24.

- [148] P. Brickman and D. Campbell. “Hedonic Relativism and Planning the Good Society”. In: *Adaptation Level Theory: A Symposium*. Ed. by M. H. Apley. New York, NY: Academic Press, 1971, pp. 287–302.
- [149] T. P. Todd, D. Vurbic, and M. E. Bouton. “Behavioral and neurobiological mechanisms of extinction in Pavlovian and instrumental learning”. In: *Neurobiology of Learning and Memory* 108 (2013), pp. 52–64. DOI: [10.1016/j.nlm.2013.08.012](https://doi.org/10.1016/j.nlm.2013.08.012).
- [150] Christopher J. Burke and Philippe N. Tobler. “Coding of Reward Probability and Risk by Single Neurons in Animals”. In: *Frontiers in Neuroscience* 5 (2011). DOI: [10.3389/fnins.2011.00121](https://doi.org/10.3389/fnins.2011.00121).
- [151] W. Schultz, P. Dayan, and P. R. Montague. “A Neural Substrate of Prediction and Reward”. In: *Science* 275.5306 (Mar. 14, 1997), pp. 1593–1598.
- [152] Wolfram Schultz. “Neuronal Reward and Decision Signals: From Theories to Data”. In: *Physiological Reviews* 95.3 (July 2015), pp. 853–951. DOI: [10.1152/physrev.00023.2014](https://doi.org/10.1152/physrev.00023.2014).
- [153] Andrew B. Barron, Eirik Søvik, and Jennifer L. Cornish. “The Roles of Dopamine and Related Compounds in Reward-Seeking Behavior Across Animal Phyla”. In: *Frontiers in Behavioral Neuroscience* 4 (Oct. 2010), p. 163. DOI: [10.3389/fnbeh.2010.00163](https://doi.org/10.3389/fnbeh.2010.00163).
- [154] Katerina Nerantzaki, Anastasia Efklides, and Panayiota Metallidou. “Epistemic Emotions: Cognitive Underpinnings and Relations with Metacognitive Feelings”. In: *New Ideas in Psychology* 63 (Dec. 2021), p. 100904. DOI: [10.1016/j.newideapsych.2021.100904](https://doi.org/10.1016/j.newideapsych.2021.100904).
- [155] H. Bless, K. Fiedler, and F. Strack. *Social cognition: How individuals construct social reality*. Hove and New York: Psychology Press, 2004, p. 2.
- [156] M. G. Haselton, D. Nettle, and P. W. Andrews. “The evolution of cognitive bias”. In: *The Handbook of Evolutionary Psychology*. Ed. by D. M. Buss. Hoboken, NJ, US: John Wiley & Sons Inc., 2005, pp. 724–746.
- [157] Robert M. Seyfarth and Dorothy L. Cheney. “Signalers and receivers in animal communication”. In: *Annual Review of Psychology* 54.1 (2003), pp. 145–173. DOI: [10.1146/annurev.psych.54.101601.145121](https://doi.org/10.1146/annurev.psych.54.101601.145121).
- [158] Q. Yang et al. “Differentiating the influence of incidental anger and fear on risk decision-making”. In: *Physiology & Behavior* 184 (2018), pp. 179–188.
- [159] A. G. Barto, R. S. Sutton, and C. W. Anderson. “Neuronlike elements that can solve difficult learning control problems”. In: *IEEE Transactions on Systems, Man, and Cybernetics* 13.5 (1983), pp. 835–846.
- [160] P. R. Montague, P. Dayan, and T. J. Sejnowski. “A framework for mesencephalic dopamine systems based on predictive Hebbian learning”. In: *The Journal of Neuroscience* 16.5 (1996), pp. 1936–1947.
- [161] Yuji Takahashi, Geoffrey Schoenbaum, and Yael Niv. “Silencing the Critics: Understanding the Effects of Cocaine Sensitization on Dorsolateral and Ventral Striatum in the Context of an Actor/Critic Model”. In: *Frontiers in Neuroscience* 2.1 (2008), pp. 86–99. DOI: [10.3389/neuro.01.014.2008](https://doi.org/10.3389/neuro.01.014.2008).
- [162] Reza Zadeh. *Twitter post*. x.com/Reza_Zadeh/status/1778641959667859744. Accessed: 2024-10-08. 2024.
- [163] Y. Bengio. “Learning Deep Architectures for AI”. In: *Foundations and Trends in Machine Learning* 2.1 (2009), pp. 1–55. DOI: [10.1561/22000000006](https://doi.org/10.1561/22000000006).

-
- [164] S. Lange and M. Riedmiller. “Deep auto-encoder neural networks in reinforcement learning”. In: *The 2010 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2010, pp. 1–8. DOI: [10.1109/IJCNN.2010.5596468](https://doi.org/10.1109/IJCNN.2010.5596468).
 - [165] Diederik P. Kingma and Max Welling. *Auto-Encoding Variational Bayes*. Preprint at <https://doi.org/10.48550/arXiv.1312.6114>. 2013.
 - [166] Sepp Hochreiter and Jürgen Schmidhuber. “Long Short-Term Memory”. In: *Neural Computation* 9.8 (1997), pp. 1735–1780. DOI: [10.1162/neco.1997.9.8.1735](https://doi.org/10.1162/neco.1997.9.8.1735).
 - [167] Kyunghyun Cho et al. “Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation”. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1724–1734. DOI: [10.3115/v1/D14-1179](https://doi.org/10.3115/v1/D14-1179).
 - [168] Daniel Kahneman. *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux, 2011.
 - [169] Craig A Smith and Phoebe C Ellsworth. “Patterns of cognitive appraisal in emotion”. In: *Journal of Personality and Social Psychology* 48 (1985), pp. 813–838.
 - [170] L Fehr and JA Russell. “Concept of emotion viewed from a prototype perspective”. In: *Journal of Experimental Psychology: General* 113.3 (1984), pp. 464–486. DOI: [10.1037/0096-3445.113.3.464](https://doi.org/10.1037/0096-3445.113.3.464).
 - [171] P. R. Kleinginna and A. M. Kleinginna. “A categorized list of emotion definitions, with suggestions for a consensual definition”. In: *Motivation and Emotion* 5.4 (1981), pp. 345–379. DOI: [10.1007/bf00992553](https://doi.org/10.1007/bf00992553).
 - [172] Greg Brockman et al. *OpenAI Gym*. Preprint at <https://arxiv.org/abs/1606.01540>. 2016.
 - [173] Alan S. Cowen and Dacher Keltner. “Self-report captures 27 distinct categories of emotion bridged by continuous gradients”. In: *Proceedings of the National Academy of Sciences of the United States of America* 114.38 (2017), E7900–E7909. DOI: [10.1073/pnas.1702247114](https://doi.org/10.1073/pnas.1702247114).
 - [174] Albert Mehrabian and James A. Russell. *An approach to environmental psychology*. Cambridge, MA: MIT Press, 1974.
 - [175] P. J. Lang. “Behavioral treatment and bio-behavioral assessment: computer applications”. In: *Technology in mental health care delivery systems*. Ed. by J. B. Sidowski, J. H. Johnson, and T. A. Williams. Norwood, NJ: Ablex, 1980, pp. 119–137.
 - [176] Terry K. Koo and Mae Y. Li. “A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research”. In: *Journal of Chiropractic Medicine* 15.2 (2016), pp. 155–163. DOI: [10.1016/j.jcm.2016.02.012](https://doi.org/10.1016/j.jcm.2016.02.012).
 - [177] Domenic V. Cicchetti. “Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology”. In: *Psychological Assessment* 6.4 (1994), pp. 284–290. DOI: [10.1037/1040-3590.6.4.284](https://doi.org/10.1037/1040-3590.6.4.284).
 - [178] M. M. Bradley and P. J. Lang. *Affective Norms for English Words (ANEW): Instruction Manual and Affective Ratings*. Technical Report C-1. Center for Research in Psychophysiology, University of Florida, 1999, pp. 25–36.
 - [179] Jaime Redondo et al. “The Spanish Adaptation of ANEW (Affective Norms for English Words)”. In: *Behavior Research Methods* 39.3 (2007), pp. 600–605. DOI: [10.3758/BF03193031](https://doi.org/10.3758/BF03193031).

- [180] Agnieszka Landowska. “Towards New Mappings between Emotion Representation Models”. In: *Applied Sciences* 8.2 (Feb. 12, 2018), p. 274. DOI: [10.3390/app8020274](https://doi.org/10.3390/app8020274).
- [181] Graham G. Scott, Anne Keitel, Marc Becirspahic, et al. “The Glasgow Norms: Ratings of 5,500 words on nine scales”. In: *Behavior Research Methods* 51 (2019), pp. 1258–1270. DOI: [10.3758/s13428-018-1099-3](https://doi.org/10.3758/s13428-018-1099-3).
- [182] Joshua Achiam. *OpenAI Spinning Up*. Github. Retrieved from <https://spinningup.openai.com>. 2018.
- [183] John Schulman et al. *Proximal Policy Optimization Algorithms*. Preprint at <https://arxiv.org/abs/1707.06347>. 2017.
- [184] François Chollet et al. *Keras*. Github. Retrieved from <https://keras.io>. 2015.
- [185] Fabian Pedregosa et al. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12.Oct (2011), pp. 2825–2830.
- [186] Bokeh Development Team. *Bokeh: Python library for interactive visualization*. <https://bokeh.org>. 2014.
- [187] Daniel Kahneman and Amos Tversky. “Advances in prospect theory: Cumulative representation of uncertainty”. In: *Journal of Risk and Uncertainty* 5.4 (1992), pp. 297–323.
- [188] Nieves Gurbindo and José Eugenio Ortega. “Adaptación de las escalas de placer, activación y dominancia de Mehrabian y Russell en sujetos hispanoparlantes”. In: *Revista de Psicología Social* 4.2 (1989), pp. 179–183.
- [189] Sandrine Detandt, Christophe Leys, and Ariane Bazan. “A French Translation of the Pleasure Arousal Dominance (PAD) Semantic Differential Scale for the Measure of Affect and Drive”. In: *Psychologica Belgica* 57.1 (2017), pp. 17–31. DOI: [10.5334/pb.340](https://doi.org/10.5334/pb.340).
- [190] Klaus R. Scherer. “What are emotions? And how can they be measured?” In: *Social Science Information* 44.4 (2005), pp. 695–729. DOI: [10.1177/0539018405058216](https://doi.org/10.1177/0539018405058216).
- [191] Thomas Dixon. “Emotion”: The History of a Keyword in Crisis”. In: *Emotion Review* 4.4 (2012), pp. 338–344. DOI: [10.1177/1754073912445814](https://doi.org/10.1177/1754073912445814).
- [192] Kate Crawford. “Time to regulate AI that interprets human emotions”. In: *Nature* 592 (2021), p. 167. DOI: [10.1038/d41586-021-00868-5](https://doi.org/10.1038/d41586-021-00868-5).
- [193] Gonçalo Assunção et al. “An Overview of Emotion in Artificial Intelligence”. In: *IEEE Transactions on Artificial Intelligence* 3.6 (2022), pp. 867–886. DOI: [10.1109/TAI.2022.3160893](https://doi.org/10.1109/TAI.2022.3160893).
- [194] Matthew Groh et al. “Computational Empathy Counteracts the Negative Effects of Anger on Creative Problem Solving”. In: *2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII)*. Nara, Japan: IEEE, 2022, pp. 1–8. DOI: [10.1109/ACII52823.2022.9896318](https://doi.org/10.1109/ACII52823.2022.9896318).
- [195] Kingson Man and Antonio Damasio. “Homeostasis and soft robotics in the design of feeling machines”. In: *Nature Machine Intelligence* 1.10 (2019), pp. 446–452.
- [196] Mariana Goya-Martinez. “The Emulation of Emotions in Artificial Intelligence”. In: *Emotions, Technology, and Design*. San Diego, CA, USA: Elsevier, 2016, pp. 171–186. DOI: [10.1016/B978-0-12-801872-9.00008-9](https://doi.org/10.1016/B978-0-12-801872-9.00008-9).