The Role of Progress-Based Intrinsic Motivation in Learning : Evidence from Human Behavior and Future Directions

Alexandr Ten

université de BORDEAUX

Inria

The Role of Progress-Based Intrinsic Motivation in Learning

Evidence from Human Behavior and Future Directions

By **Alexandr Ten**

Under the supervision of
Pierre-Yves Oudeyer and **Jacqueline Gottlieb**

In partial fulfillment of the requirements
for the degree of Doctor of Philosophy

University of Bordeaux
Graduate School of Mathematics and Computer Science
Major in Computer Science

Defended on April 19, 2022
Submitted on April 27, 2022

Composition of the jury:

Hélène Sauzéon	Professor	University of Bordeaux	President
Goren Gordon	Professor	Tel Aviv University	Reviewer
Todd Gureckis	Professor	New York University	Reviewer
Julia Leonard	Assistant Professor	Yale University	Examiner
Tom Griffiths	Professor	Princeton University	Examiner
Mathias Pessiglione	Research Director	INSERM	Examiner
Jacqueline Gottlieb	Professor	Columbia University	Supervisor
Pierre-Yves Oudeyer	Research Director	INRIA	Supervisor

The Role of Progress-Based Intrinsic Motivation in Learning

ABSTRACT

Intrinsic motivation – the desire to do things for their inherent joy and pleasure – has received its first share of scientific attention over 70 years ago, ever since we saw monkeys solving puzzles for free. Since then, research on intrinsic motivation has been steadily gaining momentum. We have come to understand, in the context of learning and discovery, that intrinsic motivation (namely, intrinsically motivated information-seeking) is foundational for the biological and technological success of our species. But where does intrinsic motivation to learn and seek information come from? Today, with the thriving synergy between perpetually advancing fields of psychology, neuroscience, and computer science, we are well positioned to investigate this question.

The *Learning Progress Hypothesis* (LPH) proposes that humans are motivated by feelings of and/or beliefs about progress in knowledge (including progress in competence). In artificial learners, progress-based intrinsic motivation enables autonomous exploration of the environment (including the agent's own body), resulting in better performance, more efficient learning, and richer skill sets. Due to similar computational challenges facing artificial and biological learners, researchers have proposed that progress-based intrinsic motivation might have evolved in humans to help us transition from babies with few skills and little knowledge to knowledgeable grownups capable of performing many sophisticated tasks. The Learning Progress Hypothesis (LPH) is attractive, not only because it is consistent with several studies of human curiosity, but also because it resonates with existing theories on metacognitive self-regulation in learning. However, the LPH has not been extensively studied using behavioral experimentation.

This thesis provides an empirical examination of the LPH. We introduce a novel experimental paradigm where participants explore multiple learning activities, some easy, others difficult. The activities involve guessing the binary category of randomly presented stimuli. To let their intrinsic motivation shine, we did not provide any material incentives encouraging specific behaviors or strategies – we simply observed which activities people engaged in and how their knowledge about these activities unfolded over time. We present statistical analyses and a computational model that support the LPH.

This thesis also suggests ideas for future investigations into progress-based motivation. These ideas are inspired by a pilot study in which we asked participants to practice a naturalistic sensorimotor skill (a video game) over the course of 3 sessions spanning 5 days. At the end

of each session, participants reported their subjective judgments of past and future progress, as well as their evolving beliefs about their perceived competence, self-efficacy beliefs, and intrinsic motivation. In support of the LPH, participants' subjective judgments correlated with the objective improvement. However, contrary to the LPH's prediction, objective and subjective progress measures did not show reliable relationships with verbal and behavioral measures of intrinsic motivation. Instead, progress measures were in strong relationships with beliefs about task learnability, which in turn predicted intrinsic motivation. Based on these findings, we suggest a novel mechanism in which learning progress interacts with intrinsic motivation via subjective beliefs.

We conclude the thesis with an extended discussion of our findings, where we examine some limitations of our experiments and propose promising future steps. In summary, we believe the behavioral paradigms introduced in this thesis should be reused to not only replicate our results, but also to advance the scientific research of intrinsically motivated information-seeking.

Keywords: intrinsic motivation / information-seeking / learning progress / curiosity / interest / self-regulated learning / active learning

Flowers Laboratory | INRIA

200 Av. de la Vieille Tour, 33405 Talence, France

Le rôle de la motivation intrinsèque basée sur le progrès dans l'apprentissage

RÉSUMÉ

La motivation intrinsèque - le désir de faire les choses pour la joie et le plaisir inhérent qu'ils procurent - a suscité l'intérêt des chercheurs pour la première fois il y a plus de 70 ans, en voyant des singes résoudre des énigmes gratuitement. Depuis, la recherche sur la motivation intrinsèque n'a cessé de prendre de l'ampleur. Dans le contexte de l'apprentissage et des comportements de découverte, il est aujourd'hui entendu , que la motivation intrinsèque (c'est-à-dire la recherche intrinsèquement motivée d'informations) est fondamentale au progrès biologique et technologique de notre espèce. Mais d'où vient la motivation intrinsèque pour apprendre et pour rechercher des informations ? Aujourd'hui, avec la synergie florissante entre les domaines en constante évolution de la psychologie, des neurosciences et de l'informatique, nous sommes en capacité d'étudier cette question.

L'*hypothèse de progrès d'apprentissage* (*Learning Progress Hypothesis*, ou LPH en anglais) propose que les humains sont motivés par leurs sentiments et/ou croyances relatifs à leurs progrès de connaissances (y compris les progrès dans les compétences). Chez les apprenants artificiels, la motivation intrinsèque basée sur le progrès d'apprentissage permet une exploration autonome de l'environnement (y compris le propre corps de l'agent), ce qui se traduit par de meilleures performances, un apprentissage plus efficace et un ensemble de compétences plus riches. En raison des défis informatiques similaires auxquels sont confrontés les apprenants artificiels et biologiques, des chercheurs ont proposé que la motivation intrinsèque basée sur le progrès ait pu évoluer chez l'Homme, pour nous aider à passer de bébés avec peu de compétences et peu de connaissances, à des adultes bien informés capables d'effectuer de nombreuses tâches sophistiquées. La LPH est attrayante, non seulement parce qu'elle est cohérente avec plusieurs études sur la curiosité humaine, mais aussi parce qu'elle résonne avec les théories existantes sur l'autorégulation métacognitive dans l'apprentissage. Cependant, la LPH n'a pas été largement étayée par l'expérimentation comportementale.

Cette thèse propose un examen empirique de la LPH. Nous introduisons un nouveau paradigme expérimental où les participants explorent plusieurs activités d'apprentissage, certaines faciles, d'autres difficiles. Les activités consistent à deviner la catégorie binaire de stimuli présentés au hasard. Pour laisser briller leur motivation intrinsèque, nous n'avons fourni aucune incitation matérielle encourageant des comportements ou des stratégies spécifiques - nous avons simplement observé dans quelles activités les personnes s'engageaient et comment leurs

connaissances sur ces activités se développaient au fil du temps. Nous présentons des analyses statistiques et un modèle de calcul prenant en charge la LPH.

Des idées pour de futures investigations sur la motivation basée sur le progrès sont également proposées dans cette thèse. Ces idées sont inspirées d'une étude pilote dans laquelle nous avons demandé aux participants d' entraîner une compétence sensorimotrice écologique (un jeu vidéo) au cours de 3 sessions réparties sur 5 jours. À la fin de chaque session, les participants ont rapporté leurs jugements subjectifs sur les progrès passés et futurs, mais aussi leurs croyances dans le temps concernant leur compétence perçue et leur auto-efficacité, et leur motivation intrinsèque. À l'appui de la LPH, les jugements subjectifs des participants étaient en corrélation avec l'amélioration objective. Cependant, contrairement à la prédiction de la LPH, les mesures de progrès objectives et subjectives n'ont pas montré de relations fiables avec les mesures verbales et comportementales de la motivation intrinsèque. Au lieu de cela, les mesures de progrès étaient en forte relation avec les croyances sur l'apprentissage des tâches, qui à leur tour prédisaient la motivation intrinsèque. Sur la base de ces résultats, nous suggérons un nouveau mécanisme dans lequel les progrès d'apprentissage interagissent avec la motivation intrinsèque via des croyances subjectives.

Nous concluons la thèse par une discussion approfondie de nos résultats, où nous examinons certaines limites de nos expériences et proposons des étapes futures prometteuses. En résumé, nous pensons que les paradigmes comportementaux introduits dans cette thèse devraient être réutilisés non seulement pour reproduire nos résultats, mais aussi pour faire avancer la recherche scientifique sur la recherche d'informations intrinsèquement motivée.

Mots-clés : motivation intrinsèque / recherche d'informations / progrès de l'apprentissage / curiosité / intérêt / apprentissage autorégulé / apprentissage actif

Équipe Flowers | INRIA

200 Av. de la Vieille Tour, 33405 Talence, France

RÉSUMÉ LONG

Au cours des trois dernières décennies, le domaine multidisciplinaire des sciences cognitives a connu une intégration approfondie entre l'intelligence artificielle (en particulier l'apprentissage automatique), la psychologie et les neurosciences. Cela est dû en partie au sujet poursuivi par ces trois sous-disciplines connues sous le nom d'*apprentissage actif* - le processus dans lequel un agent exerce un contrôle sur les situations d'apprentissage qu'il rencontre. Il est de plus en plus reconnu que l'apprentissage actif est essentiel pour rendre les agents artificiels plus autonomes. Parallèlement, les psychologues et neuroscientifiques s'intéressent de plus en plus aux processus décisionnels responsables de l'allocation des ressources cognitives (par exemple, l'attention). Le défi pour les chercheurs en cognition artificielle et humaine a été de comprendre comment l'apprentissage actif peut être organisé pour permettre une acquisition de connaissances efficace et ouverte dans le contexte d'environnements complexes et de ressources limitées.

Un concept crucial invoqué dans la littérature sur l'apprentissage actif est la *motivation intrinsèque*. La motivation intrinsèque fait référence à un type particulier d'incitation (ou de récompense) pour s'engager dans des situations d'apprentissage spécifiques. Elle peut être opposée à la *motivation extrinsèque*, qui oblige les agents à s'engager dans des situations d'apprentissage sur la base de leur utilité extrinsèque, car s'engager dans ces situations aide les agents à atteindre un résultat dissociable. La curiosité et l'intérêt sont depuis longtemps reconnus comme des manifestations de la motivation intrinsèque chez l'Homme. Lorsque nous sommes curieux ou intéressés, nous sommes motivés à rechercher des informations sans nous attendre à ce qu'elles nous apportent de la nourriture ou de l'argent - nous voulons simplement savoir et apprendre.

Cette thèse porte sur les mécanismes de l'apprentissage intrinsèquement motivé chez l'Homme. Les apprenants actifs améliorent leurs connaissances en recherchant et en traitant des informations. La recherche humaine d'informations intrinsèquement motivée est organisée, comme en témoignent des curiosités spécifiques et des intérêts pour des sujets particuliers. Cependant, nous manquons toujours d'une description complète de ce qui détermine la curiosité humaine et l'intérêt pour le monde, qui offrent bien plus que ce que tout individu peut éventuellement apprendre. Qu'est-ce qui détermine la curiosité ou l'intérêt des humains lorsqu'ils sont libres d'explorer le monde ?

Une proposition convaincante, issue de l'intelligence artificielle, soutient que les humains sont motivés à s'engager dans des situa-

tions d'apprentissage censées améliorer leurs connaissances. Cette idée - connue sous le nom de l'*hypothèse de progrès d'apprentissage* (ou *Learning Progress Hypothesis*, ou LPH, en anglais) - est plutôt nuancée. Premièrement, il existe différents types de connaissances que les apprenants capables, comme les humains, peuvent espérer améliorer. Deuxièmement, il existe plusieurs façons d'estimer l'amélioration des connaissances. Enfin, il existe plusieurs possibilités quant à la façon dont les jugements sur l'amélioration des connaissances peuvent affecter la motivation à s'engager dans des situations d'apprentissage.

Cette thèse poursuit deux objectifs principaux. Premièrement, elle vise à présenter et à discuter des preuves empiriques de la LPH. Deuxièmement, elle vise à explorer le mécanisme potentiel par lequel les jugements de progrès sont représentés et comment ces jugements interagissent avec la motivation à poursuivre des activités non instrumentales.

Nous commençons par identifier les fonctions computationnelles de la motivation intrinsèque (Chapitre 2). Une fonction importante est d'aider les apprenants autonomes à accumuler des répertoires vastes et variés de compétences. Les apprenants autonomes sont confrontés au problème de la génération, de la sélection et de l'apprentissage de leurs propres tâches. Dans des environnements suffisamment complexes, l'espace des tâches est vaste et hiérarchisé (c'est-à-dire que certaines tâches ne peuvent être accomplies que si l'apprenant a maîtrisé des tâches de niveau inférieur, par exemple, apprendre à marcher avant de pouvoir s'approcher de différents objets). Un bon système de motivation intrinsèque garantit que l'agent tente des tâches adaptées à ses capacités actuelles, et en même temps pousse les agents à se mettre au défi au-delà de ce qui est déjà familier.

Ensuite, dans le Chapitre 3, nous discutons plus en détail les raisons pour lesquelles l'accumulation de connaissances (y compris l'accumulation de compétences) est avantageuse d'un point de vue évolutif et passons en revue la littérature psychologique et neuroscientifique sur l'apprentissage intrinsèquement motivé chez l'Homme. Là, nous discutons des preuves existantes en faveur de l'idée que les humains valorisent l'information comme un bien en soi (c'est-à-dire au-delà de sa valeur *instrumentale*). Nous examinons ensuite les facteurs situationnels qui déclenchent la recherche d'informations non instrumentales (c'est-à-dire la curiosité et l'intérêt) et expliquons les mécanismes affectifs/motivationnels par lesquels la valeur intrinsèque de l'information renforce le processus d'acquisition autonome des connaissances. Ici aussi, nous fournissons une introduction complète de la LPH et évaluons certains travaux existants sur la recherche d'informations relatives à cette hypothèse. De plus, nous identifions le manque de preuves empiriques pour la LPH, préparant ainsi le terrain pour les expériences comportementales décrites dans les chapitres suivants.

Le Chapitre 4 introduit un nouveau paradigme comportemental conçu pour capturer le processus d'apprentissage autorégulé par lequel les humains explorent un espace de tâches. La tâche expérimentale consiste en 4 activités d'apprentissage dans lesquelles l'apprenant peut acquérir une règle de généralisation pour catégoriser les exemplaires de l'ensemble des stimuli de l'activité. Chaque activité est représentée par un ensemble de stimuli composé de stimuli visuels, représentés comme des personnages de monstres de dessins animés, qui varient selon une ou deux dimensions. Chaque ensemble de stimuli est associé à une règle de catégorisation unique (que les participants ne connaissent pas à l'avance, mais peuvent apprendre en interagissant avec l'activité correspondante). Les activités varient en difficulté, allant d'une activité très simple où les stimuli peuvent être classés en fonction d'une seule dimension variable, à une activité plus difficile où les stimuli varient selon deux dimensions qui déterminent conjointement la catégorie du stimulus. Fondamentalement, l'une des activités est non généralisable, car elle n'a pas de règle pour la catégorisation des stimuli. Cette activité impossible à apprendre a été incluse pour représenter des tâches de la vie réelle qui sont concevables, mais qui ne fournissent aucun progrès d'apprentissage car elles sont soit trop difficiles, soit carrément impossibles. Les participants ont eu un nombre limité d'interactions avec les activités, mais ils étaient libres de choisir l'activité à entreprendre à tout moment.

Dans cette étude, nous montrons que les individus varient considérablement dans leurs styles d'exploration. Fait important, de nombreux participants ont décidé de se mettre au défi d'aller au-delà des activités faciles à apprendre et ont passé du temps sur des problèmes plus difficiles. De plus, en utilisant un modèle choix-utilité multivarié et les données de choix d'activité, nous avons ajusté la sensibilité de chaque participant à la compétence et aux progrès d'apprentissage. Nos comparaisons de modèles révèlent que les participants avaient tendance à choisir des activités basées sur ces deux composantes, fournissant un soutien empirique a la LPH. Nous montrons également que les participants qui étaient particulièrement sensibles aux progrès d'apprentissage apprenaient mieux les activités apprenables en évitant l'activité impossible.

Dans notre seconde expérimentation (Chapitre 5), nous nous sommes attachés à approfondir le mécanisme métacognitif sous-jacent au calcul des progrès d'apprentissage. Nous introduisons un autre paradigme comportemental où les participants sont chargés de pratiquer une activité sensorimotrice présentée comme un jeu vidéo, appelée *Lunar Lander*. Notre tâche tente d'émuler le processus d'apprentissage naturaliste, dans lequel une activité est pratiquée sur plusieurs sessions et peut être interrompue par d'autres tâches quotidiennes. Ainsi, nous avons demandé aux participants de pratiquer le jeu sur 3 sessions réparties sur 5 jours. Au cours de chaque séance d'entraînement, nous

avons enregistré les taux de réussite des participants, ainsi que d'autres mesures comportementales potentiellement utiles. Après chaque session, nous avons administré un questionnaire sondant les participants sur leurs sentiments concernant le progrès (ou détérioration) de la compétence, les auto-évaluations de la compétence, diverses croyances sur l'apprentissage de la tâche et la motivation intrinsèque concernant la tâche. Aussi, à la fin de chaque session, les participants se voyaient proposer une pratique facultative. Cela a servi de mesure comportementale de la motivation intrinsèque - accepter une pratique facultative ne fournit aucune incitation, autre que de profiter du jeu ou de s'améliorer.

Les résultats de cette étude montrent que les gens sont sensibles aux progrès objectifs des compétences lorsqu'ils portent des jugements subjectifs sur les progrès. Nous montrons également que l'amélioration subjective et objective est fortement corrélée avec les croyances subjectives sur l'apprentissage de la tâche - à savoir que la pratique de la tâche peut éventuellement conduire à la maîtrise, et que les participants finiraient par apprendre la tâche. Alors que la corrélation entre l'amélioration subjective / objective et la motivation intrinsèque était faible et incohérente, cette dernière était fortement corrélée aux croyances en matière d'apprentissage. Cela suggère que l'effet des perceptions subjectives des progrès d'apprentissage est médiatisé par des croyances explicites en matière d'apprentissage - une hypothèse qui mérite d'être approfondie par des études complémentaires Enfin, nous montrons que le taux de réussite et les croyances d'apprentissage peuvent ensemble prédire la motivation intrinsèque, telle que mesurée par l'acceptation de la pratique facultative. Précisément, nos résultats suggèrent qu'une mauvaise performance sur une tâche peut entraîner un engagement plus motivé intrinsèquement dans la tâche si elle est considérée comme apprenable.

Dans le Chapitre 6, nous fournissons une discussion approfondie des contributions empiriques à la lumière de l'ensemble de la thèse. Nous évaluons de manière critique les deux paradigmes comportementaux et suggérons des poursuites de travail pour les recherches futures. Un élément important à retenir pour l'avenir est d'être conscient des multiples processus d'apprentissage qui se déroulent en parallèle lors de l'exploration autorégulée des activités d'apprentissage. Nous appelons également à des études plus approfondies des mécanismes métacognitifs basés sur la performance des jugements de progrès, et identifions deux défis importants que ces études impliquent. L'un des défis consiste à comprendre les principes généraux qui sous-tendent la représentation des tâches (c'est-à-dire comment les apprenants représentent leur compétence dans des tâches qui fournissent des retours épars et/ou fortement biaisés). Un autre défi consiste à comprendre l'étendue temporelle des jugements de progrès (c'est-à-dire à quel(s) point(s) de référence les apprenants comparent-ils leurs

niveaux actuels de performance). Dans l'ensemble, cette thèse peut être considérée comme un tremplin vers une étude plus approfondie de la motivation intrinsèque basée sur le progrès pour la recherche d'informations.

I dedicate this dissertation to the cherished memory of
Igor Ten (1963 – 2020)
who spent his selfless existence
so that I could follow my dreams and accomplish my goals.

I dedicate this dissertation to my daughter
Alissa Ten
who entered this world right before the work was finished.
Let your life be filled with curiosity and joy.

ACKNOWLEDGMENTS

I feel incredibly fortunate to have met numerous people who have helped along this journey.

First and foremost, my praise and gratitude go to my supervisors Pierre-Yves Oudeyer and Jacqueline Gottlieb, who were always generous with time, advice, and compassion. I thank them, in equal measure, for their unrivaled professionalism, for words of encouragement, and for their endless patience. Without them, this thesis would have never seen the light of day.

I would like to extend my gratitude to Clément Moulin-Frier for his help in reviewing once foreign literature; and to Hélène Sauzéon for her openness and guidance in preparing the procedure for the metacognition study.

I would like to acknowledge the contributions of Celeste Kidd and Amanda Yung in implementing the "monster task".

I would also like to thank the members of my Comité de Suivi, Frédéric Alexandre and Alexandre Zénon, for their genuine concern about my progress and for their constructive advice on how to improve.

Many thanks go to my fellow scholars and good friends, Cédric Colas, Grgur Kovac, William Schueller, Benjamin Clément, Eleni Nisioti, Masataka Sawayama, Guillermo Valle, Rania Abdelghani, Maxime Adolphe, Tallulah Gilliard, Anthony Strock, Pramod Kaushik, Mayalen Etcheverry, Laetitia Teodorescu, Tristan Karch, Remy Portelas, Rajkumar Darbar, Biswajit Das, Maxime Balan, Adrien Laversanne-Finot, Florian Golemo, Alexandre Péré, Théo Segonds, Sébastien Forestier, and Chris Reinke – some of the smartest and kindest people I could meet. Probably without realizing it, these people always inspired me to learn and improve. And it is thanks to them that I felt a sense of belonging despite being thousands of kilometers away from home.

Special thanks go to Hew Gill and Jay McClelland for their invaluable guidance in preparation for this journey. They kindly offered their time and support when they did not have to, and for that I am eternally grateful.

Last, but by no means least, I am deeply grateful to two of the most important women in my life: my wife Anastassiya Tsoy and to my mother Larissa Ten for their unwavering support, understanding, and love.

PUBLICATIONS

Some of the research leading to this thesis has appeared previously (or is in press) in the following publications:

- Ten A., Kaushik P., Oudeyer P-Y., & Gottlieb, J. (2022). **Humans monitor learning progress in curiosity-driven exploration**. *Nature Communications, 12(1)*. URL: https://www.nature.com/articles/s41467-021-26196-w

- Ten A., Gottlieb, J., & Oudeyer P-Y. (2021). **Intrinsic Rewards in Human Curiosity-Driven Exploration: An Empirical Study**. *Proceedings of the Annual Meeting of the Cognitive Science Society, 43.*. URL: https://escholarship.org/uc/item/13b6p5ms

- Ten A., Oudeyer P-Y., & Moulin-Frier, C. (to be published online in May 2022). **Curiosity-driven Exploration: Diversity of Mechanisms and Functions**. In I. Cogliati Dezza, C. M. Wu, & E. Schilz (eds). *The Drive for Knowledge: The Science of Human Information Seeking* (Cambridge University Press), URL: https://www.cambridge.org/core/books/drive-for-knowledge/B7D7B75B7C1021D2F90089CC1B4C1DBF

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

ACh	acetylcholine
AI	artificial intelligence
DA	dopaminergic
DS	dorsal striatum
dwfPC	difficulty-weighted final percent correct
EG	external goal
EXP	exposure
IG	internal goal
IMI	Intrinsic Motivation Inventory
IPE	information prediction error
JOL	judgment of learning
LC	locus coeruleus
LHb	lateral habenula
LP	learning progress
LPH	Learning Progress Hypothesis
ML	machine learning

MSLQ	Motivated Strategies for Learning Questionnaire
NAc	nucleus accumbens
NAM	number of activities mastered
NASA-TLX	NASA task-load index
NE	norepinephrine
OFC	orbitofrontal cortex
OR	odds ratio
PC	percent correct
RL	reinforcement learning
RPE	reward-prediction error
SC	self-challenge
SIMS	Situated Motivated Strategies
SL	supervised learning
UL	unsupervised learning
VP	ventral pallidum
VTA	ventral tegmental area

1

SHORT OVERVIEW OF THE THESIS

1.1 INTRODUCTION

The excellence of human mind reflects the complexity of our environment. While complex, our plane of existence is governed by persistent physical laws, biological constraints, and social conventions. The essence of human intelligence, individual and cultural, is the ability to represent some of these regularities so that we can use them to our advantage. In that sense, humans are not born intelligent, but we come into the world with a capacity to grasp certain aspects of the immense complexity that surrounds us. We explore. We discover. We learn.

We learn the regularities of the world by observing it. However, mere observation would not get us this far. Our embodiment is crucial: we causally interact with our environments, and our interactions produce a special kind of regularities – those between our actions and observations. Learning how our actions affect our observations enables us to *seek* information and not merely absorb it. The world is far too complex to be represented entirely in our minds. Active control of sensory experience, that is information-seeking, helps us to allocate our limited cognitive resources efficiently across different domains of the world's structure.

Human information-seeking is organized. That is, we do not seek any kind of information, but instead tend to pursue specific pieces that we expect to be satisfactory. Satisfaction that we derive from information can be explained in various ways. One way is to assume that it derives from states that are clearly beneficial. For example, knowing where to get water might be satisfactory because it helps us to stay hydrated. However, sometimes information we expect to satisfy us seems to do little except make us informed. Where is the benefit of knowing what the other side of the moon looks like? This thesis attempts to gain insights for understanding why and how information can satisfy us when it has not extrinsic implications.

The pursuit of information for its inherent ability to make unknown things known is sometimes called *non-instrumental* information-seeking, because the sought out information does not appear to be instrumental towards anything other than learning from that information. Accordingly, motivation to seek non-instrumental information is called *intrinsic*, because the seeking is of information is motivated by what information can offer a good in itself. Humans recognize when they seek information in an intrinsically motivated way. When

this happens, we say that we are *curious* or *interested*. Whereas it is easy to defend the affirmation that curiosity and interest do not occur randomly and are not directed towards random pieces of information, it is much more difficult to tell, in a general way, when they occur and what they are about. To gain insights into non-instrumental information-seeking, this thesis will explore causes and consequences of mental states identified as curiosity and interest.

Scientific research of curiosity (and interest) has a long history and a broad interdisciplinary outreach. Perhaps because the human mind is so self-reflective and inquisitive, people have pondered about where the desire to know comes from way before science was established as a method for inquiry. Today, we find curiosity studied by scholars from all kinds of fields, including philosophy, psychology, biology, neuroscience, economics, and artificial intelligence. This speaks to the undeniable importance and complexity of the subject. The current state of scientific affairs makes the study of curiosity ever more exciting and promising. The synergy between relatively recent technological advents in neuroscience and computer science put us right on the verge of understanding the nuts and bolts of curiosity and interest.

With an increasing appreciation for why non-instrumental information-seeking is biologically adaptive (i.e., its function) comes a need to characterize the cognitive mechanisms that conform to the implied functions. The ultimate task is to compose a detailed neuro-computational story describing how incoming sensory information interacts with the existing knowledge to engender the subsequent motivation for constructing and executing behavioral interventions. For completeness, the story also needs to include an account of how the information gathered through these behavioral interventions is evaluated with respect to goals and needs of an individual. We might still be far away from a complete story, but some important parts have already been drafted. Inspired by the existing work, this thesis aims to contribute to the developing narrative by proposing tentative computational and algorithmic accounts of motivational and evaluative processes involved in intrinsically motivated information-seeking.

Much of the research on curiosity and interest is purely epistemic – scholars genuinely want to know why they want to know. However, understanding the mechanisms of curiosity and intrinsic motivation has several utilitarian ends. One of them is to develop artificial intelligence systems capable of autonomous and open-ended development[1]. The past couple of decades have made remarkable achievements in designing algorithms that learn to perform specific tasks really well, often on a superhuman level. Combining these learning algorithms with mechanisms of intrinsic motivation holds a promise of producing

1 As we hope to demonstrate in our further discussions, developing autonomous intelligent machines can also be viewed as a purely epistemic exercise. The goal of creating a truly intelligent robot, for example, can be motivated by the very challenge of the task itself, rather than to use this robot for something utiliterian.

truly intelligent systems that can autonomously learn multiple tasks equally well. Another important domain of applying the mechanistic understanding of curiosity and interest is education. Curiosity and interest have many positive effects on learning and persistence. Knowing how curiosity and interest develop may help us design environments, interventions, and pedagogical guidelines that promote learning at schools and universities. Thus, in addition to potentially revealing deep insights about the human nature, understanding how curiosity (and interest) works may help us build better technology and better society. Of course, the concrete objectives (see below) of this thesis are much humbler in scope, but we believe they are important steps (however small) on the journey towards greater aspirations.

This thesis has two main goals. First, we want to present and discuss some original evidence for the so-called Learning Progress Hypothesis (LPH), which concerns the computational mechanism underlying intrinsically motivated learning. Initially inspired by work in developmental robotics, this hypothesis states that the motivation to engage in a particular learning activity depends on how much cognitive improvement that activity is thought to offer. Assuming that cognitive-improvement judgments come from a metacognitive process of self-modeling, the second goal of this thesis is to advance novel propositions regarding the process by which improvement judgments arise and how they influence motivation.

1.2 METHODS AND CONTRIBUTIONS

Our approach relies on behavioral experimentation inspired by ideas from the computational literature on intrinsic motivation. Accordingly, we begin by reviewing the diversity of computational mechanisms and functions of intrinsically motivated artificial agents (Chapter 2). This step allows us to accomplish several things. First, we introduce the functional perspective on intrinsic motivation (what can intrinsic motivation do? What problems can it solve?). Second, we identify various assumptions and limitations associated with different approaches to implementing intrinsic motivation in artificial agents. Additionally, the principled classification of artificial systems according to mechanistic and functional dimension allows us to recognize the relatively less popular computational aspects of intrinsically motivated information-seeking, suggesting interesting research directions.

We then proceed by discussing the relevant background details of behavioral and neuroscientific research that motivate the central hypothesis of this thesis (Chapter 3). Specifically, we first review a substantial body of evidence promoting the idea that humans value information due to its inherent ability to reduce uncertainty and improve generalizable knowledge. Later, we identify the gap in our understanding of the mechanisms by which intrinsically motivated

(i.e., curiosity- or interest-driven) information-seeking contributes to continual knowledge accumulation in humans. Finally, we call upon the previously proposed yet very modestly contested hypothesis that human motivation might be driven by learning progress.

After setting the scene, we bring forth the original contributions of this thesis. We report on two behavioral experiments relying on two original behavioral paradigms, respectively. Each paradigm was designed to capture a specific facet of self-regulated learning. Both experiments were administered online, allowing us to efficiently collect the data. The first experiment was administered through the Amazon Mechanical Turk platform, enabling us to crowdsource a large number of participants in a short period of time. For the second experiment, we designed our own platform for remote experimentation. This platform enabled us to gain absolute control over the experimental procedure, which was needed as the experiment required participants to carry out a task on three separate days.

In the first experiment (Chapter 4), we were interested in what determines activity choices in humans in a non-instrumental context. To this end, we asked participants to explore a set of learning activities with varying degrees of complexity. Importantly, participants were free to choose to engage in any activity at any given time. We recorded people's responses in each activity, along with which activities they chose. Participants were paid for participation, but none of their behavior within the task influenced the amount of compensation. To analyze the data, we relied (mostly) on regression analyses and frequentist hypothesis testing. Additionally, the chapter proposes and discusses different variants of a computational model of trial-by-trail choice utility, inspired by the relevant work on the multi-armed bandit problem. The models were based on multiple utility components measuring latent intrinsic rewards. We fitted our models (separately for each participant) using numerical optimization implemented in a popular open-source software library. Finally, we relied on the Akaike Information Criterion to compare different model forms in order to infer which model provides the best account of our data. Although nuanced, our analyses lend empirical support for the progress-based motivation in self-regulated exploration and motivate the search for more specific process accounts. The exact details of our methodology in this study can be found in Chapter 4.

The second experiment (Chapter 5) is a stepping stone towards a detailed process account of progress-based motivation, explaining (1) how progress is computed and represented in a real-world sensorimotor learning activity, and (2) what is the mechanism by which this (potentially conscious) representation affects motivation. To begin addressing these questions, we designed and piloted a naturalistic video-game control task in which we can closely track participants' performances during learning. In order to explore performance-related

factors potentially influencing progress judgments, we asked participants to report their subjective feelings of improvement throughout practice, which spanned several minutes across three days. We broke down the practice time into multiple bouts in order to evaluate the fidelity of self-reported improvement judgments over different time periods. In addition to soliciting improvement judgments, we measured people's perceptions of task load, different aspects of situation intrinsic motivation, and situational learning/achievement beliefs. Being an exploratory study, we mostly relied on correlational analyses and frequentist hypothesis testing of regression models. Our analyses demonstrate the metacognitive sensitivity to feedback-based progress and the relationships between progress, beliefs about learning, and intrinsic motivation. Another significant finding of this study highlights the importance of considering the relationship between intrinsic motivation and competence-related aspects of the self-concept. Further details about this study are provided in Chapter 5.

Finally, in Chapter 6, we provide an extended discussion reflecting on our results in the context of the state-of-the-art. We critically examine the conclusions drawn from our studies and evaluate the potential of both behavioral paradigms for future research. Specifically, we call for a more precise and transparent control of different kinds of learning processes induced by experimental tasks that provide behavioral autonomy (e.g., what kinds of knowledge structures are involved? what kinds of objective are pursued?). We also motivate further research on feedback-based metacognitive evaluation of competence progress, and emphasize two important challenges for this research: (1) to understand the process of subjective task representation, and (2) to understand the temporal extent of progress computation.

Part I

STATE OF THE ART

2

COMPUTATIONAL MODELS OF INTRINSICALLY MOTIVATED EXPLORATION

Although defining learning can be controversial (Barron et al., 2015), in this chapter, we shall adopt a straightforward formulation from ML (machine learning; Jordan and Mitchell, 2015), where learning is defined as improvement on a task (or a set of tasks) with experience. An artificial agent is said to have improved on a task if – according to some well-defined criterion – it is able to perform the task better than it did prior to receiving learning experience. Thus, what drives improvement is a combination of the experience that comes in the form of data that the agent can represent and the learning algorithm (also called learning, or update rule) that specifies how the agent's innards change by processing the incoming data.

For a long time, major machine learning (ML) paradigms have been elaborating increasingly efficient and powerful learning algorithms that optimize pre-specified task criteria. For example, all traditional algorithms of supervised learning (SL) and unsupervised learning (UL) depend on a formally defined objective function which evaluates the agent's responses to stimuli and thus drives structural changes that result in better responses in the future. While these algorithms can learn many different tasks (e.g. image classification, natural language processing, visual scene parsing etc.), they are typically trained on datasets and objective functions assigned by the engineer. On the other hand, active learning agents (Cohn, Ghahramani, and Jordan, 1995; Thrun, 1994) feature algorithms that actively sample the data they learn from. This is particularly the case in reinforcement learning (RL) (Sutton and Barto, 2018) where agents have control over the sampled data by virtue of causal interactions with their environments. Here, learning is driven by an evaluative objective criterion which comes in the form of a reward function. Like in SL and UL, what the agent ends up learning is determined by a predefined criterion, but additional complications arise because of the need to sample experiences that help the agent improve. In realistic settings, only a tiny fraction of all possible experiences are relevant, which can be further complicated by the sparsity or deceptiveness of rewards (see Oudeyer, 2018). Sampling relevant experiences while avoiding noise is an important problem which will receive much attention in this chapter.

In contrast to ML agents that learn externally assigned tasks, biological agents, particularly humans, often have autonomy not only in how they choose experiences to learn tasks, but also in choosing what tasks (i.e., goals) to learn. Moreover, we often seek information when we

become curious/interested without being told what to be curious/interested about. Seeking information out of curiosity, rather than to achieve a separable outcome like food or money, is characterized in psychology as intrinsically motivated (Harlow, Harlow, and Meyer, 1950; Ryan and Deci, 2000). Intrinsically motivated behavior – including information-seeking – also includes pursuing self-selected goals that do not have proximal benefits for biological fitness (e.g., learning tango or climbing Everest). In addition to information-seeking phenomena like curiosity and interest, intrinsic motivation is also linked to other hallmarks of human behavior, such as creativity (Gross, Zedelius, and Schooler, 2020) and play (Chu and Schulz, 2020a).

Similarly to humans but unlike the traditional ML systems mentioned above, intrinsically motivated artificial agents control what they learn through autonomous and non-instrumental sampling of learning experiences. This sampling is achieved by considering various features of what we call *learning situations*. To understand what we mean by "learning situations" better and to illustrate how they relate to various mechanisms and functions of intrinsic motivation, consider the following scenarios:

- A robot trying random actions in its environment
- A rat exploring a maze to get familiar with its environment
- A toddler trying to build the tallest possible tower with toy blocks
- A curious student raising his hand to pose a question to her teacher
- A infant looking at where her mother is pointing
- A philosopher considering what book to read

All of these scenarios describe an agent interacting with its environment and thereby engaging in a learning situation[1]. What differentiates these learning situations is the mechanism by which information happens to be sampled. We can identify two core components of computational mechanisms of curiosity-driven exploration. The first component is the interface through which agents actively sample learning situations, i.e. the space in which they can make choices. The second component is the principle by which agents rank learning situations within this space. We describe these components more thoroughly in Section 2.1 (Mechanisms), but the examples above already illustrate that agents can sample learning situations completely at random (like the robot), by considering what actions to take (like the rat), what goals to pursue (like the toddler), or whom to ask (the student).

1 Of course, the agent may not learn upon every single interaction with the environment, but any interaction creates a situation where learning could happen.

Moreover, decisions between alternative actions, goals, or social peers, may be based on prior knowledge (in case of the philosopher), driven by competence (in case of the toddler), or influenced externally, e.g., by others (in case of the infant). Note different mechanisms illustrated by the examples imply distinct consequences, suggesting that they may serve distinct functions. In Section 2.2 (FUNCTIONS), we organize these functions along the axis of causal proximity to evolutionary fitness: from the rat exploring a maze to the philosopher pondering the most obscure questions. Finally, in Section 2.3 (USAGES) we discuss the relationships between artificial intelligence (AI) and psychology and briefly survey some of the use cases for various intrinsically motivated algorithms outside the realm of AI research.

2.1 MECHANISMS

As mentioned in the introduction, researchers in AI have proposed a wealth of algorithms for intrinsically-motivated exploration. These algorithms differ by how they address two related subproblems:

1. How to *parameterize* learning situations?
2. How to *choose* learning situations?

The first subproblem is addressed by defining the choice space to drive exploration (Moulin-Frier and Oudeyer, 2013). Intuitively, a choice space serves as a basis for accessing and assessing different learning situations. Actualizing choices in a choice space results in agent-environment interactions from which the agent can learn. This chapter reviews three kinds of choice spaces that consist of either actions, goal states, or social partners. These choice spaces correspond to different ways in which learning situations can be parameterized. Parameterizing learning situations on the basis of *actions* lets the agent consider what can be learned by performing these actions; parameterizing learning situations on the basis of *goals* lets the agent consider what can be learned by pursuing goals; parameterizing learning situations on the basis of *social partners* lets the agent consider what can be learned by interacting with others. We discuss the main ways in which choice spaces are specified and how they differ in Section 2.1.1 (MECHANISMS/*Exploration Bases*).

Given a fully specified choice space, the agent can make decisions within that space. Specifying this decision-making process addresses the second subproblem. Agents choose what to learn by following a certain strategy by which they assign "interestingness" to the available choices, thereby determining which learning situations are more likely to be approached. For example, choices can be driven by features of learning situations, such as the amount of knowledge a situation might bring or how novel it is; they can also be made completely at random e.g., Colas, Sigaud, and Oudeyer, 2018; Colas et al.,

2020. We provide a brief survey of these methods in Section 2.1.2 (MECHANISMS/*Exploration Strategies*).

2.1.1 *Exploration Bases*

We stated before that a learning situation arises whenever an agent interacts with its environment. Active interactions entail that the agent has to decide how to act in a given context, and upon deciding and acting, gets to observe the effects of its actions. How the agent explores, therefore, depends primarily on how the agent contextualizes its interactions. Specifically, the agent can explore by considering what actions to take, what goals to pursue, or what social partners to engage. Sets of actions, goals, or social partners provide a basis for comparing potential interactions. The rest of this section reviews different ways in which such bases are defined, used, and represented.

2.1.1.1 *Exploring by Choosing Actions*

One family of approaches considers agents which observe the current state of the environment as a context, choose actions to execute, and observe the resulting following state. Here, the objective of exploration is to select the actions which generate informative data for learning an internal model of causal dynamics through SL, UL or RL. For example, several SL-based robotic agents (Baranes and Oudeyer, 2009; Caligiore et al., 2008; Lefort and Gepperth, 2015; Oudeyer, Kaplan, and Hafner, 2007; Saegusa et al., 2009) maintain world models representing their knowledge about either forward dynamics (inferring future states from specific actions), or inverse dynamics (inferring the right actions to bring about specific states), or both. These systems learn from observations borne out of the actions they choose. Therefore, exploration of learning situations in these approaches corresponds to making choices in the action space. Examples of action-space exploration can also be found in RL settings (Bellemare et al., 2016; Bougie and Ichise, 2020a; Burda et al., 2018; Haber et al., 2018; Jaderberg et al., 2016; Pathak et al., 2017; Singh et al., 2010; Tang et al., 2017). In intrinsically motivated RL, agents learn behavioral policies by maximizing intrinsic rewards (e.g., rewards based on state novelty as in (Tang et al., 2017); on model prediction error as in (Pathak et al., 2017); or on surprise as in (Berseth et al., 2021)). In these systems, actions that bring about rewarding states get reinforced. Thus, the agent collects learning data (state transitions) by ranking actions according to their capacity to yield intrinsic rewards.

2.1.1.2 *Exploring by Choosing Goals*

Another family of approaches considers agents making choices in a goal space instead of an action space. In the general case, such

agents learn to represent and sample their own goals, i.e., they are autotelic (Colas et al., 2021b). This family of approaches is referred to as Intrinsically Motivated Goal-Exploration Processes (or IMGEP for short; (Colas et al., 2021b; Forestier et al., 2020)). Here, the notion of a goal is generalized: it refers to any set of constraints on any set of future sensorimotor representations (Colas et al., 2021b). This abstract conceptualization enables researchers to express all kinds of goals, ranging from particular world states (e.g., coordinates of the agent's hand must be equal to a specific x, y, and z), to constraints on entire behavioral trajectories and their linguistic descriptions (e.g., "water the plant and then feed the dog"; (Colas et al., 2020)). In all cases, goals are specified by two essential components: (1) goal representation that specifies the criteria and (2) goal-achievement function that signals whether the criteria are met. Goals are usually represented as numerical vectors (sometimes called goal embeddings) that comprise abstract goal spaces from which specific goals can be sampled, while goal achievement is evaluated using logical operations.

Examples of IMGEPs can be found in SL and UL contexts (Baranes, 2013; Forestier et al., 2020; Jordan and Rumelhart, 1992; Laversanne-Finot, Péré, and Oudeyer, 2018; Moulin-Frier and Oudeyer, 2013; Reinke, Etcheverry, and Oudeyer, 2020; Rolf, Steil, and Gienger, 2010; Takahashi et al., 2017). In these frameworks, agents autonomously engage in learning situations by attempting to reach self-selected goals. Note that this process is markedly different from the one described above, where the agent accesses learning situations by choosing among the available actions. In IMGEPs, the agent can consider any goal it may imagine, not just the states that its actions may bring about.

The domain of goal-conditioned RL works with agents that learn action policies conditioned on goals. These agents base their actions not only on the current state (as in traditional RL) but also the goal encoding, which means they can act differently in the same situation depending on what they are after (Schaul et al., 2016). Intrinsically motivated goal-conditioned RL builds upon that framework and allows agents to generate their own goals (see Colas et al., 2021b, for a recent review). Some notable examples include (Colas, Sigaud, and Oudeyer, 2018; Colas et al., 2019, 2020; Nair et al., 2018; Pong et al., 2020). While there is a great deal of variability among strategies for sampling interesting goals (see Section 2.1.2, Mechanisms/*Exploration Strategies*), all of these goal-oriented agents make decisions in a goal space rather than in an action space. Goal-oriented intrinsically motivated learning has a number of advantages compared to action-oriented intrinsically-motivated learning: it improves the performance and convergence time when learning inverse models in high-dimensional spaces with highly non-linear mappings (Baranes, 2013), it automatically generates learning curricula from easy to more complex skills (Moulin-Frier, Nguyen, and Oudeyer, 2014); finally, it enables hindsight learning

(Andrychowicz et al., 2018; Colas et al., 2019; Forestier et al., 2020). We discuss these positive practical implications in more detail in Section 2.2, FUNCTIONS.

2.1.1.3 *Exploring by Choosing Social Partners*

A few contributions have explored how IM can be coupled with social interaction. The SGIM-ACTS architecture (Nguyen and Oudeyer, 2012) considers an IM agent that is able to choose whether and when to learn from a social peer or by autonomous goal generation, from which social peer to learn, and what to ask the chosen social peer. Interacting with the social peer becomes part of the choice space where the agent makes decisions hierarchically: it first decides to interact or to self-explore its own goals, then which social peer or self-generated goal to focus on. The SGIM-ACTS framework was also applied to agents equipped with a realistic computer model of the human vocal tract and was able to reproduce the main developmental stages of infant vocal development (Moulin-Frier, Nguyen, and Oudeyer, 2014).

More recent contributions consider social influence as intrinsic motivation for achieving coordination and communication in multi-agent RL (Jaques et al., 2019). Other work in multi-agent RL have theorized and demonstrated how competition and cooperation display intrinsic dynamics, resulting in a naturally emergent curriculum (Baker et al., 2020; Leibo et al., 2019).

In humans, not only does the choice of a social partner can be intrinsically motivated, but the social partner can also influence intrinsic motivation. As humans often model the behavior of social peers, curiosity of a social partner can "spread" onto the learner, making them more likely to seek for information (Gordon, Breazeal, and Engel, 2015). This interplay between the curiosity of the agent and the behavior of a curious social partner has not yet been modeled in artificial systems. However, it can be a part of the mechanism which enables agents to modulate their own exploratory behavior, not by sampling instructions or directions from their peers, but by inferring mimicking their motivations.

2.1.1.4 *Representing Choice Spaces*

Precisely what do choices in choice spaces represent? In the case of action-space exploration, actions can represent micro-actions responsible for transitions between temporally adjacent momentary states, like pixel images (e.g., Bellemare et al., 2016; Pathak et al., 2017). In this case, the agent can only learn from short-term transitions which limits their ability to efficiently learn regularities spanning larger time scales[2]. However, the agent can sample learning situations by making

2 Model-free RL agents (e.g., Bellemare et al., 2016; Pathak et al., 2017) do not learn from the transitions per se – they learn from rewards. Specifically, the agent usually

decisions in the space of macro-actions (e.g., action-policies), rather than micro-actions, which enables exploration of more temporally extended effects of its actions (see Baranes, 2013).

In the case of goal sampling, the choice space can correspond to the entirety of the state space in which goals correspond to any representable state. This can be problematic for very high-dimensional spaces (see Kovač, Laversanne-Finot, and Oudeyer, 2020). Consider an example where the state space is a space of pixel images. Most (random) samples from such a space are white-noise images (Nair et al., 2018). Analogously, blindly picking letters from the alphabet will produce mostly gibberish. One solution to this problem is to define the goal space as some high-level feature space (also called latent or embedding space) of the raw-image space (e.g., Laversanne-Finot, Péré, and Oudeyer, 2021). Following the letter-picking analogy, this would correspond to composing and sampling from a higher-level space of syllables, words, or phrases, or sentences, which is more likely to produce meaningful strings. Thus, goal spaces can be further differentiated according to their modality (e.g., images, sounds, proprioceptive states) and internal organization (low-level or abstract states). When the goal space is defined over some abstraction over the raw sensory experience (e.g., 2D positions of objects vs raw pixel images) further design choices become relevant. One needs to consider whether to assume that this abstract space is given to the agent (e.g., Forestier et al., 2020), or whether the agent needs to learn a latent goal space from scratch (e.g., Laversanne-Finot, Péré, and Oudeyer, 2021; Nair et al., 2018). Note that learning of a latent goal-space adds another level of complexity to the autonomous learning process.

2.1.2 *Exploration Strategies*

Given a well-defined choice space for exploration, what strategies can an artificial agent follow to decide which learning situations are more or less interesting? Comprehensive reviews of different approaches can be found elsewhere (Aubret, Matignon, and Hassas, 2019; Linke et al., 2020; Mirolli and Baldassarre, 2013; Oudeyer and Kaplan, 2007), so we only provide a short survey of different approaches with a focus on their diversity rather than precise implementations.

We group existing approaches into three main categories: undirected, knowledge-based, and competence-based exploration. While all learning situations are equally interesting in undirected exploration, directed exploration strategies scale interestingness with the agent's abilities. Directed exploration can be divided into two broad classes:

learns a value function V, mapping states to their expected cumulative reward. Even if the transitions are momentary, the agent can still maximize their long-term cumulative reward using techniques like bootstrapping (Sutton and Barto, 2018). Another exception is when the world model is represented by a recurrent neural network (RNN; Takahashi et al., 2017) as RNNs can encode time series.

knowledge-based and competence-based strategies (Oudeyer and Kaplan, 2007). Sometimes the distinction between the two can be subtle because knowledgeable systems can also be competent and competent systems can be knowledgeable (Mirolli and Baldassarre, 2013). The point of divergence for these families of mechanisms is that to a knowledge-based system, the interestingness of a learning situation is determined by its relation to the system's knowledge. On the other hand, to a competence-based system, a given learning situation may be more or less interesting because it relates to the system's ability to reach a specific self-generated goal.

2.1.2.1 *Undirected Exploration*

Undirected exploration (sometimes random or uniform exploration) refers to a strategy that assigns interest uniformly across the choice space (making all learning situations equally interesting). The effectiveness of this simple strategy is inconsistent across different settings. When applied to action-space exploration, for example, undirected exploration is only effective for simplistic problems (Baranes, 2013; Benureau and Oudeyer, 2016), for example, when the environment provides dense rewards or when actions have simple and consistent effects. In more challenging settings, where the mapping between actions and their effects exhibits a combination of non-linearity, stochasticity, and redundancy, motor exploration is not sufficient for effective learning, but goal exploration could be (Moulin-Frier and Oudeyer, 2013). This is largely because learning how to reach goals contributes to the agent's competence and thus its ability to control the environment, while learning about various outcomes of all of one's actions in all possible contexts may be worthless in practice (Mirolli and Baldassarre, 2013). Besides, trying out random actions from a particular state may be futile for reaching certain hard-to-reach regions of the state space. Think of how hard it would be to learn how to drive from home to work by performing random actions (you would end up crashing your car most of the time). If instead you were to learn how to drive to various places from home (your driveway, a corner shop, a nearby gas station) your chances of finding your way to your office eventually would be much higher.

Despite its simplicity, random goal exploration has proven to be surprisingly efficient, leading to some forms of novelty search as an emergent feature and surpassing directed approaches operating in the action space in the learning of redundant inverse mappings (Benureau and Oudeyer, 2016; Colas, Sigaud, and Oudeyer, 2018). Because it is simple and computationally cheap, random goal exploration is often combined with other strategies, either to jumpstart the primary mode of directed exploration by collecting initial data, or sometimes as a complementary strategy at a certain level of hierarchical sampling decisions in modular spaces (e.g., Forestier et al., 2020), and sometimes

as an epsilon-greedy strategy (see Sutton and Barto, 2018) to balance between random and directed exploration (e.g., Colas et al., 2019). Still, in many situations, undirected exploration may not be as efficient or effective as more sophisticated guided exploration approaches. For example, random (goal) exploration performs poorly when the space of effects has a hierarchical structure, so that certain states are only accessible through reaching some prerequisite states (e.g., Forestier et al., 2020). An agent exploring goals randomly is unlikely to ever get to practice these "out-of-reach" goals, unless there is a mechanism for imagining them that leverages structured representations of goals, such as natural language encodings (e.g., Colas et al., 2020).

2.1.2.2 *Knowledge-Based Exploration*

Knowledge-based exploration is perhaps the most diverse family of intrinsically motivated strategies. Not only are there many ways in which one can characterize a relation between learning situations and the agent's knowledge, but there are many kinds of knowledge that agents represent. For example, an agent can maintain a predictive causal model of the effects of its actions and have a metacognitive monitoring system track errors that this predictive model commits. Equipped with such a system, the agent can measure the interestingness of actions based on, for example, outcome prediction error of the forward model (Gordon and Ahissar, 2012; Saegusa et al., 2009). Specifically, the agent can be more (or less) interested in taking actions for which the forward model does not accurately predict the consequences (known as prediction-error strategy). Alternatively, the agent can track changes in prediction accuracy (a strategy known as learning progress; e.g. Kim et al., 2020; Oudeyer, Kaplan, and Hafner, 2007; Schmidhuber, 1991b). Here the agent would be more interested in actions, for which the predictions get more accurate with time. While the measure of interestingness in these approaches is based on the predictions from a forward model, exploring agents can also monitor the behavior of an inverse dynamics model (Haber et al., 2018; Pathak et al., 2017).

Following the temporal derivative of prediction error, rather than its instantaneous values, protects the agent from spending time on uncorrelated regions of the sensorimotor space. To explain, agents that are motivated to take actions with unpredictable consequences can get stuck in a situation where action predictability does not improve, e.g., trying to predict the next image on a TV-screen showing white noise. Agents sensitive to prediction-error dynamics, on the other hand, will eventually associate the "noisy TV" with low LP and lose interest in the futile activity. An alternative clever solution to the "noisy TV" problem is to explore in a choice space is to focus on prediction errors that are due to the agent's actions rather than the stochasticity inherent in the environment. Pathak et al. (2017) achieved this by using a space

of latent features encoding parts of the sensory space that the agent can control. The space was obtained by learning to predict actions from temporally adjacent states (an inference known as the inverse-dynamics problem). The correct action can only be reliably inferred from state features that actually change due to that action. Thus, if the agent is set up to be curious about actions that result in erroneous predictions of the next latent feature (instead of the next raw state), it will tend to reproduce unpredictable but controllable actions. This approach provides a remedy for the noisy-TV problem, but does not escape a other perils of unpredictability. For example, if the agent can control the TV by switching what appears on the screen but has no control over what the new frame shows, it will be distracted by this "unpredictable-TV" (Burda et al., 2018). LP based solutions are generally better at protecting agents against different "unpredictability traps" (Kim et al., 2020; Kovač, Laversanne-Finot, and Oudeyer, 2020).

A distinct subclass of knowledge-based strategies relies on knowledge about frequencies of observed states. In so-called count-based approaches (Bellemare et al., 2016; Tang et al., 2017), this knowledge is represented explicitly, allowing the agent to selectively explore over- or under-visited states. Other systems in this subclass of approaches incorporate different variants of autoencoder networks to learn latent spaces (Bougie and Ichise, 2020a; Twomey and Westermann, 2018) or generative models (Nair et al., 2018; Pong et al., 2020) of states observed by the agent. Specifically, Twomey and Westermann (2018) defined several interestingness measures based on backpropagation computation of their category-learning neural network, including measures of weight update, prediction error, and activation-function derivative (which model, respectively, the system's curiosity, novelty, and plasticity). Bougie and Ichise (2020) introduced auxiliary tasks of image reconstruction with context-based autoencoders and defined an intrinsic reward measure derived from reconstruction errors from these tasks. Others employed variational autoencoders to estimate valid-state distributions in order to guide exploration (Nair et al., 2018; Pong et al., 2020). Although these models do not explicitly encode state visitation counts, the interestingness measures defined on their basis are related to frequencies of the observed states. Autoencoder prediction error, for instance, should decrease with repeated exposure to a given state. Variational autoencoders additionally represent the statistics of the latent space – a feature that can be used to estimate the likelihood of any state (including completely novel states) given what has been observed in the past.

2.1.2.3 *Competence-Based Exploration*

Another family of strategies assigns interestingness based on competence. These approaches do not need to assume any explicit world-dynamics knowledge model (although they can), so they can be readily

incorporated in model-free learning systems without the need to introduce any auxiliary tasks. Competence-based exploration strategies are especially well-suited for intrinsically motivated agents that explore self-imposed goals, since the notion of competence corresponds naturally to goal achievement. A simple competence-based heuristic in a goal-oriented context is to sample goals according to the agent's ability to achieve them. For example, Bougie and Ichise reward the agent in a given state based on how incompetent the agent perceives itself to be in that state (Bougie and Ichise, 2019). Here the agent generates its data set by taking interest in actions that lead it to states at which it deems itself incompetent, i.e. by choosing actions that maximize incompetence. In a different approach, Florensa and colleagues leveraged the generative power of adversarial networks for generating goals, which the agent evaluates based on the probability of reaching them (Florensa et al., 2018). Thanks to this evaluation, their agent can prioritize goals of intermediate difficulty, thereby avoiding goals that it already knows how to achieve and goals that it knows it cannot achieve. Santucci et al. compared the efficacy of several interestingness measures for autonomous mastering a set of tasks that included unreachable distractor tasks. The best-performing measure that allowed their agent to learn all learnable tasks in the least amount of time was based on competence prediction-error (Santucci, Baldassarre, and Mirolli, 2013). Several other teams have explored modular goal spaces using measures of competence progress (Colas et al., 2019; Forestier et al., 2020; Stout and Barto, 2010). Agents in these studies were incentivized to sample goals from the predefined regions (modules) of the goal space where competence was either improving or deteriorating. Oudeyer and colleagues used a similar competence progress-based strategy but in settings where the singular goal-space was progressively modularized (Baranes, 2013; Moulin-Frier, Nguyen, and Oudeyer, 2014).

2.2 FUNCTIONS

The previous section reviewed the diversity of approaches to specifying intrinsically-motivated mechanisms in AI. We have mentioned in the introduction that these mechanisms are useful for autonomous learning in the absence of externally assigned objectives, but what specific functional consequences do different intrinsic motivation mechanisms confer? This section focuses on the functional aspects of intrinsically motivated systems addressing the question of how different intrinsic-motivational systems are useful for both artificial and biological agents.

Singh et al. provide a useful connection between the concepts of extrinsic and intrinsic motivation and the concepts of primary and secondary rewards (Singh et al., 2010). While the reception of pri-

mary rewards (e.g., related to nutrients, sex, or pain) contributes to an agent's survival and, thus, its reproductive success, secondary reinforcers signal anticipation of primary rewards, but are neutral a priori and have to be learned. In RL, outputs of the reward function can be analogous to primary reward signals, because the reward function is given to the agent. However, value functions evaluate states based on their learned expected cumulative reward, and therefore outputs of the value function are analogous to secondary reinforcers. In this formulation, intrinsic motivators (e.g., novelty, uncertainty, learning progress etc.) are primary reinforcers because their "rewardingness" is a given. On the other hand, predictors of intrinsic primary reinforcers can acquire rewarding qualities through the learning of secondary reinforcers, in the same way as non-rewarding states gain value due to extrinsic reinforcers. Therefore, intrinsic primary rewards differ from extrinsic primary rewards, mostly due to their causal proximity to evolutionary success. As it is often the case in biology (Dobzhansky, 1973), it is sensible to analyze the functional aspect of artificial intrinsic-motivational mechanisms in light of evolution – an exercise that allows us to consider the possible bio-ecological roles of engineered curiosity-driven systems.

In what follows, we build upon Singh et al.'s framework as well as the taxonomy of mechanisms we proposed in the previous section in order to extract key functional aspects of intrinsic motivation in both biological and artificial agents. We start from the more evolutionarily proximal functions (direct procurement of primary rewards) and gradually consider increasingly distal ones (learning of internal models, goal discovery, and cultural innovation).

PROCUREMENT OF EXTRINSIC PRIMARY REWARDS In relatively dense primary-reward environments, unstructured random exploration in the action space is sufficient to efficiently elicit extrinsic primary rewards (as it is the case in some standard AI benchmarks based on video games, e.g., Mnih et al., 2015). This points to the most proximal function of intrinsic motivation: the generation of diverse sensorimotor experiences. The direct benefit of generating diverse experiences is to increase the probability of eliciting primary extrinsic rewards. However, the diversity generated by random exploration is usually not sufficient in sparser reward environments (e.g., Pathak et al., 2017). Knowledge-based exploration can increase the diversity of learning experiences by guiding exploration based on different measures of interestingness. Discovering extrinsic primary rewards such as food, water, or shelter through exploration is clearly linked to the agent's well-being.

LEARNING INTERNAL MODELS A more distal function of intrinsic motivation is the learning of internal models of the agent-environment

interaction (e.g. forward or inverse models) that can enhance the agent's decision-making abilities. Learning such a model can be formalized as an UL or an SL problem and strongly relies on the information contained in the training dataset. Autonomous agents generate this dataset by interacting with the environment, and the key role of intrinsic motivation is to generate informative training data. Research in intrinsically motivated SL extensively studied two main cases. On the one hand, knowledge-based approaches have proven to be efficient in generating informative data for learning forward models (Oudeyer, Kaplan, and Hafner, 2007). On the other hand, competence-based approaches have proven to be more efficient than knowledge-based approaches for learning inverse models (Baranes, 2013). RL agents can also benefit from internal models, be it in the form of a value function and an action policy in model-free RL (Pathak et al., 2017), or in the form of a world-dynamics model (forward or inverse) in model-based RL (Haber et al., 2018). Understanding how the world works per se does not put proverbial food in the agent's mouth, but it allows the agent to act more intelligently in novel situations and plan ahead in order to obtain what it needs more reliably.

GOAL DISCOVERY The third level of functions we propose is related to the discovery and learning of novel goals and the associated skills to achieve them. This is the main function of competence-based approaches, where exploration is guided by the pursuit of self-imposed goals. These approaches can automatically organize exploration from simple to more complex skills (Forestier et al., 2020; Gordon and Ahissar, 2012; Pong et al., 2020), as well as discover the full range of the achievable behavioral repertoire, possibly in an open-ended manner (Colas et al., 2019). There are multiple functional advantages to discovering and mastering novel goals that are not extrinsically rewarding. First, in environments where eliciting extrinsic primary rewards require the acquisition of complex skills (e.g., hunting), it is crucial to structure learning in a curriculum from simple (e.g., locomotion) to more complex skills (shooting a projectile). Complex skill sets often display a hierarchical structure, where mastering easier ones is prerequisite for acquiring more complex ones. Second, the ability to autonomously explore and discover new goals and skills provides a crucial advantage in changing environments. For example, paleoclimatological data provides evidence for strongly varying climate conditions in the Rift Valley of East Africa, approximately 7 million years ago, and it is hypothesized that the ability to rapidly and flexibly reorganize a diverse behavioral repertoire was a key requirement to adapt to such unprecedented conditions (Potts, 2013). Thus, the ability to autonomously generate and master novel goals and thereby acquire a diverse repertoire of complex skills provides a crucial advantage for a species' success in such settings of strong environmental variability

(see Nisioti and Moulin-Frier, 2020, for a recent proposition to apply this principle in AI).

CULTURAL INNOVATION Finally, the fourth and most distal level of functions we propose concerns cultural innovation. Several theoretical contributions proposed a potential role of curiosity-driven exploration in both language acquisition (Oller, 2000) and evolution (Oudeyer and Smith, 2016). From a sensorimotor perspective, active exploration can spontaneously generate diverse behaviors from modality-independent and task-independent internal drives. Such spontaneous behavior can result in vocal activity that may have bootstrapped the emergence of communication. This hypothesis is supported by computational simulations showing a role of curiosity-driven exploration in vocal development (Moulin-Frier, Nguyen, and Oudeyer, 2014), social affordance discovery (Oudeyer and Kaplan, 2006), and active control of the emerging conventions in social lexicon (Schueller, Loreto, and Oudeyer, 2018). From a cognitive perspective, compositional language itself is a powerful cognitive tool for imagining novel out-of-distribution goals in competence-based intrinsic motivation (Colas et al., 2020). Moreover, recent contributions in multiagent RL have shown how an auto-curriculum of increasingly complex behaviors displaying features of open-ended innovation can emerge from agents' co-adaptation in mixed cooperative-competitive environments (Baker et al., 2020). Such mechanisms are potential precursors of cultural evolution in the human species. Cultural evolution has triggered increasingly complex technological innovation across generations (Fogarty and Creanza, 2017). A prime example of this is the industrial revolution, which has resulted in a rapid acceleration of the global population growth in the 19th century (Lucas, 2004).

2.3 USAGES

Other than helping artificial agents explore learning situations in abstract task-independent contexts, how can intrinsically-motivated learning algorithms be used? We identify two main directions in which such algorithms can have high impact. On the one hand, they can be a great tool for advancing research in cognitive psychology and neuroscience. Outside cognitive research, these algorithms can be applied directly to problems that require intelligent automated exploration. We review both of these domains of application in the rest of this section.

2.3.1 *Cognitive Modeling*

The notion of intrinsically-motivated exploration in psychology has been developing – for the most part – independently of AI (Ka-

plan and Oudeyer, 2007). Psychological investigations can be traced back to psycho-physiological perspectives on exploration in the early 1940s (Hull, 1943). Since then, psychological views on curiosity and information-seeking have undergone multiple changes (Bazhydai, Twomey, and Westermann, 2020; Loewenstein, 1994) and are now becoming more integrated with the computational perspective (Gottlieb et al., 2013; Kaplan and Oudeyer, 2007). On the other hand, early formulations of intrinsic motivation in AI were either "discovered by accident" (Andreae, 1977, p. 5) or influenced by relatively distant research areas, like biological autopoiesis (Maturana and Varela, 1980) and aesthetic information theory (Nake, 1974, as cited in Schmidhuber, 1991a), yet not the aforementioned psychological literature (see Kaplan and Oudeyer, 2007, for a historical overview). Over the course of history, psychology and AI have actually been converging on similar ideas for why certain behaviors could be intrinsically rewarding: due to some kind of mismatch between bottom-up observations and top-down predictions (Kaplan and Oudeyer, 2007).

Evolutionary implications of artificial intrinsic-motivational systems discussed earlier (Section 2.2, FUNCTIONS) raise the need to seriously consider them as candidate models for human non-instrumental learning. A major advantage that comes naturally with these systems is their precise formulation. Such a formal description unambiguously discloses crucial structural and functional properties of the system in question, and thus enables to advance the related theory more efficiently (McClelland, 2009).

Cognitive models based on exploring artificial agents are becoming increasingly frequent. For instance, Moulin-Frier, Nguyen, and Oudeyer (2014) explained the progression of human vocal behavior through distinct developmental stages as an intrinsically motivated, goal-exploration process based on competence-progress motivation. Such computational accounts of curiosity-driven learning have led to novel hypotheses about the mechanisms of intrinsic motivation in humans (Kaplan and Oudeyer, 2007) and the role of curiosity in the evolution of language (see Oudeyer and Smith, 2016). Gordon and colleagues implemented an intrinsically motivated RL agent to account for multiple features of the development of exploratory whisking behavior in rats (Gordon, Fonio, and Ahissar, 2014; Gordon and Ahissar, 2012). More recently, Twomey and Westermann (2018) used an actively exploring autoencoder network to hypothesize an algorithmic-level description of visual exploration in infants. Poli et al. (2020) compared several knowledge-based sampling strategies to predict visual-attention control in infants. Moreover, computational models of intrinsic motivation are invoked to explain self-determined instrumental (Gershman, 2018b) and non-instrumental (Ten et al., 2021) choices in human adults.

This chapter relates several mechanistic implementations of curiosity-driven exploratory systems to a common underlying structure. This can be beneficial for revealing potential theoretical and empirical gaps in the scientific understanding of human information-seeking. For instance, while information-seeking literature is brimming with work investigating episodic curiosity-driven sampling over short time scales corresponding to exploration by sampling micro-actions, fewer studies have looked at time-extended exploration of learning activities. Motivation to engage in time-extended learning activities cannot be easily reduced to unpredictability or information-gains conferred by the decisions to pursue them. For example, taking a math course is difficult to explain in terms of the uncertainty or information-gain expected from the act of enrolling into a course or attending a class. Some of the long term engagement in learning activities can be at least partially explained by goal pursuit (e.g., wanting to be good at math), but often times such activities are engaged because of the inherent fun or enjoyment expected from them.

Take a moment to appreciate the semantic distinction between "action" and "activity".

Our taxonomy identifies a salient dimension of variability among computational models of curiosity-driven exploration: the choice space through which agents sample learning situations. This aspect of exploratory mechanisms is not discussed explicitly in psychology. What choice spaces are used by computational *cognitive* models of human exploration? The dominant formal framework of human inquiry – the so-called Optimal Experiment Design (OED, Coenen, Nelson, and Gureckis, 2019) – concerns situations in which humans evaluate a set of potential queries (i.e., actions) to decide which query to execute. In comparison, hardly any studies investigate intrinsically motivated exploration of goals. Specifically, how humans generate and choose new goals in unfamiliar environments? In that regard, the IMGEP (Forestier et al., 2020) framework offers means to model goal-based exploration in humans.

2.3.2 *Practical Applications*

Functional diversity of intrinsic-motivational mechanisms makes them useful for practical applications, such as automated knowledge discovery and education. Mechanisms of intrinsic motivation help autonomous agents learn in settings with complex sensorimotor spaces and sparse or non-existent rewards. They are crucial for building autonomous control systems that can learn efficiently in open-ended environments. The practical effectiveness of these mechanisms has been recently demonstrated in studies of automated discovery in complex systems, where curious agents learn to control diverse effects in complex non-linear settings, such as smartphone applications (Pan et al., 2020), continuous cellular automata (Etcheverry, Moulin-Frier, and Oudeyer, 2021), and real-world chemical systems (Grizou et al., 2020).

Automated discovery can have a high societal impact by assisting both scientific research and artistic creation. In another line of work, artificial intrinsic motivation has been used to foster generalization in deep RL agents by guiding them through automatic-curriculum learning (Portelas et al., 2020b). Moreover, since one of the functions of curiosity-driven systems is knowledge acquisition, directed exploration strategies of such systems can be used to assist learners when they behave suboptimally. Importantly, such intelligent tutoring systems can be tailored to the current levels of knowledge or competence of individual learners (Clément, Oudeyer, and Lopes, 2016) and have shown promising results in pedagogical settings (Clément et al., 2015; Delmas et al., 2018), where they assist learners in selecting topics and exercises that maximize their individual learning progress.

2.4 CONCLUSION

The goal of this chapter was to familiarize the reader with the diversity of computational mechanisms and possible evolutionary functions of curiosity-driven exploration. We identified an important problem facing autonomous agents that have control over their learning experiences. Specifically, such agents must decide how to sample learning data in the absence of externally imposed tasks. We briefly reviewed several ways in which artificial agents choose actions, goals, or social peers in order to engage in learning situations. We presented distinct families of exploration strategies, including undirected, knowledge-based, competence-based and socially-influenced approaches. We then discussed how these mechanisms can contribute to evolutionary success at different levels: by helping agents to approach primary rewards, acquire world models, discover goals, and bootstrap a cultural repertoire. Finally, we provided some contemporary examples of how intrinsic motivation algorithms are used in practical applications as well as in cognitive research.

We hope that this concise bird's-eye perspective – organized along the proposed mechanistic, functional, and pragmatic dimensions of curiosity-driven exploration – can serve as a stepping stone towards a unified taxonomy of this fascinating and important field. Recently, Gordon (2020) proposed a related framework that organizes different curiosity-driven artificial systems along hierarchical levels of cognitive development. In this framework, curious artificial agents can be understood as instantiations of a so-called "curiosity loop" (Gordon and Ahissar, 2012) consisting of an embodied learner, an action/goal selection mechanism, and an intrinsic reward. Thus, specific curiosity loops operate at different levels of an autonomous-learning hierarchy, where each level is associated with a specific function, including exploration of the self, exploration of the environment, object interaction, and social interaction.

We believe that our taxonomical organization and the "curiosity loops" framework are highly complementary. For instance, our taxonomy highlights the parameterization of choice spaces for decision-making (the action selection component in a curiosity loop), pointing out not only the diversity of mechanisms, but also the common underlying structure. At the same time, the consideration of the learner component's learning problem in Gordon's (2020) framework enables to draw more fine-grained functional distinctions between different agents. Specifically, at different points of the developmental trajectory, the learner might prioritize learning different aspects of the world structure, which leads to the acquisition of increasingly complex world models, from self-models, to models of objects, to models of other sentient beings.

3

ELEMENTS OF INFORMATION-SEEKING

Where is the wisdom we have lost in knowledge?
Where is the knowledge we have lost in information?
— T. S. Eliot (2014)

The above quote from T.S. Eliot is one of the earliest sources that inspire the famous Data≠Information≠Knowledge≠Wisdom model (DIKW) – a framework that draws distinctions and defines relationships between the concepts of data, information, knowledge, and wisdom (Sharma, 2008). While the DIKW model is tossed around in the context of business administration, distinguishing between information and knowledge is also important for the scientific study of information-seeking. This chapter reviews the current literature (mostly) about non-instrumental information-seeking and presents a theoretical perspective, arguing that what makes information inherently valuable for curious beings like humans is its implications for knowledge (and perhaps, wisdom).

3.1 INFORMATION AND INFORMATION-SEEKING

Sense organs are ubiquitous in the living world. Indeed, they are so useful that not only animals but plants, fungi, and many other organisms have evolved them (Bourret, 2006; Braunsdorf, Mailänder-Sánchez, and Schaller, 2016; Schwab, 2018; Trewavas, 2005). Moreover, humans endow their tools with artificial sensors to make them more responsive and thus "smarter". It is no mystery why sensors are so powerful: they *inform* their owners about what is "out there" so that they can adapt their behavior accordingly. Information, then, is the basis of intelligent behavior. It is also a central concept in this thesis, so we need to establish what it means more precisely.

Information is an important concept across many disciplines. Even though precise definitions differ, it is possible to identify three major conceptual stances on what information means (Adriaans and van Benthem, 2008). According to one perspective, information refers to declarative descriptions of the mentally represented world, which can be obtained, for example, through empirical observation, linguistic communication, or "armchair" deduction. This is the sense in which the word 'information' tends to be used in lay conversation. Another view characterizes information in terms of uncertainty. Here, information is viewed as an abstract communication process by which uncertainty about some random event can be reduced. This formulation is commonly adopted in Information Theory (Shannon, 1948).

Finally, there is an approach that treats information as the complexity of the simplest possible representation of an object in a given (i.e., fixed) system. This is a "Kolmogorov-complexity" stance on information (Kolmogorov, 1965). It underlies the intuition that simple objects require less information to be described in, say, natural language or neural code, compared to complex objects.

While these three perspectives may seem quite far apart, they are demonstrably and rather intricately connected (Adriaans and van Benthem, 2008). This is particularly clear when we consider how these stances converge within a single information-processing sequence of events, such as sensing. Through senses and neural substrates, organisms can systematically represent entities in the environment (stance 3) thereby reducing uncertainty about presence or absence of particular stimuli (stance 2) and *forming* an *in*ternal description of the surrounding world (stance 1). In other words, sensing is a process of representing, communicating, and interpreting information[1]

Since the world is so incredibly complex, there is probably more potentially observable and thinkable information in it then organisms can possibly represent and process (Kolmogorov, 1965). Given the limited computational resources, the overabundance of potential information in the world means that it needs to be somehow funneled for senses to be useful (Gottlieb and Oudeyer, 2018). Mechanical features of sense organs responsible for domain specialization (e.g., light, sound, pressure) and sensitivity to specific intensity ranges within domains (e.g., Schwab, 2018) can be viewed as passive, structural funnels of information. Another, active kind of funneling is enabled by the ability to selectively expose one's external and internal sensors to specific stimulation via actions such as movement (e.g. Gottlieb et al., 2013) and neuromodulation (e.g., Yu and Dayan, 2005). This *active* sensing behavior can be characterized as *information-seeking*. How do organisms control their behavior to expose themselves to "good" information?

3.2 VALUE OF INFORMATION

Information is beneficial only insofar as it communicates what the organisms should care about. Thus, information-seeking must be deployed strategically and ultimately optimize the biological fitness of a species. Individual organisms, of course, do not know how to optimize this global fitness function directly (Berge and van Hezewijk, 1999; Gottlieb et al., 2013; Singh et al., 2010). Instead, phenotypes implementing biologically advantageous behavioral tendencies emerge through evolution and ontogenetic development. Information-seeking

1 Mnemonic retrieval of information can be viewed as a kind of "internal sensing" by which one system "observes" and/or encodes information from another, all within the same brain.

is likely to be one such tendency, but it is not necessarily obvious how information that organisms seek contributes to the biological fitness of their species.

It is easy to see why information-seeking is biologically advantageous in *instrumental* contexts. Instrumental information-seeking includes behaviors driven by *extrinsically valuable* states that cannot be achieved by sensory stimulation alone. For example, although foraging for food (or tracking a predator) requires seeking information, the ulterior biologically rewarding end is the consumption of nutrients (or avoidance of predation) – something that cannot be achieved solely by observing or thinking about food (or predators). However, sometimes, information is sought in the apparent absence of an extrinsically valuable state. Valuable states that lack extrinsic value are *intrinsically* valuable by definition. Information-seeking that is driven by such states is called *non-instrumental*.

The most recent evidence for the intrinsic value of information comes from neuroscience. It has revealed common substrates for processing intrinsically and extrinsically rewarding stimuli. This work has relied on different variants of the "observing task", where a subject decision-maker can choose to observe or forego information about the outcome of a maximally uncertain gamble (reviewed in Cervera, Wang, and Hayden, 2020; Kidd and Hayden, 2015). One early study showed that consuming a water-reward and observing information about the upcoming water-reward is processed by the same structure in the macaque monkey brain (Bromberg-Martin and Hikosaka, 2009). Specifically, the well-studied "reward-signaling" dopaminergic (DA) neurons of the midbrain (Schultz, Dayan, and Montague, 1997) responded to more-than- and less-than-expected amounts of water in the same way they responded to information about the corresponding amounts of water (Bromberg-Martin and Hikosaka, 2009). Moreover, some neurons in the lateral habenula (LHb) region encoded information prediction errors (IPEs) similarly to how these neurons encoded prediction errors about extrinsic rewards: their activity was increased when less than expected information was promised to the subjects, and it decreased when they were promised more information than expected (Bromberg-Martin, 2011). In another study, Blanchard et al. found that distinct subpopulations of neurons in the orbitofrontal cortex (OFC) orthogonally encoded the potential amount of water-reward and the validity of the related cue, i.e., its "informativeness" with respect to an extrinsic reward (Blanchard, 2015). These results demonstrate that informative states (or stimuli) can be valued for their inherent property of reducing uncertainty, regardless of the extrinsic consequences.

The encoding of informativeness by the OFC is especially revealing in light of recent integrative accounts of this cortical structure. One contemporary view suggests that the OFC functions like a "map" (or, perhaps, more like a GPS tracker) that tells an organism where it is

positioned in the so-called *task space* (Wilson et al., 2014). Specifically, the OFC is proposed to combine multimodal sensory information to represent the organism's current state in relation to the goal(s) of the task at hand. The same sensory context can be used to generate model-based predictions of future states for which the OFC can represent their *implied* value (see Stalnaker, Cooch, and Schoenbaum, 2015, for details). Outside the narrow settings of specific laboratory tasks, the OFC might have an even more general function. It might provide a substrate for incorporating incoming sensory information into the process of arbitrating between competing high-level goals (Fine and Hayden, 2022) – goals that ultimately serve transient yet consistent biological and psychological needs (Juechems and Summerfield, 2019). According to Fine and Hayden (2022), the value of states encoded by the OFC might actually represent state proximity to the current goal. The authors propose that OFC is the apex of the goal-abstraction hierarchy that cascades down from abstract needs to high-level goals to low-level actions. At each level of this extended premotor hierarchy, the cortex implements a goal-selection policy that chooses lower-level goals/actions to optimize higher-level objectives, given the current context. These contemporary accounts have intriguing implications. First, they suggest that the valuation of informativeness at the level of the OFC (e.g. Blanchard, 2015) encodes proximity to fundamental needs, implying that being informed is a basic motive. Second, they suggest that states of being informed can enter the lower premotor levels as goals. That is, states of being informed can be pursued independently of extrinsically rewarding states. Finally, these considerations imply that being informed can compete with other fundamental goals such having food or staying warm.

One of the observing task's main takeaways is that the brain processes non-instrumental information similarly to how it deals with information about extrinsically valuable stimuli (e.g., as if it was experiencing water consumption). This suggests that information *about* an upcoming reward and information *from* the reward itself have similar roles. The most salient role of information *from* rewards is to drive learning (Schultz, 2016). That is, information generated from the reward consumption (e.g., the sweetness or bitterness of food) is used to adjust context-dependent predictions (e.g., the sight of food) and reinforce or suppress the preceding appetitive behavior. The study by Bromberg-Martin and Hikosaka (2009) suggests that the brain might similarly reinforce or suppress non-instrumental information-sampling behaviors and maintain a model for predicting when and how much of such information can be expected.

But how does the intrinsic value of information elevate biological fitness? For example, does the "observing task" demonstrate an evolutionary adaptive behavior? One possibility is that advance information about upcoming rewards is useful for arbitrating between alternative

goals that the organism might be pursuing, which is conducive to allostatic regulation – a process through which organisms prepare for future challenges before they arise (Fine and Hayden, 2022; Sterling, 2012). Concretely, inferring that the food is coming enables individuals to focus their mental activity on tasks other than foraging or prepare for the expected food delivery.

More fundamentally, information is useful because it contributes to generalizable declarative knowledge about the environment (including the organism's own body). One salient function of such knowledge is to enable effective planning. Planning can be understood as a hierarchical optimization of lower-level actions with respect to higher-level goals (Fine and Hayden, 2022). Simply put, planning is a process of breaking down a high-level objective (e.g., buy food) into progressively lower-level actions (e.g., arrive at supermarket → ... → drive to supermarket → ... → find keys → ...). Hierarchical breakdown of high-level objectives is a central principle in hierarchical reinforcement learning (Pateria et al., 2021) and sensorimotor control (Todorov and Jordan, 2002). Planning depends on the ability to anticipate future states, because at any given level of the hierarchy, the best action is determined by the corresponding sensory context (e.g., predicting that the shortest path will be jammed might cause you to take a longer path). The ability to accurately predict future states, and, therefore, plan, improves with knowledge. And since the same piece of knowledge can be used to plan out different goals, declarative knowledge accumulation via non-instrumental information-seeking is conducive to robust achievement of arbitrary goals.

Recall that "robust achievement of arbitrary goals" was the central problem facing intrinsically motivated artificial agents in Chapter 2

Whereas declarative knowledge accumulation is clearly beneficial, it is not the only adaptive functional consequence of the value of information. In fact, declarative knowledge accumulation alone does not guarantee the generation of adaptive behaviors (Mirolli and Baldassarre, 2013). The problem is that while all declarative knowledge is *potentially* useful, not all such knowledge is useful in practice. Several authors proposed that a major benefit of value of information is that it facilitates acquisition of skills (i.e., procedural knowledge; Gottlieb et al., 2013; Mirolli and Baldassarre, 2013). But how could non-instrumental information-seeking promote skill acquisition? To answer this question, it is helpful to think of information (abstractly) as uncertainty reduction. Under this perspective, learning a skill can be conceived as a process of reducing the uncertainty (i.e., seeking information) regarding which actions to take in order to achieve a goal (Gottlieb and Oudeyer, 2018). One might argue that the value of such information is instrumental towards achieving a certain goal, and thus does not demonstrate the usefulness of valuing of information as a terminal end. However, the subtle point is that *knowing* how to achieve a goal might be more important than achieving the goal itself, which is consistent with the idea that competence proper is a fundamental

psychological need in humans (Deci and Moller, 2005). Thus, non-instrumental information-seeking can be conducive to accumulation of diverse sets of skills, especially if organisms are motivated to playfully explore arbitrary goals (Chu and Schulz, 2020a).

The idea that non-instrumental information-seeking increases biological fitness through accumulation of knowledge (declarative and procedural) is intriguing because it provides a biological explanation for phenomena like curiosity, interest, and exploratory play (Chu and Schulz, 2020b; Gottlieb and Oudeyer, 2018; Murayama, FitzGibbon, and Sakaki, 2019). However, this idea only describes the phylogenetic mechanism by which organisms can evolve non-instrumental information-seeking behavior. It does not describe the actual features and computational principles enabling knowledge acquisition for the individual. That is, it does not tell us how non-instrumental information-seeking is initiated, sustained, and terminated. The rest of this thesis will revolve around the work – including the original contributions in Chapters 4 and 5 – aiming to elucidate *how* intrinsically motivated information-seeking operates to enrich knowledge in a useful way. We will begin by discussing a contemporary perspective on the situational determinants of curiosity and interest. We will then deliberate on how the affective and motivational aspects of curiosity/interest can reinforce information-seeking behaviors, thus setting the stage for the central hypothesis of this thesis – the idea that the motivation to pursue information comes from the expected gains in knowledge.

3.3 CURIOSITY AND INTEREST

The terms *curiosity* and *interest* lack universally agreed upon technical definitions (Dubey and Griffiths, 2020; Kidd and Hayden, 2015; Murayama, FitzGibbon, and Sakaki, 2019). It is still possible to delineate features of two distinct motivational states that the two labels map onto. For example, an aversive state experienced as deprivation and wanting to resolve one's salient awareness of ignorance can be contrasted with an appetitive state experienced as positive anticipation or the actual enjoyment of learning something new. It is tempting to call the former "curiosity" and the latter "interest", but let us refrain from committing to specific definitions and use these terms rather informally. Both curiosity and interest play important roles in the continuous process by which humans acquire knowledge (Murayama, FitzGibbon, and Sakaki, 2019). Although they coincide more often than not, motivational states differ in what triggers them, what neural and behavioral responses ensue, and how they are affectively experienced (Day, 1982; Grossnickle, 2016; Hidi and Renninger, 2019; Litman, 2019; Shin and Kim, 2019).

3.3.1 *Situational determinants*

Throughout history, researchers have proposed different explanations for what triggers the motivation to seek information in the absence of extrinsic value. These include conflict (Berlyne, 1954a), ambiguity (Ellsberg, 1961), incongruity/dissonance (Festinger, 1962; Hunt, 1960), knowledge-gap (Loewenstein, 1994), unpredictability (Shin and Kim, 2019), and more (see Bazhydai, Twomey, and Westermann, 2020; Loewenstein, 1994; Oudeyer and Kaplan, 2007, for a review). A common denominator for all of these proposals seems to be uncertainty. Indeed, uncertainty appears to be a necessary ingredient for sparking curiosity[2]. One reason for the diversity of propositions is that uncertainty comes in many "shapes" and "sizes" (Payzan-LeNestour and Bossaerts, 2011).

For instance, Yu and Dayan (2005); (2003) proposed a qualitative distinction between *expected* and *unexpected* uncertainty. According to their account, expected uncertainty arises when an agent holds ambivalent expectations about any particular future outcome (conflict in predictions) or holds ambivalent beliefs about potential causes of an observed outcome (ambiguity of explanation). Unexpected uncertainty arises when an agent observes an outcome or explanation that violates its learned expectations or beliefs. The authors speculated that expected and unexpected uncertainty are processed differently by the brain. Expected uncertainty is associated with the arousal of cholinergic activity resulting in elevated levels of acetylcholine (ACh), while unexpected uncertainty corresponds to the arousal of the noradrenergic system resulting in higher levels of norepinephrine (NE). As we discuss below, expected and unexpected uncertainties are transient states in the spiral of learning. A recent computational theory of executive function holds that expected uncertainty is also signaled by NE, which serves a dual function of invigorating effort and enhancing sensory processing (Silvetti et al., 2021).

The status of expected and unexpected uncertainty as triggers of curiosity is yet to be systematically investigated empirically, there are clues to suggest that both might be involved. Diminished curiosity was observed in patients with a probable early Alzheimer's disease (Daffner et al., 1992). These patients spent less time looking at curiosity-inducing stimuli compared to healthy controls. Alzheimer's disease is associated with severe damage to the cholinergic system (Ferreira-Vieira et al., 2016) involved in the processing of expected uncertainty (Yu and Dayan, 2005). On the other hand, activity in one of the main noradrenergic structures, the locus coeruleus (LC), correlates with heightened arousal, pupil dilation, and more efficient learning (Breton-Provencher, Drummond, and Sur, 2021) – prominent associates of curiosity. Moreover, LC interacts with the major DA structure in the

2 I hope you can agree that the idea of a curious omniscient being is oxymoronic.

ventral tegmental area (VTA) and can even disseminate dopamine in addition to NE across the brain (Ranjbar-Slamloo and Fazlali, 2020). As discussed above, dopamine is involved in reward anticipation and reward processing, and is a major component of the 'wanting' system in the brain (Berridge, 2007).

Pertinent to the current discussion is Shin and Kim's (Shin and Kim, 2019) distinction between two kinds of uncertainty: unpredictability and incongruity. Triggering the so-called *forward* curiosity, unpredictability corresponds to states of not knowing "when, where, or how an event has occurred or will occur" (Shin and Kim, 2019, p. 13). On the other hand, incongruity causes *backward* curiosity, and it arises in situations that violate expectations. It should be clear from this characterization that unpredictability and incongruity are similar (if not equivalent) to expected and unexpected uncertainty, respectively.

Whether a person experiences a state of forward or backward curiosity depends on whether information is anticipated "forwardly" or contemplated "backwardly". This implies that while unpredictability and incongruity arise in two different situations, they can fluidly morph into one another. Suppose I am scammed into buying a "loaded" coin that is falsely advertised to turn up heads 99.9% of the time. If I am indeed fooled, my expected uncertainty about any future toss of that coin is low. If I then observe 8 out of 10 unexpected tails, not only will I be surprised by this incongruent event, but I will also update my belief about the coin so that the expected uncertainty about future tosses will increase (I will appreciate their unpredictability). This example illustrates that the term "unexpected uncertainty" conflates two distinct concepts. The "unexpected" part refers to the surprise (see Barto, Mirolli, and Baldassarre, 2013) caused by an observation that is incongruent with expectation. Furthermore, the "uncertainty" part is rather ambiguous. It does not specify whether it refers to the expected uncertainty that follows a surprising event (i.e., expected uncertainty) or the uncertainty in knowledge that generates the expectation.

To avoid confusion, we will abandon this misleading expected-unexpected dichotomy. To adopt a less ambiguous terminology, we first need to introduce another prominent distinction that differentiates between *aleatoric* and *epistemic* uncertainty (Hüllermeier and Waegeman, 2021). The Latin word "aleatory" refers to gambling activity (especially dice playing), so aleatoric uncertainty refers to the knowingly irreducible uncertainty *of* an expected event (e.g., expecting randomness of a die roll). Epistemic uncertainty, on the other hand, refers to the uncertainty *in* expectation, as opposed to uncertainty *of* expectation. For a Bayesian inference aficionado, the distinction is easy to appreciate by thinking about the uncertainty expected in the data by a given generative model (aleatoric) and the uncertainty of the prior/posterior distribution of model parameters (epistemic). "Aleatoric uncertainty" corresponds completely to the previously introduced

"unpredictability" and "expected uncertainty". Epistemic uncertainty, by contrast, refers unambiguously to the uncertainty in knowledge. In sum, we will use the following terminology:

- **Unpredictability** (or **aleatoric/expected uncertainty**; also **ambiguity**) will refer to uncertainty about a future event (e.g., not knowing how a coin will land) or past event (e.g., not knowing who dropped a coin)
- **Incongruity/surprise** will refer to an event of observing something that violates expectation (e.g., observing 20 heads in a row using an ordinary coin)
- **Epistemic uncertainty** will refer to uncertainty about an expectation itself (e.g., being unsure if a coin is fair)

We believe that epistemic uncertainty is not given as much attention in curiosity research as aleatoric uncertainty and surprise. We will come back to this point later when we discuss the idea that uncertainty reduction reinforces uncertainty-seeking behaviors, thereby directing learners towards states in which their knowledge can improve.

As Shin and Kim (2019) note, unpredictability and incongruity have different relationships with curiosity. Unpredictability has an inverted U-shape relationship with (forward) curiosity (Berlyne, 1954a; Day, 1982; Loewenstein, 1994), while incongruity has a positive monotonic relationship with (backward) curiosity (Horstmann, 2015). The monotonic relationship between incongruity/surprise and curiosity is easier to understand. Surprise signals the inadequacy of one's knowledge, so it makes sense for organisms to minimize it, for example, by being curious and seeking information (Schwartenbeck et al., 2019). This is in line with Friston's (2009) free-energy principle – a mathematical theory of the sustainability of life itself. Surprise minimization is a biological implementation of this principle. Several empirical studies have demonstrated the positive relationship between surprise and curiosity using a range of behavioral paradigms (Berlyne, 1954b; Itti and Baldi, 2009; Poli et al., 2020).

The prediction of an inverted U-shape relationship between unpredictability/uncertainty and forward curiosity is less intuitive. In a nutshell, it maintains that prior knowledge determines the subjective intensity of curious states in a nonlinear way: one is predicted to be most motivated to obtain information if one has some incomplete idea as to what this information might be (intermediate uncertainty); conversely, curiosity is predicted to be low when one already has all the relevant knowledge (low uncertainty) or when one has too little knowledge (high uncertainty). The prediction that perfect knowledge should spark no curiosity is trivial, but it is not obvious why very poor knowledge is unlikely to result in curiosity. Yet, there is considerable evidence supporting this prediction (Baranes, Oudeyer, and Gottlieb, 2015; Berlyne, 1954b; Day et al., 1972; Kang et al., 2009; Loewenstein, 1994).

While researchers seem to agree on the general prediction, the proposed mechanisms may vary considerably. For example, Loewenstein's (Loewenstein, 1994) proposed mechanism is based on awareness. He argues that poor knowledge prevents an individual from attending to the knowledge gap because information that *is* known is relatively more salient than information that *could* be known. This idea resonates with the Dunning-Kruger effect that describes a tendency of ignorant people (not to be conflated with unintelligent people) to be unaware of their deficiencies in knowledge (Dunning, 2011). Note, however, that this mechanism does not describe a direct link between curiosity and uncertainty. Rather, it shows how the latter relates to the former via the mediating effect of awareness. On the other hand, Berlyne's (1954) conflict-based mechanism proposes a direct interaction between uncertainty and curiosity. Conflict is the "disagreement" between competing behavioral/internal responses to a stimulus (Berlyne, 1954a, 1957). Highly unfamiliar stimuli fail to arouse conflict because there are no sufficiently activated responses to clash with each other. Curiosity is maximal when there are several equipotential responses. Berlyne's proposal sits well with Hebb's physiological theory of arousal (Hebb, 1955). Specifically, Hebb's notion of *disturbances* in activation patterns of acquired cortical representations – what he calls phase sequences and cell assemblies, respectively (Hebb, 2002) – seems to correspond to Berlyne's concept of conflict. One proposed source of such disturbances is an "unfamiliar combination of familiar things (fear of the strange)" (Hebb, 2002, p. 250), which can disrupt normal responses to either of the familiar things, i.e., create conflict. Hebb's physiological theory, in turn, is in line with the Yu and Dayan's more recent and more precise theories of uncertainty processing in the brain (Yu and Dayan, 2005; Yu and Dayan, 2003).

Theoretical and empirical studies above show us how quantitative variability in situational unpredictability is related to forward curiosity, but the precise mechanism(s) underlying this relationship remain speculative and await empirical validation. Additionally, while Yu and Dayan's account explains how uncertainty and incongruity can be computed, neurophysiological implementation of uncertainty representation is yet to be fully specified (although many important advances have been achieved in the previous decade; see Kepecs and Mainen, 2012; Ma and Jazayeri, 2014; van Bergen et al., 2015). Furthermore, although the involvement of cholinergic and noradrenergic systems is likely, we still lack the description of precise neural pathways from uncertainty and surprise processing to subsequent motivational and affective responses that we experience as curiosity. These are promising avenues for future research on neural mechanisms of curiosity.

If curiosity and interest are signaled by uncertainty, what might generate specific types of uncertainty in the first place? So far, we

have reviewed two propositions concerning two types of uncertainty. Expected uncertainty arises as a result of conflict or ambiguity between competing representations. Epistemic uncertainty follows surprising observations that violate expectations. Let us now speculate over how these ideas might relate to other important determinants of curiosity implicated in the literature: namely, novelty and complexity (Berlyne, 1966; Dubey and Griffiths, 2020).

In general, uncertainty is a product of perception of stimuli (or events) that we encounter. While researchers often talk about incongruity, ambiguity, complexity and novelty as properties of stimuli, psychologically, these terms refer to representational states. They are what individuals perceive from the observed stimuli through the subjective lenses of their prior knowledge. Observers' context-dependent expectations and the stimuli under observation jointly determine the state of their perceptual system. It is possible to conceptually map different states of the perceptual system onto the corresponding constructs implicated in the induction of uncertainty. While these are speculative ideas about the potential representation of different kinds of uncertainty in the perceptual system, we think these ideas merit further discussion, because they provide a basis to formalize and unify concepts whose relationships have been implied but to this day remain tacit.

To start our (speculative) conceptual unification, we need a simple model of perception. Consider a few generic details of contemporary neuro-computational models of perception. These models characterize perception as a hierarchical inferential process (McClelland et al., 2014; Olshausen, 2013; Quian Quiroga, 2016, e.g.,). In such accounts, objects are perceived as concepts through a sequence of inferences on progressively more abstract neuronal representations. Low-level neural populations encode less abstract features (e.g., surfaces, edges, edge orientations) while more abstract neurons encode more abstract features (e.g., shapes, concepts). Neurons within the same processing level are mutually inhibitory, creating within-level competition; neurons on different levels are mutually excitatory, resulting in the bidirectional spread of activation. Perceptual inference consists of parallel integrative guesswork executed at different levels of abstraction: at any given step of encoding, a neuron has to "decide" whether to fire or not based on the bottom-up inputs it receives as data, top-down inputs it receives as context-dependent expectations, and same-level inputs from competing units. To exemplify, one might perceive the same footprint (same sensory image) as belonging to a dog or a wolf, depending on whether it is encountered in a city park or in a wild forest. An unambiguous perception of a stimulus is achieved by the complete agreement between the bottom-up and top-down input streams, so that the competition at the same level reaches a unanimous

consensus. However, there are several ways in which this harmony can be disrupted:

- There might be a strong prior expectation (strong top-down priming) of one feature and a strong bottom-up evidence for another feature. Intuitively, this seems to correspond to a state of *surprise*. Surprising observations can signal epistemic uncertainty which, in turn, can attenuate expectations and thus increase unpredictability. Suppose someone told you, "Every night before going to bed, I go to the bathroom and brush my iguana". The beginning of a sentence creates a strong expectation for the word "teeth", but the actual ending is surprising. At the risk of abusing this already ridiculous example, we further remark that after learning about the person's nocturnal rituals, you might think him somewhat unpredictable.

- The bottom-up signaling might strongly activate several competing higher-level representations. We can say that the data or observation is *ambiguous*. In a very early investigation (Berlyne, 1958), participants looked longer at stimuli that combined visual features of two distinct objects (i.e., were perceptually ambiguous; e.g., a chimera with a camel back, dog head, and elephant legs). To provide a more bona fide example, think of a person who shares a flat with two other people. The sound of someone coming in through the front door would suggest that it is either one of two people, but the information would be ambiguous and perhaps compel the person to peek outside their room.

- The bottom-up data might activate a large number of representations at a relative mid-level of the abstraction hierarchy which fails to converge on a single representation due to the lack of input from the higher-level. This might indicate the lack of high-level encoding of the stimulus, i.e., that the individual has not acquired an efficient way to represent it. Such stimuli might be perceived as *complex* (but also, *relatively novel*; see the next bullet-point). A good example of this is perceiving a string of Chinese characters as complex (e.g., 复 杂 的 [3]). The simple strokes that go into this string are relatively high-level visual features, but they are not further compressed into a word or concept representation. Presumably, a literate Chinese speaker do not perceive this string to be as complex as those who cannot read and speak Chinese.

Relative novelty *refers to a novel combination of familiar features (Barto, Mirolli, and Baldassarre, 2013)*

- Similarly to complexity representation, the bottom-up data might not sufficiently activate a high-level encoding, while activating relatively few mid-level representations. This could indicate

3 If you are curious about this complex stimulus, this is how Google translates the English word "complex"

relative novelty without complexity. Alternatively, one could conceive novelty as residual competition at some level of encoding due to the lack of convergence at higher-levels. For example, someone who cannot read Cyrillic and does not know Russian, the character string с л о ж н ы й might appear novel. Like in the example with Chinese characters, this string consists of fewer simple features that are arranged in a somewhat more familiar way. Both 复 杂 的 and с л о ж н ы й are relatively novel, but the latter should appear less complex to a literate English speaker who does not know Chinese or Russian. This suggests how relative novelty and complexity might be related.

Having discussed the situational factors influencing curiosity/interest, we can now turn to how this state is experienced. We can identify two distinct experiential aspects. The affective/emotional aspect pertains to how the state of curiosity/interest is experienced from a subjective (phenomenological) view-point. The motivational aspect concerns what the individual does as he or she experiences curiosity/interest.

3.3.2 *Affect and Motivation*

Curiosity, is often assumed to be the motivation to reduce an aversive state. From early on, it has been conceptualized as a drive that people are motivated to eliminate (Berlyne, 1954a; Loewenstein, 1994). Observations that uncertainty can induce fear and anxiety (Carleton, 2016; Hebb, 2002) lend further support of this view. However, like any other state, uncertainty must undergo a cognitive appraisal that determines the affective profile of the ensuing motivational state (Anderson et al., 2019).

If deemed irrelevant by the appraisal process, uncertainty may fail to elicit any affective response. There are many things about which we are knowingly ignorant; and even if we have some faint knowledge about such things, we do not necessarily react to our ignorance positively or negatively. I doubt that not knowing the first letter of my great-grandfather's name provokes any kind of emotion in the reader. Besides, if uncertainty was anxiety or fear-inducing, young children would have to be in a constant emotional distress. In order to elicit any affective response, the situation's personal relevance – its significance for one's goals and needs – must be evaluated (Cunningham and Brosch, 2012; Lazarus, 1991). Accounts that portray uncertainty as anxiety-inducing assume the presence of perceived threats (Grupe and Nitschke, 2013). Without referring to peoples' current goals and needs, it is impossible to make a general statement about how they feel about uncertainty.

If the perceived uncertainty is somehow relevant to the individual, it may be accompanied by positive or negative feelings. For instance,

participants in Noordewier and van Dijk's (2017) study reported positive feelings about curiously anticipating an intriguing video clip, but only when they expected to watch it after only one minute of waiting, in contrast to those who expected a 30-minute delay and experienced their curiosity negatively (but see van Lieshout, de Lange, and Cools, 2020). These discrepancies are in line with a more recent theory of I/D-type curiosity (Litman, 2019), which maintains that curiosity can be experienced as a feeling of deprivation (D-type) or a feeling of interest (I-type). The D-type curiosity is associated with negative affect, while the I-type is experienced positively. The I/D-type taxonomy resonates with contemporary views on curiosity and interest (e.g., Hidi and Renninger, 2019; Murayama, FitzGibbon, and Sakaki, 2019; Shin and Kim, 2019), which propose that there are two distinct affective states that fuel information-seeking behavior. One state is accompanied by an unpleasant feeling of a knowledge-gap and another is characterized by a pleasant anticipation of learning. Although distinct, these states tend to coincide with one another (Hidi and Renninger, 2019).

A useful framework for understanding affective and motivational states underlying non-instrumental information-seeking is the INCENTIVE MOTIVATION theory (Berridge, 2009; Robinson et al., 2016). The theory revolves around a distinction between systems of "wanting" and "liking", also referred to as *incentive salience* and *hedonic impact*, respectively. Traditionally, the two terms are marked with quotes to distinguish them from more intuitive, everyday meanings of wanting and liking. Incentive salience ("wanting") refers to the visceral desire usually directed at a specific state or stimulus. Notably, "wanting" is distinguished from cognitive desire – a more explicit and *intentional* kind of wanting that involves conscious thoughts about the object of desire and its emotional significance, akin to goal representation. On the other hand, hedonic impact ("liking") refers to the objective hedonic responses to stimuli such as activation of hedonic "hotspots", certain facial expressions, and other physiological manifestations of pleasure (Berridge and Kringelbach, 2015). As such, hedonic impact is distinguished from the conscious feelings of liking. Normally, "wanting" coincides with cognitive desire and both are intricately related to "liking" and the concurrent conscious hedonic experience. While the unconscious "wanting" and "liking" systems might be phylogenetically and ontogenetically older, conscious desires and feelings confer additional evolutionary value by enabling more flexible and allostatic behavior (Damasio and Carvalho, 2013).

Intentionality *refers to the contents or representational targets of certain mental states, their "aboutness" or "directedness".*

Incentive salience and hedonic impact are implemented by anatomically overlapping but functionally distinct systems in the brain (Berridge, Robinson, and Aldridge, 2009). The "wanting" system resides in the circuitry that includes midbrain DA nuclei and their mesolimbic projections. It is responsible for maintaining seeking behaviors that can be directed towards rewards and reward-conditioned cues, but can also

be completely non-intentional (Berridge, 2009). The "liking" system includes relatively small neuronal populations distributed throughout the frontal and insular cortices and the midbrain (Berridge and Kringelbach, 2015). It functions to communicate value of experienced states to other systems in the brain (Damasio and Carvalho, 2013). These distinct functions, as one might expect, are highly complementary: hedonic responses of "liking" signal what the organism should "want". However, the activity of the two motivational components is dissociable, which can lead to confusing behaviors, such as "wanting" something without "liking" it or "liking" something without having "wanted" it. Since voluntary information-seeking is a type of motivated behavior, the incentive motivation theory has a potential to explain why people can be "seduced" by either pleasant or unpleasant information (FitzGibbon, Lau, and Murayama, 2020).

Incentive motivation theory has gradually replaced the previously dominant drive theories of motivation on the grounds that drive reduction is a weak (negative) reinforcer at best (Berridge, 2018). According to the incentive motivation theory, while the unpleasantness of a drive (e.g., a feeling of deprivation) can be experienced – and to a high degree – it does not spark or fuel motivation, but serves to intensify it. Recall that the affect associated with the "wanting" state is determined by an independent appraisal process. This implies that "wanting" information is not an inherently unpleasant feeling, as drive theories of curiosity (Berlyne, 1954a; Loewenstein, 1994) have assumed. Thus, information-seeking is incentivized by information and can be modulated by uncertainty that can be affectively positive, negative, or neutral. Crucially, the desired information does not have to be available to activate the "wanting" system (just like we don't need to see a cup of coffee in order to start craving one). As mentioned earlier, incentive salience can be initiated by learned associations between rewarding stimuli and arbitrary cues.

Uncertainty can be regarded as a cue that triggers the "wanting" system, while information that resolves uncertainty can be viewed as an input to "liking" system. As noted above, the degree of uncertainty can have a moderating effect on the intensity of "wanting". Several empirical findings support this view. In addition to the "observing task" study reviewed earlier (Bromberg-Martin and Hikosaka, 2009), Aron et al. (2004) found a correlation between categorization uncertainty and midbrain activity. Gruber et al. (Gruber, Gelman, and Ranganath, 2014) reported a positive relationship between curiosity ratings about trivia questions and activation levels in the DA midbrain and the nucleus accumbens (NAc). White et al. (White et al., 2019) used single-cell recordings to show graded activity related to reward uncertainty in the dorsal striatum (DS) and the ventral pallidum (VP), both involved in incentive salience (Smith et al., 2009; Volkow et al., 2002). These are only indirect clues that raise the appeal of the idea that

uncertainty moderates information-"wanting". Not all imaging studies on human curiosity report associations between uncertainty/curiosity and the incentive salience system (Jepma, 2012; van Lieshout et al., 2018). To make things clearer, we need more research focusing on the relationships between degrees of uncertainty, curiosity, and incentive salience.

How does uncertainty get associated with information to become a cue for information-"wanting"? The answer to this question might explain individual differences in personality traits like tolerance to uncertainty (Hillen et al., 2017), need for cognition (Cacioppo and Petty, 1982), developed individual interests (Hidi and Renninger, 2006) and more. The mechanism proposed by the incentive motivation theory is conditioning (we might as well call it reinforcement learning (RL); Maia, 2009; Murayama, FitzGibbon, and Sakaki, 2019). Actions that reduce a previously registered state of uncertainty are reinforced by "liked" states or stimuli and discouraged by "disliked" ones. The agent can thus develop abstract generalizable strategies (e.g., read a book, search in Google, ask a parent, etc.) to obtain information cued by uncertainty in various contexts. Interestingly, the theory of incentive motivation suggests that through this associative process, uncertainty itself can eventually become the target of "wanting" so that the agent will be willing to work to get to this state. However, it should be emphasized that it is the initial "liking" of information that is rewarding, and as such, it is the foundation of motivated information-seeking in response to uncertainty. An intriguing and empirically testable prediction follows: without a sufficiently frequent pairing of uncertainty and "liked" information, no appreciation of uncertainty states can develop in a non-instrumental setting.

3.4 "LIKING" INFORMATION

Let us now take a step back and recall the idea that valuing non-instrumental information is biologically adaptive because it enables individuals to enrich procedural knowledge, resulting in large and diverse skill sets. The motivational mechanism of uncertainty-triggered, non-instrumental information-seeking provides a plausible but not very detailed explanation for how the knowledge accumulation process is instantiated in humans, and perhaps other animals. To render a more complete picture, we need to understand the sufficient conditions for "liking" information. This is by no means an unexplored territory, as researchers have proposed several potential explanations. It is important to mention that different propositions reviewed below are compatible and should not be regarded as alternative explanations of motivated information-seeking. A more productive mindset is to consider how different aspects of informational appeal jointly determine situational curiosity and the development of more per-

sistent interests (see Bromberg-Martin and Sharot, 2020; Daddaoua, Lopes, and Gottlieb, 2016; FitzGibbon, Komiya, and Murayama, 2021; Kobayashi et al., 2019).

One proposition is that information is "liked" because it induces an inherently pleasing anticipation about a desired event in the future – a phenomenon known as *savoring* (Loewenstein, 1987). The preference to learn about upcoming rewards has been empirically demonstrated in humans (Iigaya et al., 2016; Kobayashi et al., 2019; van Lieshout et al., 2018) and other animals, including monkeys (Bromberg-Martin, 2011; Bromberg-Martin and Hikosaka, 2009; Daddaoua, Lopes, and Gottlieb, 2016), and pigeons (Gipson et al., 2009; Spetch et al., 1990). For instance, when presented with two options for requesting a time-delayed rewards (one more probable and one less probable), pigeons prefer the less probable option when it offers the chance to learn with certainty about the delayed outcome (Gipson et al., 2009; Spetch et al., 1990). For a more intuitive understanding, imagine you were offered two lottery tickets, A and B. A has a 50% chance of winning \$100 while B has a 75% chance. B seems like a no-brainer. However, you can take ticket A (but not ticket B) to a fortune-teller, who will tell you in advance whether it is a winning ticket or not. Does it make ticket A more attractive? What if you had to wait a week, or a month, or a year until the draw? Pigeons were shown to choose "ticket A". Sometimes, if they have to wait long enough, they can even prefer a 50% reward to a certain reward (Spetch et al., 1990). Informally, the "savoring" account of this phenomenon holds that (under certain conditions) the inherent value of the reward-predicting cue outweighs the value of the reward itself.

Similar savoring phenomena were demonstrated in monkeys (Daddaoua, Lopes, and Gottlieb, 2016) and humans (Iigaya et al., 2016). In Daddaoua, Lopes, and Gottlieb (2016), monkeys sampled a redundant reward-predicting cue even when it carried no added predictive benefit (i.e., the reward was anticipated with certainty in the first place). The authors explain in this behavior in terms of conditioning (RL) based on a composite reward function consisting of Pavlovian, operant, and informational value components. Like in the previous research, the monkeys searched for information more vigorously, when they were uncertain about it. Iigaya et al. (2016) offered adult male participants an opportunity to sample valid advance information about the presentation of an extrinsic (graphic) reward after a delay. When the delay was short, participants did not care to sample advance information; however, the tendency to request the informative cue increased with the delay in reward presentation. This effect was non-monotonic due to the temporal discounting of the reward value. The authors explain the preference for advance information about a delayed uncertain reward in terms of reward-prediction errors (RPEs) that boost the temporally discounted anticipatory utility. In a related human behavior

study, Kobayashi et al. (2019) participants could choose to partially reveal a two-part monetary prize by either maximally reducing uncertainty about the total amount or by savoring the greater part of the prize expected with certainty. The majority (though not all) of participants chose to savor[4] the highly paying part of the prize and remain ignorant about the less lucrative part. The authors account for the diversity of informational preference using a choice-utility model with two free parameters describing a preference for uncertainty reduction or anticipatory savoring, respectively. All three of the aforementioned studies propose very different computational accounts for the savoring phenomenon. Future work should aim at unifying these accounts in a single framework that accounts for similar but not identical behaviors displayed in different studies of anticipatory utility of information.

The "savoring" account of "liking" information can be viewed as an instance of a broader phenomenon: information can be "liked" for its pleasing implications. For example, some pieces of information support desirable beliefs that people are personally invested in (e.g., confirmation bias Nickerson, 1998). Let us call it the "hedonic value" conjecture. The taste for self-pleasing information might have adaptive value beyond knowledge acquisition, for example, by virtue of elevating self-efficacy (Bandura, 1977) (the subjective belief of achieving certain outcomes in certain situations) for accomplishing desirable tasks (Bromberg-Martin and Sharot, 2020). This idea is certainly plausible, but the "hedonic value" conjecture is insufficient for explaining many instances of non-instrumental information-seeking where information does not inherently imply anything positive for the individual. One example is counter-factual information: humans (FitzGibbon, Komiya, and Murayama, 2021) and monkeys (Wang and Hayden, 2019) sample information about what happened in the past to learn what could have happened in the present – something that can actually induce negative feelings of regret. Other examples include related behaviors such as explanation-seeking (Coenen, Nelson, and Gureckis, 2019; Liquin and Lombrozo, 2020) and exploratory play (Chu and Schulz, 2020a; Cook, Goodman, and Schulz, 2011). Upon encountering observations that are surprising or have ambiguous causal precedence, humans engage in investigatory activity aimed to resolve the ambiguity. We seem to seek observations that convey accurate rather than pleasing information about the world. Intuitively, "liking" such information should facilitate autonomous knowledge acquisition.

Note:, no single "conjecture" can account for all of motivated information-seeking.

To advance the idea that "liking" knowledge-enhancing information, we need to characterize what we mean by "knowledge" more precisely. In cognitive-computational literature, knowledge is instantiated in the parameters of formal cognitive models. Two broad classes of

4 Interestingly, when the outcomes were framed as monetary losses, rather than gains, fewer participants chose the high-loss option which is consistent with the notion of anticipatory dread – the counterpart of savoring.

models dominate the field: probabilistic models (Griffiths et al., 2010) based on the assumptions about the computational problems facing intelligent systems, and connectionist models (McClelland et al., 2010) based on algorithmic and implementational constraints of intelligent systems. Regardless of the modeling approach, knowledge in these computational systems serves the same purpose: enabling inferences under uncertainty. These inferences can be perceptual (what is "out there"? what am I hearing/feeling/looking at etc.?), causal (what has led to the current state?), predictive (what will be the next state?), or instrumental (what to do to get to the desired state?), to name a few[5]. Importantly, inferences are improved by generalization from data (i.e., induction).

Following this more precise characterization, we can now introduce the "epistemic value" conjecture – *information is "liked" when it improves knowledge*. To improve knowledge is to improve the accuracy of knowledge-based inferences. We have mentioned before that aleatoric uncertainty, or unpredictability, refers to "irreducible uncertainty", which is how it is often characterized (Hüllermeier and Waegeman, 2021; Kendall and Gal, 2017). While there is a certain sense in which this characterization is undeniable, we can also say that we "reduce" the uncertainty of a die roll or a coin toss by observing their outcomes. In this sense, reducing aleatoric uncertainty is much simpler than reducing epistemic uncertainty, as it can be done in a relatively short period of time. On the other hand, epistemic uncertainty consists of multiple episodes of aleatoric uncertainty reduction, just like learning that a coin is fair involves multiple coin tosses. Epistemic uncertainty reduction appears to be better aligned with the "epistemic value" conjecture than does aleatoric uncertainty reduction. Curiosity experiments often measure uncertainty to relate it to people's self-reported ratings (of curiosity), but they do not always explicitly state which kind of uncertainty is being studied. As we briefly discuss in the next section, this can lead to confusion and incompatibility in results.

3.5 LEARNING PROGRESS HYPOTHESIS

The "epistemic value" conjecture is closely related to the concept of LP originating in AI (e.g., Oudeyer, Kaplan, and Hafner, 2007; Schmidhuber, 1991b). LP-inspired intrinsic-reward functions push agents to progressively explore their environments by engaging in learning situations (see Chapter 2) where their knowledge is improving. Thus, despite revolving around the principle of prediction-error reduction, LP-based reward functions are similar to other so-called heterostatic adaptive reward functions, like competence-progress or information-

5 The limited number of query-forming words (see Chu and Schulz, 2020a) suggests that there is a finite amount of inference classes that people might make, so it might be useful to taxonomize them.

gain motivations (Oudeyer and Kaplan, 2007). At a higher level, all of these reward functions are designed to encourage agents to explore states where their knowledge is improving.

By re-engaging in situations where their predictions or competence are improving, artificial agents are able to efficiently address two non-trivial computational challenges of open-ended learning (Gottlieb et al., 2013). The first challenge is the virtually infinite amount of things one can learn and the concurrent limited amount of resources individual organisms possess. Given an ability to perceive the similarity of various sensorimotor contexts, LP-guided agents can fragment the vast learning space and delineate "progress niches" inside it that are appropriate for their level of knowledge or ability (e.g. Etcheverry, Moulin-Frier, and Oudeyer, 2021; Forestier et al., 2020; Oudeyer, Kaplan, and Hafner, 2007). The second challenge is the prevalence of uncorrelated events and unreachable goals that agents cannot learn or learn to achieve in principle. The strategy of pursuing LP helps agents divert their resources from trying to improve predictability of uncorrelated variables and attempting to reach states that agents are not ready to or will never be able to reach. Furthermore, progress-based heuristics for time allocation across multiple tasks have been shown to be optimal in certain conditions (Lopes and Oudeyer, 2012; Son and Sethi, 2006). Thus, LP-based motivation presents a potential solution to tough computational challenges faced by artificial and biological agents alike (Gottlieb and Oudeyer, 2018; Gottlieb et al., 2013; Oudeyer, 2018).

Following the advancements in implementing self-organized exploration in robots, Kaplan and Oudeyer (2007) put forth the the Learning Progress Hypothesis (LPH). In a nutshell, the original LPH contends that the human brain features a system for detecting reduction of prediction errors that signify "progress niches" – sets of sensorimotor states that promise to improve the brain's forward-predictive knowledge. However, there are numerous ways in which knowledge improvement can be formalized (Graves et al., 2017; Linke et al., 2020; Oudeyer and Kaplan, 2007; Santucci, Baldassarre, and Mirolli, 2013) and over the years, the notion of LP has broadened to include other signals that indicate knowledge improvement. The "epistemic value" conjecture introduced in this chapter is entirely consistent with this broader notion of LP. Thus, the LPH which we will advance in this thesis proposes that the brain monitors its own knowledge and that states that are associated with knowledge-improvement (e.g., uncertainty reduction, prediction-error reduction, competence gain, etc.) are intrinsically valuable, and thus motivating. Whereas the original LPH involves prediction errors as the metric of knowledge and prediction-error reduction as the metric of learning, our "epistemic value" conjecture refrains from specifying representational details.

Knowledge-improvement expectations are also tightly related to the concept of *self-efficacy* (Bandura, 1977). Evidence for improving performance – a form of LP called *competence progress* (see Oudeyer and Kaplan, 2007) – might foster feelings of self-efficacy which bear on the psychological need for competence (Blain and Sharot, 2021) postulated by the Self-Determination Theory (Ryan and Deci, 2000). According to this view, mastery-oriented effort is fueled by self-efficacy beliefs that are, in turn, supported by LP. This self-efficacy account of intrinsic motivation adds potentially important details to the current theoretical perspective. The relationship between LP and self-efficacy will be discussed in more detail in Chapter 5.

Before proceeding to discussing the work that supports the LPH, let us first discuss how it relates to exploration in instrumental settings. It is worth examining information-seeking in instrumental contexts because it might have similar mechanisms with non-instrumental information-seeking (Gottlieb and Oudeyer, 2018; Gottlieb et al., 2020). In the so-called "multi-armed bandit" task (see Averbeck, 2015), participants sample from 2 or more sources of stochastic rewards, called "bandits". Of interest is how participants use their finite sampling opportunities to maximize the amount of total reward. On any given sampling trial, they can sample from bandits that they believe are most rewarding (i.e., exploit) or try less familiar ones in order to learn more about them (i.e., explore). Several studies report that it is common for people to explore bandits that they are least certain about (Gershman, 2018a; Schulz et al., 2019; Speekenbrink and Konstantinidis, 2015). Importantly, the uncertainty pursued by participants is the posterior (epistemic) uncertainty regarding expected-value estimates of the bandits, not the irreducible expected (aleatoric) uncertainty of independent bandit outcomes (Schulz and Gershman, 2019). Epistemic uncertainty is always reducible (Hüllermeier and Waegeman, 2021), even if a bandit is unpredictable. Moreover, it is typical to assume uninformative or weakly informative priors about the expected values of the bandits. Therefore, a bandit of maximal (reducible) uncertainty offers the most LP, because low-entropy priors (more strongly held beliefs) are less likely to be influenced by new data compared to less informative priors. Thus, uncertainty maximization strategies observed in instrumental exploration contexts of multi-armed bandits seems compatible with the LPH.

Several results from research on non-instrumental information-seeking are also compatible with the LPH. For instance, van Lieshout et al. (2018) had participants play multiple trials of risky lotteries. Participants could sample accurate information about the amount of a given lottery in advance by paying a known price. The results showed a *monotonic* relationship between uncertainty of a lottery and the requesting of information. Since the experimenters delivered the promised information reliably upon request, participants could expect

to improve their prediction about the upcoming reward. That is to say, all learning situations encountered in the study were knowingly learnable. Therefore, maximal-uncertainty lotteries can be reasonably speculated to correspond to learning situations with the highest expected LP. However, it should be noted that these results are also compatible with another information-sampling strategy, which predicts that decision-makers are most curious about maximally uncertain situations.

In contrast to the results from van Lieshout et al. (2018) suggest, studies using trivia questions (Baranes, Oudeyer, and Gottlieb, 2015; Kang et al., 2009) report a *non-monotonic* inverted-U relationship between uncertainty in knowing the answer and self-reported curiosity – participants were more interested in questions for which they reported *intermediate* certainty of knowing the answer. This finding is also consistent with the LPH, as it predicts the strongest motivation for seeking intermediate-complexity stimuli (Baranes, Oudeyer, and Gottlieb, 2014; Oudeyer, Gottlieb, and Lopes, 2016a). The apparent contradiction with van Lieshout et al. (2018) can be explained by noting the crucial differences in experimental stimuli. Learning situations in the lottery task are unrelated to participants' broader knowledge beyond the experiment or even individual trials. In contrast, trivia facts typically make us more knowledgeable, regardless of whether they are learned inside or outside the lav. Since committing declarative propositions to memory (i.e., learning them) depends on prior knowledge (Brod, Werkle-Bergner, and Shing, 2013), intermediate-confidence questions could induce maximal curiosity because their answers are more likely to be committed to memory (thus promising more LP), compared to the answers to maximally uncertain questions.

While findings from both trivia-question and lottery tasks are compatible with the LPH, they are not designed to probe knowledge acquisition over extended periods of time, as it happens in natural settings. There are several studies compatible with the LPH that investigate time-extended information-seeking. For instance, Gerken, Balcomb, and Minton (2011) showed that infants attend to stimulus sets with learnable structure longer compared to sets with unlearnable structure. Kidd, Piantadosi, and Aslin (2012) showed that infants look longer at sequences of intermediate-complexity compared to highly predictable and excessively unpredictable sequences. In a similar study, Poli et al. (2020) provide a more direct support to the LP-based engagement by showing that infants' tend to look away from stochastic sequences when they no longer update an assumed predictive model. Geana et al. (2016) showed that human adults rate invariant and random-uniform number generators as more boring than random-but-predictable number generators based on normal distributions. Leonard et al. (2021) showed that young children who saw evidence for gradual improvement were more likely to stick to a challenging task compared to

children whose improvement record was held constant. While these studies support the LPH, they measure task engagement by first forcing learning situations on participants and then observing how participants disengage. On the other hand, the LPH predicts how people decide to actively engage in learning activities.

A few studies have examined the relationship between free active engagement in time-extended learning tasks and variables closely linked to LP. Son and Metcalfe (2000) gave participants multiple biographical texts to explore over a fixed amount of time. Before free exploration, participants sampled each text and provided metacognitive judgments for how easy they thought the texts were. Participants spent more time on and gave priority to texts that they judged to be easier. Metcalfe and Kornell (2005) subsequently reported a positive correlation between study time and a crude temporal derivative of explicit JOLs: people spent more time studying Spanish-English word pairs for which their judgments of learning were increasing. Interestingly, Metcalfe and Kornell's *Region of Proximal Learning* theory (Metcalfe and Kornell, 2005) predicts learning activities are given priority with respect to their JOL, similarly to how uncertainty can cue and modulate the motivation for information-seeking. Their theory also predicts that perseverance on a learning activity is determined by a temporal derivative of JOL, which resonates with the idea that low-LP do not arouse interest. While Metcalfe and colleagues' work is clearly related to the LP-hypothesis, task designs in the reviewed studies did not permit a close tracking of learning rates and task engagement.

In metacognition research, a JOL is a self-reported confidence in retrieving a stimulus in the future.

It is still unclear whether the human brain hosts a mechanism (or several mechanisms) to compute LP and generate the motivation for seeking it. Considering the necessary computational features of LP algorithms as well as a number of studies that are compatible with their operation, studying the role of LP in a dedicated task is an important step in advancing our understanding of intrinsically motivated information-seeking. In Chapter 4, we report an original study that introduces a behavioral paradigm for studying the process of learning of multiple non-instrumental activities as it unfolds concurrently with active exploration. Using a computational model of trial-by-trial choice utility, we demonstrate that humans do show sensitivity to LP while exploring multiple learning activities. Next in Chapter 5 (also in Chapter 6), we explore the potential metacognitive mechanisms for LP computation and consider a belief-based process by which LP representation shapes the intrinsic motivation to practice a real-world sensorimotor task.

Part II

EMPIRICAL CONTRIBUTIONS

4

PROGRESS-BASED EXPLORATION IN HUMANS

4.1 INTRODUCTION

As explained extensively in Chapter 2, curiosity is a fundamental drive in human behavior and a topic of great interest in neuroscience and cognitive psychology. The vast majority of recent research on curiosity has operationalized it as intrinsically motivated information demand, using tasks in which participants can request information about future events but do not have the opportunity to exploit (act on) the information. The studies have shown that humans and other animals seek to obtain information as a good in itself and this preference is encoded in neural systems of reward and motivation, suggesting that information is rewarding independently of material gains (Bromberg-Martin and Hikosaka, 2009; Duan et al., 2020; Kang et al., 2009; Lau et al., 2020).

While these findings tap into the intrinsic motivation behind curiosity, they are yet to capture the full scope of curiosity-driven investigations (Gottlieb and Oudeyer, 2018). Specifically, in natural settings, humans investigate questions on much longer time scales relative to those tested in the laboratory. In contrast with tasks of information demand in which participants request information about brief unrelated events – e.g., a forthcoming reward or a trivia question – in natural behavior, learners maintain sustained focus on specific activities such as reading an article, conducting an online search, or taking a course. Operating from early infant development (Bazhydai, Twomey, and Westermann, 2020), this ability for sustained investigations may underlie the most important ecological role of curiosity, as it allows people to develop individual interests and skills and, ultimately, discover explanatory models and latent structures of the world (Dubey and Griffiths, 2020; Hidi and Renninger, 2019; Schwartenbeck et al., 2019).

Very little is known about how people self-organize investigations to achieve learning on longer time scales. Natural environments afford a practically infinite number of activities that a curious learner can in principle investigate. However, given the limited time and resources available for investigation, the learner must carefully select which activity to engage with to enable discovery. Formal treatment of this "strategic student" problem prescribe how learners should allocate study time to maximize learning across a set of the activities (Lopes and Oudeyer, 2012; Son and Sethi, 2006) but show that the optimal allocation is very sensitive to the shape of the expected learning trajectory, which is not available to learners in practice (Son and Sethi, 2006).

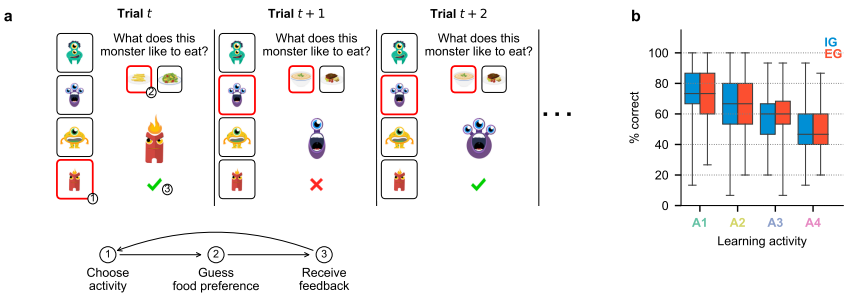

Figure 4.1: **Task design and difficulty manipulation. a**, Trial structure during free play. The panels show 3 example free-choice trials consisting of 3 steps each. Each trial began with a choice among 4 "monster families" depicted as visual icons (1). This was followed by the presentation of a randomly drawn individual from that family and a prompt to guess which of two possible foods the individual liked to eat (2). After guessing (2), the participant received immediate feedback (3) and the next trial began. Participants were free to repeat the previously sampled activity (e.g., trial $t + 2$ in this figure) or switch to any other monster family (e.g., trial $t + 1$) as they wished. **b**, Performance during the forced-choice familiarization stage. Each box plot shows the percent correct (PC) during the 15 familiarization trials in which participants had to play each activity for the internal goal (IG) (blue; $N = 186$) and external goal (EG) (red; $N = 196$) groups. Horizontal bars inside boxes show the median values across all participants in a group; box boundaries show the 1st and the 3rd quartiles; whiskers show sample minima and maxima. Image credits (Fig. 1, **a**): monster character designs by macrovector/Freepik; food-item designs by brgfx/Freepik.

A common proposal for how people resolve this conundrum is that they prioritize study items based on their perceived difficulty, i.e., their perceived level of knowledge or competence on a task, but the precise form of this prioritization is under debate. Several studies have shown that people prioritize tasks with high difficulty or high uncertainty (Loewenstein, 1994; Schulz et al., 2019). In contrast, an expanding literature proposes that people prefer intermediate difficulty (Berlyne, 1960) in a range of conditions including curiosity about trivia questions (Baranes, Oudeyer, and Gottlieb, 2015; Kang et al., 2009), choices among sensorimotor activities (Baranes, Oudeyer, and Gottlieb, 2014), infant attention (Kidd, Piantadosi, and Aslin, 2012) and aesthetic appreciation (Gold et al., 2019; Tsutsui and Ohmi, 2011).

Strategies that prioritize high versus intermediate difficulty activities may have different computational bases and ecological roles. A preference for high difficulty tasks may emerge from computational architectures that assign intrinsic utility to prediction errors or uncertainty, thus motivating agents to venture beyond familiar activities (Bellemare et al., 2016; Dayan and Sejnowski, 1996; Pathak et al., 2017;

Schulz et al., 2019). In contrast, a strategy prioritizing activities with intermediate difficulty may emerge from control architectures based on learning progress (LP; Colas et al., 2019; Graves et al., 2017; Kaplan and Oudeyer, 2007; Kim et al., 2020; Schmidhuber, 2010; Twomey and Westermann, 2018) that monitor the temporal derivative of performance - e.g., percent correct (PC) - and generate intrinsic rewards for activities in which the agent's performance changes with practice.

LP-based algorithms are particularly important in naturalistic environments because they allow agents to avoid not only highly familiar tasks but also unlearnable tasks – i.e., activities that are intrinsically random or cannot be mastered with the learners' current knowledge or skills (Forestier et al., 2020; Kim et al., 2020; Oudeyer, Kaplan, and Hafner, 2007). Unlike PC-based algorithms that steer agents toward tasks of maximum difficulty, LP-based algorithms help to avoid random or too-difficult activities. Moreover, these algorithms provide realistic solutions for optimizing study time allocation - by maximizing the progress that an agent experiences in practice without precise knowledge of one's future learning curve (Lopes and Oudeyer, 2012; Son and Sethi, 2006) - and have been applied to automate curriculum learning in difficult machine learning problems (Graves et al., 2017; Matiisen et al., 2019; Portelas et al., 2020a) and personalize sequences of learning activities in educational technologies (Clément et al., 2015; Mu et al., 2018; Oudeyer, Gottlieb, and Lopes, 2016b).

Despite the potential importance of LP-based control strategies, there is no empirical evidence of whether, and how, people use such strategies. In the studies conducted so far, people were asked to estimate the difficulty of study materials based on their familiarity with the topic (e.g., biographical text or foreign vocabulary; Son and Metcalfe, 2000). However, no study has tested whether participants can dynamically monitor their performance on an arbitrary activity and use dynamic estimates of PC or its temporal derivative (LP) as predicted by computational algorithms.

Here, we examined this question using computational modeling and a behavioral task in which people self-organized their study curricula based on trial-by-trial feedback about their performance on a set of novel activities. We provide direct evidence that humans show bona fide sensitivity to LP – the change in performance on novel activities – which coexists with a sensitivity to PC and steers people away from unlearnable tasks consistent with computational theories.

4.2 RESULTS

We analyzed data from 382 participants who performed an online task in which they could freely engage with a set of learning activities (Fig. 4.1, **a**). Each trial started with a free-choice panel prompting the participant to choose one of 4 activities depicted as families of

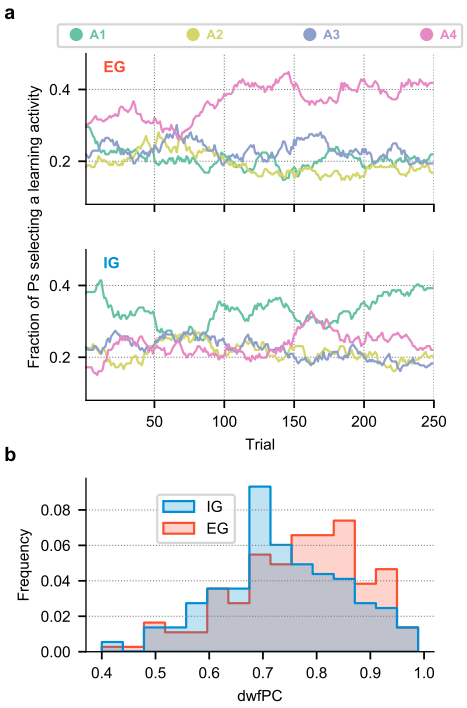

Figure 4.2: **Free play behavior. a**, The fraction of participants selecting each learning activity in the EG ($N = 196$) and IG ($N = 186$) groups (respectively, top and bottom panels) as a function of trial number during the free play stage (no smoothing) demonstrate that group differences in choice patterns persisted throughout the task. **b**, Histograms of difficulty-weighted final percent correct (dwfPC) for each instruction group. The EG group ($N = 196$) achieved better dwfPC scores than the IG group ($N = 186$), but the distributions were broad and overlapping, highlighting important individual variability. The difference between groups was significant with both dwfPC and unweighted average PC scores.

"monsters" (Fig. 4.1, **a**, (1)). After making a choice, the participant received a randomly drawn member from the chosen family, made a binary guess about which food that member liked to eat (Fig. 4.1, **a**, (2)), and received immediate feedback regarding their guess (Fig. 4.1, **a**, (3)). To understand how participants self-organized their learning curriculum, we required them to complete 250 trials but did not impose any other constraint on their choice of activity.

Our key questions were (1) how people self-organize their exploration over a set of activities of variable difficulty, and (2) whether they spontaneously adopt learning maximization objectives when they do not receive explicit instructions. To examine these questions, we manipulated the difficulty of the available activities as a within-participant variable, and the instructions that participants received as an across-participant variable. Difficulty was controlled by the complexity of the categorization rule governing the food preferences. In

the easiest activity (A1), individual monster-family members differed in only one feature and that feature governed their food preference (e.g., a red monster with big flame liked fries and a red monster with small flame liked salad; 1-dimensional categorization). In the next easiest level (A2), family members varied along two features, but only one feature determined preference (1-dimensional with an irrelevant feature). In the most difficult learnable activity (A3) food preferences were determined by a conjunction of 2 variable features (2-dimensional categorization). Finally, the 4^{th} activity (A4) was random and unlearnable: individual monsters had two variable features, but their food preferences were assigned randomly each time a new monster was sampled, and were thus unpredictable with either a rule-based or rote memorization strategy.

Learning objectives were manipulated across two randomly selected participant groups. Participants assigned to the "external goal" group (EG; N = 196) were asked to maximize learning across all the activities and were told that they will be tested at the end of the session. In contrast, participants in the "internal goal" group (IG, N = 186) were told to choose any activity they wished with no constraint except for completing 250 trials. Except for this difference in instructions (and the fact that the EG group received the announced test), the two groups received identical treatments. Each group started with 15 forced-choice familiarization trials on each activity, followed by a 250-trial free-play stage, and gave several subjective ratings of the activities before and after the free play stage (see Appendix 4.B, Self-reported Ratings).

Performance on the forced-choice familiarization stage verified that these manipulations worked as intended. The EG and IG groups had equivalent performance during this stage (Fig. 4.1, **b**; mixed-design ANOVA on percent correct (PC) with group and difficulty as factors; EG vs IG, $F(1,380) = 1.829,\ p = .177$; group $\times$ difficulty interaction, $F(3,1140) = 0.820,\ p = .483$). For both EG and IG participants, performance on each activity was significantly different from all others, suggesting that both groups could use performance feedback as an index of activity difficulty (Fig. 4.1, **b**; mixed-design ANOVA, main effect of activity, $F(3,1140) = 158.400,\ p < .001$; post-hoc pairwise Tukey's HSD tests between all activity levels within each group were significant with all p-values smaller than $p = .01$). Additional evidence from the ratings obtained at the end of the task showed that the EG and IG groups provided similar retrospective ratings of time spent, progress made and interest in learning activities (Appendix 4.B.1), suggesting that they had equivalent engagement and self-monitoring while performing the task.

4.2.1 *Self-challenge*

Despite their equivalent learning ability, EG and IG participants showed different choice patterns and substantial individual variability in the extent to which they challenged themselves and mastered the available tasks.

Analysis of group-level activity choices showed that, while the EG group focused strongly on the most difficult activity (the unlearnable activity that had the lowest PC), the IG group showed a more uniform preference with only a slight bias toward the easiest activity (Fig. 4.2, **a**). Across the entire session, the EG group had significant below-chance time allocation to the two easiest activities and above-chance allocation to the random (lowest-PC) activity (relative to 25%; linear model with sum contrasts: A1: 20.61%, $t(1520) = -3.002$, $p = .003$; A2: 19.29%; $t(1520) = -3.910$, $p = .048$; A4: 36.92%; $t(1520) = 8.156$, $p < .001$). In contrast, the IG group had a significant above-chance allocation for the easiest (A1) activity (A1: 33.00%, $t(1520) = 5.330$, $p < .001$) while spending less time on other activities (A2: 21.42%; $t(1520) = -2.387$, $p = .017$; A3: 22.16%; $p > .05$; A4: 23.43%; $p > .05$; Fig. 4.2, **a**). According to a significant interaction between instruction-group $\times$ activity-type interaction, revealed by a 2-way mixed design ANOVA of time allocation, these differences were reliable ($F(3, 1140) = 14.578$, $p < .001$).

Consistent with their higher self-challenge, average learning achieved by the end of the free-play stage was greater in the EG relative to the IG group (Fig. 4.2, **b**). A measure of difficulty-weighted final PC (dwfPC: the average PC in the last 15 trials spent on each activity scaled by its difficulty rank (see Section 4.4.2.1, METHODS/*Difficulty-weighted final performance*) was significantly higher for the EG group (M = 0.756, SD = 0.127) relative to the IG group (Fig. 4.2, **b**; M = 0.721, SD = 0.126; $t(379.4) = 2.679$, $p = 0.008$, Welch two-sample t-test), and the same result held if we used unweighted average PC (EG: M = 0.787, SD = .118; IG M = 0.756, SD = 0.120; $t(378.1) = 2.539$, $p = .011$, Welch two-sample t-test).

Notwithstanding these group-level differences, participants showed substantial individual variability and, importantly, a subset of those in the IG group adopted levels of self-challenge similar to the EG group. To investigate this variability we categorized each participant based on the number of activities they mastered to a learning criterion - i.e., whether they mastered 1, 2 or all 3 learnable activities (NAM1, NAM2 or NAM3; see Section 4.4.2.2, METHODS/*NAM designation*). The dwfPC score within each NAM group was not affected by instructions, showing that the NAM designation effectively captured the variability in learning achievement (Fig. 4.3, a; pairwise contrasts IG vs. EG conditioned on NAM were nonsignificant, $p > .05$, at all levels of NAM).

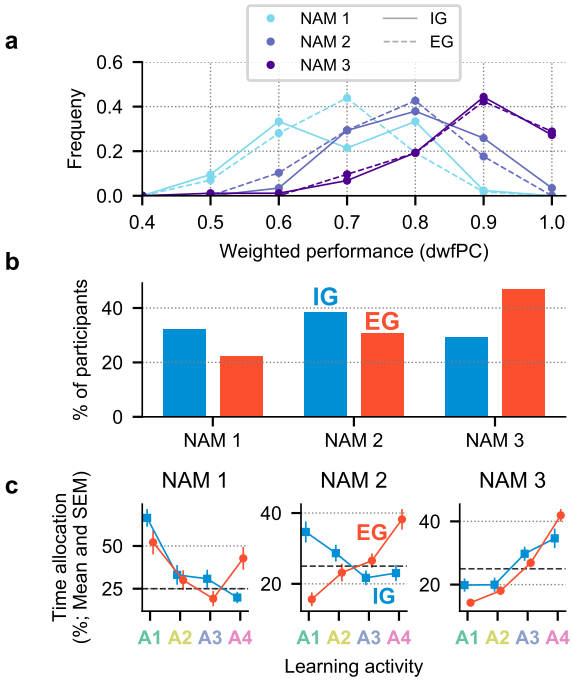

Figure 4.3: **Individual variability within groups. a**, Final performance was the same across instruction groups when accounting for the number of activities mastered (NAM). As expected, the NAM designation captured well the learning achievement of our participants. In light of **b**, this demonstrates that many participants achieved a high performance across learning activities, even without an explicit instruction to learn. **b**, Distributions of participants mastering 1, 2, or 3 activities in each instruction group. Whereas half of the participants in the EG group achieved high performance across learnable tasks, a sizable portion of the IG participants (almost 1/3) were motivated enough to self-challenge and learn without being asked to do so. Only 8 participants in the EG and 9 participants in the IG group failed to master even one activity. Thus, 99 participants mastered only 1 activity ($N_{EG} = 42$; $N_{IG} = 57$), 126 mastered two ($N_{EG} = 58$; $N_{IG} = 68$), and 140 mastered all three ($N_{EG} = 88$; $N_{IG} = 52$) **c**, Time allocation patterns differed by instruction and level of achievement. The three panels show the average time allocation patterns in IG ($N = 177$) and EG ($N = 188$) groups observed over the free-play trials separately for each level of NAM (from left to right, NAM1, NAM2, and NAM3). Circle (EG) and square (IG) symbols represent the average percentage of time spent on an activity in the respective NAM-instruction group; error bars indicate the standard error; the horizontal dashed lines show random time allocation (25%). Time allocation was consistent across the levels of NAM towards harder activities in the EG group. In contrast, only the best learners in the IG group displayed a similar preference, whereas NAM1 NAM2 participants tended towards easier activities.

Importantly, despite not being instructed to study for a test, 64.52% of IG participants mastered more than one activity (NAM2 and NAM3) and 29.59% mastered all 3 activities (Fig. 4.3, b). These percentages were comparable to learning achievements in the EG group, where 74.49% mastered at least 2 activities, and 36.56% mastered all three. The relative proportions of participants at each achievement level were comparable between the two groups across a range of mastery criteria (see Appendix 4.C, for a detailed analysis). Thus, while changing the criterion modified the number of participants who achieved mastery, it left intact the relative fractions of NAM subgroups in the IG and EG groups. This shows that our conclusions are independent of a specific definition of mastery.

While NAM1 and NAM2 participants in the IG group showed choices consistent with the group average – favoring the easiest activity – NAM3 participants showed a distinct preference for A3 and A4 activities that more closely resembled the EG group (Fig. 4.3, **c**). A two-way mixed ANOVAs of time allocation in the IG group showed a significant main effect of activity ($F(3, 525) = 8.847, p < .001$) and a highly significant interaction between activity and NAM ($F(3, 525) = 14.791, p < .001$). In the EG group there was also a significant main effect of activity ($F(3, 525) = 19.407,\ p < .001$) and a significant interaction with NAM ($F(3, 525) = 7.197,\ p < .001$). As Fig. 4.3 (**c**) shows, while participants in NAM1 and NAM2 groups differed in activity selection across the instruction conditions, those who mastered all 3 learnable activities allocated their time similarly. Importantly, a sizeable fraction of the IG group behaved in the same way as people who were instructed to learn and prepare for a test.

To further examine the relationship between learning achievement and activity choices, we created an index of self-challenge (SC) measuring the extent to which each participant tended to challenge themselves. This index was defined as the recent PC of the activity selected on each trial, normalized to the entire range of PC levels the participant experienced so far (see Section 4.4.2.3, METHODS/*Self-challenge index*). Thus, SC values close to 0 denote participants who tended to choose the easiest of the activities they experienced; SC close to 1 denote participants who tended to choose the most difficult activities; and SC near 0.5 denote participants who preferred activities of intermediate difficulty. Supplementary analyses verified that the SC index is a more efficient measure of the tendency to choose challenging activities compared to simple contrasts between pairs of activities (Appendix 4.D).

Plotting dwfPC versus SC (Fig. 4.4) reveals two important insights. First, dwfPC has a strong inverted-U relationship with SC, suggesting that the best learning outcomes were associated with intermediate SC. An additive model of dwfPC that included both linear and quadratic SC-index terms (as well as control variables of initial performance and

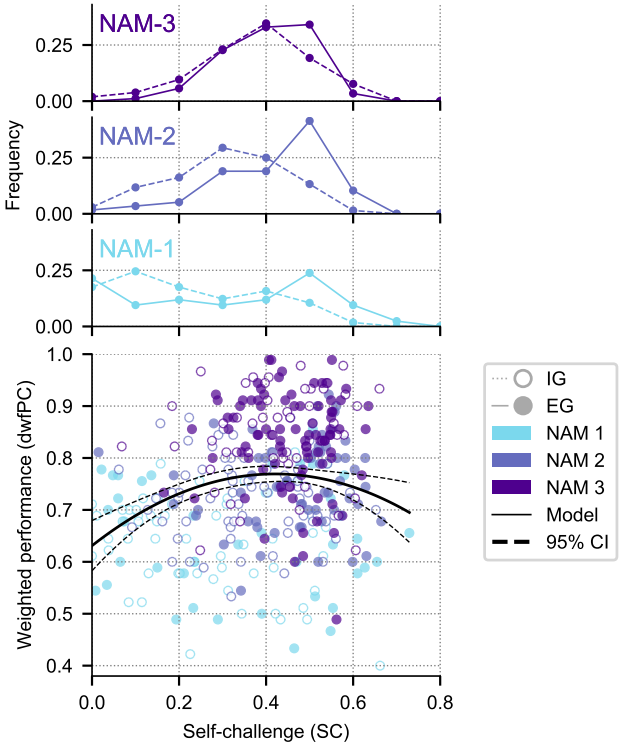

Figure 4.4: **Relationship between self-challenge and final performance.** The scatter plot shows the difficulty-weighted final score (dwfPC; y-axis) as a function of the self-challenge index (SC; x-axis). Each point is one participant. Colors indicate the number of activities mastered: NAM1, $N = 99$ ($N_{EG} = 42$; $N_{IG} = 57$); NAM2, $N = 126$ ($N_{EG} = 58$; $N_{IG} = 68$); and NAM3, $N = 140$ ($N_{EG} = 88$; $N_{IG} = 52$); filled and unfilled circles indicate, respectively, EG ($N = 188$) and IG ($N = 177$) groups. The black curve shows the line of best fit from a linear-quadratic regression model, with 95% confidence intervals represented by the strip surrounded by black dashed lines. The marginal histograms on the top show the distributions of SC scores for each NAM (color) and group (solid and dashed traces). SC was higher for EG relative to IG groups in participants who mastered only 1 or 2 activities (NAM1 and NAM2), and was equivalent, with intermediate values, for participants who mastered all 3 activities (NAM3; top histogram).

instruction) was superior to its counterpart with only a linear term, $\Delta_{AIC} = 11.775$). The linear-quadratic model accounted for a significant fraction of variance ($R^2_{adjusted} = .159$, $F(4, 360) = 18.238$, $p < .001$) and produced a significant negative coefficient for the quadratic term (-0.016, $t(360) = -1.966$, $p < .001$). We replicated this finding when we repeated the analysis using unweighted final PC scores ($R^2_{adjusted} = .191$, $F(4, 360) = 13.642$, $p < .001$, with the coefficient for the quadratic term $= -0.017$, $t(360) = -3.561$, $p = .007$) and when replacing SC with pairwise contrast of activity choices (Appendix 4.D.1,

b). This shows that the finding was not an artifact of the specific ways we measured PC or SC.

Second, participants with different instructions and learning achievements fell on different portions of the inverted-U curve. Participants who did not master all 3 activities (NAM1 and NAM2) fell on the rising and falling arms of the inverted-U curve if they were in, respectively, the IG or the EG group (Fig. 4.4). These participants had equivalent dwfPC but higher SC in the EG relative to the IG group (multiplicative linear model; NAM1, $t(359) = 2.856$, $p = .005$; NAM2 ($t(359) = 4.377$, $p < .001$; Tukey's HSD; see the marginal histograms in Fig. 4.4). Thus, EG participants who failed to master all 3 tasks did so because they over-challenged themselves and those in the IG group did so because they under-challenged themselves. In contrast, participants who mastered all 3 activities were at the top of the inverted-U curve and had equivalent (intermediate) SC in the IG and EG groups (Fig. 4.3, **c**; no significant pairwise contrasts between EG and IG for NAM3, $t(359) = 1.236$, $p = .217$; see the top marginal histogram). Thus, consistent with the activity preferences (Fig. 4.3, **c**): a subset of participants spontaneously adopted intermediate self-challenge strategies and maximized learning regardless of external instructions.

4.2.2 *Computational Modeling and Sensitivity to Learning Progress*

While empirical studies demonstrate preferences for activities of intermediate complexity, they have yet to report specific sensitivity to LP. One study (Son and Metcalfe, 2000) reports that people choose study words that are judged to have intermediate difficulty, but did not measure dynamic sensitivity to LP - the change in performance over time - either alone or in combination with PC.

To examine this question, we fit the participants' activity choices by leveraging the formalism of intrinsically motivated reinforcement learning models (Colas et al., 2019; Graves et al., 2017; Linke et al., 2020; Lopes and Oudeyer, 2012). Such models typically include three major components: (1) a space of learning activities, (2) an intrinsic utility function for each activity, associated with a decision-making mechanism, modeling how they are sampled, and (3) a model of learning mechanisms that improve skills after practicing an activity. Here, we already know the space of learning activities and we can observe the evolution of performance as learners engage in the activities. Thus, we can ask which intrinsic utility function could best explain the participants' choices. To do so, we consider a standard softmax model (in a bandit setting (Linke et al., 2020), in which the utility of an activity is a linear combination of PC and LP:

$$U_{i,t} = w_{\mathrm{PC}} \times \mathrm{PC}_{i,t} + w_{\mathrm{LP}} \times \mathrm{LP}_{i,t} \tag{4.1}$$

PC and LP were dynamically evaluated for each activity i at each trial t based on the recent feedback history. PC was defined as the number of correct guesses over the last 15 trials of activity i, and LP was defined as the difference in PC between first versus second parts of the same interval. We fitted each participants' data (excluding 8 EG and 9 IG participants who did not master even a single activity) as a probabilistic (softmax) choice over 4 discrete classes, using maximum likelihood estimation with 3 free parameters - the softmax temperature (capturing choice stochasticity) and weights w_{PC}, w_{LP} indicating the extent to which each participant was sensitive to, respectively, PC and LP (Section 4.4.2.4, METHODS/*Computational modeling*). Appendix 4.E.1 illustrates the model fitting procedure for an example participant's data.

Note that we use a fixed-size memory to compute LP and PC. The memory size of 15 recent trials was chosen to match the number of practice trials, but there are no reasons to assume that this is a veridical temporal extent of self-assessments. We explore a more flexible approach with freely parameterized LP and PC computation (and additional components based on performance variance) in Section 4.G. While parameterizing PC/LP computation and adding more utility components provides improves fit, increasingly complex models are hard to interpret. Therefore, we stick to simpler models for the remainder of this chapter.

The bivariate form of the model that included both LP and PC (Eq. (4.13)) provided a superior fit to the data in both EG and IG groups. The bivariate model average AIC score ($\text{M} = 491.992$, $\text{SD} = 200.389$) was lower than that of an alternative model based on random selection ($\text{M} = 693.147$; $\text{SD} = 0.0$; the baseline model yields the same likelihood regardless of participants' choices; see Eq. (4.6)) and, importantly, also outperformed univariate models that included only LP or only PC terms (Fig. 4.5, **a**). A 2-way ANOVA of AIC scores showed a significant effect of model form ($F(2, 1089) = 43.992$, $p < .001$), a marginal effect of instruction ($p = .054$), but no interaction between model form and EG/IG groups ($p = .716$). The bivariate model had the lowest AIC scores in a large majority of participants in both groups (EG: 70.74%; IG: 74.01%). Finally, in each group, the bivariate model had a significantly lower AIC relative to each participant's next-best model (Wilcoxon signed-rank test, EG: mean difference $= 21.503$, $\text{SD} = 41.433$; $Z(188) = 55$, $p < .001$; IG: mean difference $= 21.882$, $\text{SD} = 45.383$; $Z(177) = 46$, $p < .001$) and was at least 2 AIC points away from the next-best model in a majority of participants (EG: 58.51%; IG: 62.71%).

The fact that the bivariate model fits free-choice data better than univariate models provides direct evidence that participants are sensitive to LP – a heuristic for the temporal derivative of PC – above and beyond overall error rates. Importantly, the lack of interaction between model

form and instruction shows that participants do not need to be explicitly instructed to maximize learning to demonstrate sensitivity to LP. Additional analyses showed that the PC and LP coefficients remained important even after including a term representing task familiarity (the reciprocal of novelty) in the utility function. As discussed in Appendix 4.F (FAMILIARITY COMPONENT), we focus on models without the familiarity term because in our task, novelty/familiarity is defined only by past choices and is thus circular if used to model choices. Modeling familiarity accounts for choice autocorrelation, but does not explain it. We note, however, that in computational RL studies (Bougie and Ichise, 2020b; Pathak et al., 2017), measures of competence (like our PC measure) are used as a proxy for novelty preference that guides agents towards unfamiliar states.

As a final validation of our models, we conducted model simulations of time-allocation using the coefficients fitted by the bivariate models. We simulated activity choices over 250 trials in each NAM and EG/IG group using the observed success rates in conjunction with the fitted coefficients (randomly sampled with replacement over 500 iterations). As shown in Fig. 4.5 (**b**), the simulations reproduced the main patterns of time allocation, including the preference for activity A4 in the EG and IG NAM3 groups, and the preference for activity A1 in the NAM1 and NAM2 IG groups (see Fig. 4.3, **c**, for comparison), confirming that the bivariate models captured the main features of the empirical data.

Computational theories suggest that sensitivities to PC and LP will have distinct contributions to activity choices and learning. While a sensitivity to PC can motivate people to learn by steering them away from overly easy activities, a sensitivity to LP may protect them from focusing on overly difficult or impossible activities. Several aspects of the w_{PC} and w_{LP} coefficients in our task support these hypotheses.

First, w_{PC} and w_{LP} coefficients were uncorrelated and showed different effects of instructions, suggesting that they capture different influences on choice strategies. We found no correlation between the w_{PC} and w_{LP} coefficients in the IG group (Pearson correlation of normalized coefficients, IG group: $r(186) = -.077,\ p = .298$); EG group: $r(175) = .062,\ p = .399$; the normalization procedure is described in Section 4.4.2.4, METHODS/*Computational modeling*). Moreover, the PC coefficients were on average positive in the IG group and negative in the EG group (consistent with the groups' relative preferences for easier versus harder activities) while the LP coefficients showed no effects of instructions (mean normalized PC coefficient in IG: $\text{M}_{\text{norm}} = 0.255,\ \text{SD} = 0.724$; in EG: $\text{M}_{\text{norm}} = -0.232,\ \text{SD} = 0.741$; 1-way ANOVA, $F(1, 363) = 40.240,\ p < .001$; mean normalized LP coefficient in IG: $\text{M}_{\text{norm}} = 0.079, \text{SD} = 0.640$; in EG: $\text{M}_{\text{norm}} = 0.062,\ \text{SD} = 0.631$; 1-way ANOVA, $F(1, 363) = 0.065,\ p = .799$).

Additional analyses supported the view that while both PC and LP coefficients correlate with higher self-challenge (Appendix 4.D.1, **c**),

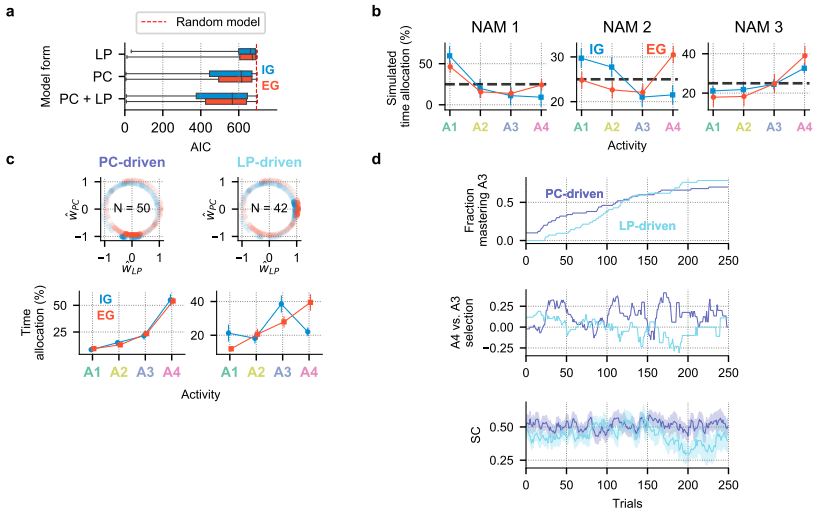

Figure 4.5: **Computational modeling results. a**, The bivariate models had better AIC scores both across and within groups ($N_{EG} = 188$; $N_{IG} = 177$), compared to random-choice and univariate baselines univariate models. Box boundaries represent the 1st and the 3rd quartiles, and the lines inside show median scores; whiskers represent the full sample range. The dotted red line shows the AIC of the random-choice model. **b**, Fitted coefficients reproduce choice patterns across instruction and NAM groups. The panels show the average time allocation patterns obtained by simulating activity choices over 250 trials using $N = 500$ randomly sampled coefficients from the pool of all fitted bivariate models. **c**, Models of two distinct activity-selection strategies. The top row shows the joint distributions of normalized bivariate-utility coefficients. Subsets of these distributions whose data is presented below are highlighted with solid colors. These subsets were formed by first grouping all fitted models into three segments along $\hat{w}_{PC}$ and $\hat{w}_{LP}$, and then selecting groups corresponding to PC-driven and LP-driven profiles. Sample sizes of each subset are shown their respective subpanels. The bottom row shows mean relative frequencies of selecting each activity in the corresponding subset of participants depicted immediately above. LP-driven participants sampled the unlearnable activity (A4) in relative moderation compared the PC-driven group. **d**, LP-driven participants selected allocated time more efficiently for learning and had better learning outcomes. The top row shows fractions of participants in the two groups that reached an objective criterion of 13/15 trials on the hardest learnable activity (A3) at least once in the experiment. The middle row shows the relative preference for activity A4 over A3, defined as the difference between fractions of participants (that still have not mastered A3) who selected A4 minus the fraction selecting A3. The bottom row shows average SC scores in the two groups (shaded regions indicate the standard error).

a sensitivity to LP can steer people away from unlearnable activities.

We first conducted a group-level analysis of the correlation between the coefficients and two model-free measures of task choices: the difference between the time devoted to A3 versus easier activities (indexing the tendency to choose more challenging learnable activities) and the difference between the time devoted to activity A4 relative to the other activities (indexing the tendency to choose the unlearnable activity). Across all participants, lower PC coefficients coincided with a preference for choosing both A3 and A4, but $\hat{w}_{\text{LP}}$ coefficients correlated only with a preference for the learnable, A3 activity (Appendix 4.H.1).

To more closely examine the specific contribution of LP sensitivity we focused on two subsets of participants whose choices were driven predominantly by, respectively, PC or LP. As shown in Fig. 4.5 (**c**, top), PC-driven participants had negative PC coefficients but near-zero LP coefficients and LP-driven participants had positive LP coefficients but near-zero PC coefficients (see Section 4.4.2.4, METHODS/*Computational modeling*, for more details on the grouping procedure). While both groups preferred more difficult activities, the preference for A4 was lower in LP-driven relative to PC-driven participants. Linear regression models of time allocation as a function of activity (A3 or A4) and type of drive showed that PC-driven people engaged in activity A4 more often relative to A3 in both the EG and IG groups (EG: slope $= 76.485$, $t(104) = 7.019$, $p < .001$; IG: slope $= 83.941$, $t(72) = 5.199$, $p < .001$) but this preference was lower or absent in LP-driven participants as shown by its negative interaction with the type of drive (EG: interaction slope $= -47.628$, $t(104) = -2.726$, $p < .001$; IG: interaction slope $= -125.179$, $t(72) = -5.764$, $p < .001$).

Importantly, the lower preference for A4 enhanced learning outcomes in the LP-driven relative to the PC-driven group. As shown in Fig. 4.5 (**d**), after approximately trial 80, PC-driven participants showed a prominent increase in choices of A4 in favor of A3 but this was not seen in the LP-driven participants (Fig. 4.5, **d**, middle row, captured as a decline in SC in the latter group (Fig. 4.5, **d**, bottom row). At around the same time, the fraction of participants mastering A3 in the LP-driven group exceeded that in the PC-driven group (Fig. 4.5, **d**, top row). By the end of the free-play stage, the probability of mastering at least 2 activities was 90.48% in the LP-driven group versus 70.59% the PC-driven group, and the probability of mastering all 3 tasks was, respectively, 64.29% versus 34.98%. Thus, consistent with theoretical predictions, LP-driven choices increase the efficiency of active learning by steering participants away from unlearnable activities.

4.3 DISCUSSION

While the ability to self-organize study time is critical for learning success, finding an efficient organization poses a daunting computational challenge. Prominent theories such as the free energy prin-

ciple postulate that animals are intrinsically motivated to optimize their explanatory models of the environment (Collins, Cavanagh, and Frank, 2014; Schwartenbeck et al., 2019). However, the strategies for optimal exploration that are proposed by these theories are limited to highly simplified laboratory conditions while being typically too complex to be computed in real-world situations (Cohen, McClure, and Yu, 2007). Similarly, mathematical models prescribing how students should allocate study time across competing activities show that optimal allocation is strongly sensitive to the precise shape of the learning trajectory, but this shape is typically unknown to the learner in advance (Son and Sethi, 2006).

LP-based algorithms solve this conundrum by generating intrinsic rewards for activities in which learning recently occurred in practice, and thus provide a uniquely powerful means to optimize choices of study activity using a biologically plausible mechanism. And yet, it is unknown whether or how such choice strategies influence human behavior. Here we use a free-choice paradigm in which participants allocate study time based on dynamic feedback history and provide direct empirical evidence that humans are sensitive to LP.

Converging evidence suggests that humans tend to choose activities of intermediate complexity in a range of disparate settings - e.g., when spontaneously allocating visual attention in infancy (Kidd, Piantadosi, and Aslin, 2012) or declaring aesthetic preference (Gold et al., 2019; Sauvé and Pearce, 2019; Tsutsui and Ohmi, 2011). Our present results show that the preference for intermediate complexity extends to choices of learning activities (see also Baranes, Oudeyer, and Gottlieb, 2014) and, most importantly, that it may be a manifestation of an underlying LP-based mechanism. Thus, the ubiquitous preference for intermediate complexity reported in different settings may reflect an underlying mechanism that steers organisms toward activities that provide learning maximization.

Two major ideas in the literature postulate that exploration is structured based on the learner's competence (prediction errors or error rates) or, alternatively, based on changes in competence over time (learning progress). However, whereas these strategies are typically framed as mutually exclusive alternatives, (Kaplan and Oudeyer, 2003, 2007; Mirolli and Baldassarre, 2013; Santucci, Baldassarre, and Mirolli, 2013) our findings suggest that these two factors are uncorrelated and can jointly shape activity choices and contribute to different aspects of an investigative policy. A sensitivity to PC – with a preference for higher error rates – motivates people to explore more difficult unfamiliar activities, while a sensitivity to LP - the temporal derivative of PC - allows people to avoid unlearnable activities.

The properties of PC- and LP-based control mechanisms in our data suggests that the relative influence of each type of control may depend on the set of available learning activities. Here we used a relatively

simple setting in which the available activities can be quickly mastered, and found that a PC-based strategy strongly contributed to the drive to choose challenging activities rather than stick with already-mastered tasks. However, if the environment is replete with challenging and unlearnable tasks, e.g., during realistic scientific investigation, an LP-based strategy may be more critical for steering learners toward tasks where progress is made as proposed in artificial curiosity (Colas et al., 2019; Kaplan and Oudeyer, 2007; Schmidhuber, 2010).

Our results also pertain to the relation between extrinsic and intrinsic motivation - and specifically the debate whether extrinsic rewards bolster (Duan et al., 2020) or suppress (Murayama et al., 2019) the intrinsic motivation to learn. Our findings suggest that the answer is more complex, as external objectives both enhanced and impaired different aspects of our learners' study strategy. On one hand, external objectives motivated participants to greater self-challenge, as people who were told to study for a test showed a greater tolerance for errors and better learning outcomes than those who did not. On the other hand, external instructions dampened learning achievement by inducing some participants to labor in vain on a random activity rather than learnable activity.

It is important to note that, while previous studies pitted intrinsic motivation against extrinsic monetary incentives (e.g., Murayama et al., 2019), the extrinsic motivation for the EG group in our task came from the specification of a learning objective. In addition, rather than rewarding participants for individual correct answers, our external instruction specified the end-goal but not the local strategy for achieving the goal; this allowed people to choose their activities and commit errors in the short term, in the interest of maximizing learning in the long term. This greater autonomy, we believe, contributed to the synergism we observed, whereby externally imposed goals enhanced the eventual learning outcomes, rather than hindering them. Our findings support two key postulates of Self-Determination Theory stating that intrinsic and extrinsic motivations are not dichotomous but fall on a continuum, and that a sense of agency is a strong factor that motivates people to internalize and meet externally imposed goals (Ryan and Deci, 2020). Thus, the most critical question may not be whether external objectives have beneficial or detrimental effects - but how to balance these objectives to support the investigative strategy that is most efficient in a particular context.

Last but not least, by examining investigations on longer time-scales, our results bear on the increasingly recognized distinction between momentary curiosity and sustained learning and interest (Hidi and Renninger, 2019; Murayama, FitzGibbon, and Sakaki, 2019). Beyond the brief satisfaction offered by fleeting (diversive) curiosity, long-term sustained interest, and the willingness to exert sustained effort in pursuit of such interests, can have profound influence on the life-

long acquisition of competence and skills (Hidi and Renninger, 2019; Hidi and Renninger, 2006). Hidi and Renninger (2019); (2006) proposed a four-stage model of interest development, whereby situational interests is initially triggered and sustained (or dampened) by the environment but with time gives way to well-developed interest in which people spontaneously generate new questions and initiate investigations (Son and Metcalfe, 2000). The fact that many people in our IG group mastered two or more tasks and reported subjective interest proportional to their time allocation (Appendix 4.B.1), suggests that the activities we provided may have triggered their situational interest in the absence of explicit instructions to learn. The fact that higher achievements were more common in the EG group suggests that external instructions help support that fledgling interest. Thus, important questions for future research concern the relation between the mechanisms by which people self-organize their activities, their subjective feelings of interest and the impact of both factors on the development of lifelong interests and skills.

4.4 METHODS

4.4.1 *Participants and Procedure*

Four-hundred participants (including 208 female, 187 male, and 5 participants of undisclosed gender) were recruited for the study on the online platform Amazon Mechanical Turk. Participants were between 19 and 71 years of age, with an average age of 36.15 years, SD = 10.54). All participants provided informed consent. All the procedures were approved by the Institutional Review Board of the University of Rochester.

All participants were told that the experiment will last 45 min to 1 hour and, upon completion, they will be compensated $1 regardless of performance. This scheme was consistent with prevailing rates on Amazon MTurk and with our goals of minimizing the role of monetary incentives and avoiding biasing participants toward activities with consistently high performance. All participants were asked to complete the task on their own in a quiet environment and eliminate external distractions (e.g., turn off cell phones, TV sets, music players, etc.). After receiving detailed written instructions, each participant completed 4 task modules in sequence: (1) 15 forced-choice familiarization trials with each activity; (2) rating of prospective learnability for each task (see below); (3) a free-play stage with 250 trials of free-choice of activity; (4) 6 additional subjective ratings (see below).

Before delivery of the instruction, participants were randomly assigned to either EG or IG group The groups received identical treatments except for the initial instruction. The IG participants received a task description that did not communicate any expectations or ob-

jectives on the part of the experimenters: "In each family there are several individuals, and the appearance of an individual might predict what food they like to eat. When you interact with a monster family, different individuals will be presented to you. For each individual, two food items will be displayed, and you can click on the one you think it prefers. You will receive feedback whether your guess was correct or not", which was followed by brief descriptions of familiarization, free-choice, and questionnaire stages. The EG participants' instructions were identical, except for two additional sentences that included an explicit prescription of a learning goal: "In the main section of the task, we ask you to play for 250 trials and try to maximize your learning *about all the 4 families*" followed by information on the post-session testing module re-emphasizing their objective: "We will briefly test how well you learned to predict the food preferences within each family". After the free-play stage, participants in the EG group received the announced test (between steps 3 and 4) consisting of 15 forced-choice trials on each activity. (However, in our analyses we used the last 15 trials on the free-play stage rather than the test data, as the latter were not available for the IG group). Participants in both groups also provided several ratings of the activities, described in detail in Appendix 4.B.

4.4.2 *Data analyses*

All the *t*-tests reported throughout this chapter (including the Appendix) are two-tailed. We excluded a total of 18 participants – 5 in the EG and 13 in the IG group – who did not appear to be sufficiently engaged in the task based on a response bias criterion (see supplementary Appendix 4.A.1 for more details). This criterion measured the participants' tendency to choose a single response category in each activity (i.e., always guessing the same food item, regardless of the stimulus).

4.4.2.1 *Difficulty-weighted final performance*

Difficulty weighted final PC (dwfPC) is a weighted average of each participant's final PC (fPC) on the learnable activities over the last 15 trials played on the activity. The weights are equal to the activity rank (1, 2 and 3) divided by the sum of the ranks (6). Thus, dwfPC for participant i is $\text{dwfPC}_i = \frac{1}{6}\text{fPC}_{i,A1} + \frac{1}{3}\text{fPC}_{i,A2} + \frac{1}{2}\text{fPC}_{i,A3}$. (Here and in all subsequent analyses we chose a 15-trial time window that was equal to the number of familiarization trials each participant played).

4.4.2.2 *NAM designation*

We divided participants into discrete groups based on the number of activities on which they reached a mastery criterion. The data

presented in this article are based on a criterion of 13/15 correct trials (86.7% correct), which, in a binomial distribution with discrete outcomes, corresponds to $p = .0037$ of arising by chance. Additional analyses verified that the conclusions are robust over a range of criteria (see Appendix 4.C.1). Ten participants (5/154 in the IG group and 5/176 in the EG group) did not master any activity and were excluded from NAM-related analyses and computational modeling.

4.4.2.3 *Self-challenge index*

For each participant, we defined a self-challenge (SC) index for each trial t and activity i as:

$$SC_{t,i} = 1 - \frac{PC_{t,i} - \min_{\forall k \in K} PC_{:t,k}}{\max_{\forall k \in K} PC_{:t,k} - \min_{\forall k \in K} PC_{:t,k}} \tag{4.2}$$

where $PC_{t,i}$ is the recent PC of the selected activity the participant selected on trial t (measured over the last 15 trials on that activity, including familiarization trials) and where $\min_{\forall k \in K} PC_{:t,k}$ and $\max_{\forall k \in K} PC_{:t,k}$ are the minimum and maximum PC experienced by the participant over the entire set of trials (including both free- and forced choice) prior to trial t and over the entire set of activities K. Thus, SC values close to 1 indicate a tendency to select activities that yield the minimum PC ("over-challenging") and values closer to 0 indicate a tendency to select activities with the highest PC ("under-challenging"). To get a single SC index for each participant, we averaged each participants' the trial-wise SC scores across the entire free-play stage. Supplementary analyses verified that the SC index was a better, more concise measure of the preference for challenging tasks relative to the pairwise preferences between different combinations of activities (see Appendix 4.D.1).

4.4.2.4 *Computational modeling*

To understand which intrinsic utility function could best explain the task sampling behavior, we consider a model in the bandit setting ((Linke et al., 2020)), where an intrinsic utility function for each task, measuring its value, is used to decide which task to sample probabilistically. The sampling mechanism used here is the softmax function, following prior models of human decision-making in (Nussenbaum and Hartley, 2019). This softmax function simultaneously translates the underlying choice utilities into selection probabilities and scales the correspondence between utility and probability:

$$p_t(\text{choice}_i) = \frac{e^{U_{i,t} \times \tau}}{\sum_{\forall k \in K} e^{U_{k,t} \times \tau}} \tag{4.3}$$

U_i is the subjective value of choice i, and k indexes the utilities of all items in the set of available activities K (including i); the parameter τ,

known as temperature, controls how strongly the item values determine the probability of their selection. U was defined for each trial as described in the Results section (Computational modeling and sensitivity to LP), as a linear combination of two quantities that represent two aspects of learning: competence and change in competence. Both signals were defined for a retrospective time window of the last 15 trials played on the activity chosen at trial i (including familiarization trials early in the free-play epoch):

$$\mathrm{PC}_{i,t} = \frac{1}{15} \sum_{t'=t-15}^{t} y_{t'} \tag{4.4}$$

$$\mathrm{LP}_{i,t} = \left| \left(\frac{1}{10} \sum_{t'=t-15}^{t-5} y_{t'} \right) - \left(\frac{1}{9} \sum_{t'=t-9}^{t} y_{t'} \right) \right| \tag{4.5}$$

where $y_{t'}$ equals 1 or 0 if the participant guessed, respectively, correctly or in error at time t'. Hence, PC was defined as the proportion of correct guesses over the last 15 trials, while LP was defined as the absolute value of the difference in PC over the first 10 and the last 9 of the same stretch of 15 trials. This implementation of PC and LP signals is similar to machine learning models in (Colas et al., 2019; Linke et al., 2020; Oudeyer, Kaplan, and Hafner, 2007). In particular, one follows these computational approaches in using the absolute value of LP, which was shown to enable learners to detect tasks where performances decrease, e.g., due to forgetting, and re-gain interest to re-focus on them (Colas et al., 2019).

An individual set of parameters was estimated for each participant by minimizing the negative sum of log likelihood values over the free play trials (see (Daw et al., 2011)). Assuming that choice probabilities on a trial come from a categorical probability distribution, the likelihood of a model equals the probability (provided by the model) of the observed choice. The categorical distribution is a special case of the multinomial probability distribution, which provides the probabilities of K discrete outcomes in a single sample. Thus, the likelihood of a model that predicts choices with probabilities p_t is:

$$L(\mathbf{p}_t|\mathrm{choice}_i) = f(\mathrm{choice}_i|\mathbf{p}_t) = \prod_{j=1}^{K} p_t(\mathrm{choice}_j)^{[i=j]} \tag{4.6}$$

Where $\mathbf{p}_t$ is a vector of probabilities at time t associated with K items indexed by j, and the term $[j = i]$ evaluates to 1 when i is the activity that was chosen and to 0 otherwise. Thus, at the level of a single trial, higher likelihood is attributed to the model that assigns higher utility to the option chosen on the subsequent trial. For two and more trials, the likelihood of a model increases with the utility of the observed choices across trials. Therefore, in a maximum-likelihood model, a highly positive coefficient for a given learning signal reflects

a tendency to choose options with higher values along that signal. Conversely, a highly negative coefficient for a feature indicates a tendency to choose options that have lower values along that feature, while coefficients close to zero reflect the indifference to the feature. The total likelihood of observing all choices from a participant is given by the product of likelihoods from individual trials, $\prod_t^T L(\mathbf{p}_t|\text{choice})$. We take a logarithm of each individual trial's likelihood value in order to compute the overall model likelihood per individual as the sum of single-trial log likelihoods, $\sum_t^T \log L(\mathbf{p}_t|\text{choice})$, rather than their product. Finally, we maximized this summed likelihood by minimizing its negative value using the L-BFGS-B nonlinear numerical optimization method (Byrd et al., 1995). The same optimization tool was used in the extended computational modeling described in the next section (**??**).

Values of the estimated parameters vary not only due to different choice data between participants, but also as a function of initialization of starting values in the parameter space. Because of this variability, we estimated a model multiple times for each participant using different parameter initializations for every fit, until a convergence criterion was reached. The utility parameters were initialized from a random uniform distribution between -1 and 1, and softmax temperature was randomly sampled from [0, 100]. Convergence was reached by repeatedly fitting a model with different random initializations until 50 maximum likelihood models were found. Concretely, the algorithm updated the current "best model" each time a model better the current best was found, and stopped when it found a model just as good as the current best 50 times.

For analyses of the relation between the coefficients, instructions and choices, we normalized each coefficient pair $[w_{\text{PC}}, w_{\text{LP}}]$ by their Euclidean norm, allowing us to interpret the coefficients as relative preferences for PC and LP, respectively.

To select participants driven predominantly by PC or LP (Fig. 4.5, **c** and **d**), we categorized all participants into equally-spaced bins ($\text{bin}_1 = [-1.00, -0.33)$; $\text{bin}_2 = [-0.33, 0.33)$; $\text{bin}_3 = [0.33, 1.00]$) along each (normalized) coefficient. The PC-driven group (Fig. 4.5, **c**, left) had negative PC coefficients but near-zero influence of LP (intersection of bin_1 along PC and bin_2 along LP i.e., $\hat{w}_{\text{LP}} \approx 0$, $\hat{w}_{\text{PC}} \approx -1$), while the LP-driven group (Fig. 4.5, **c**, right) had a high preference for LP but little preference to PC ($\hat{w}_{\text{LP}} \approx 1$, $\hat{w}_{\text{PC}} \approx 0$).

4.A EXCLUSION CRITERIA

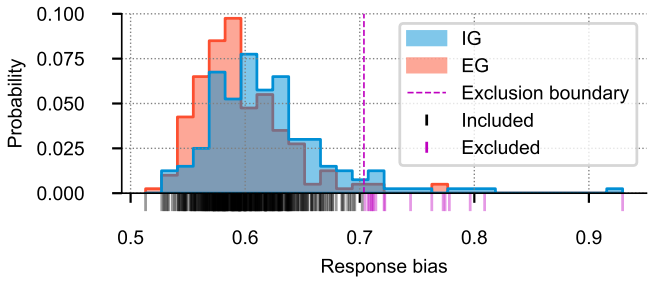

Figure 4.A.1: **Exclusion analyses.** Distribution of response bias scores ($N = 400$) used to exclude participants who were disengaged in performing the task. Vertical bars underneath the plot show individual data-points. The exclusion criterion, depicted as a vertical dashed line, was set to 2 standard deviations.

We excluded 15 participants (11 in EG and 4 in IG group) based on a response bias criterion that characterized the level of engagement in the task. Response bias was defined as:

$$\text{response bias} = \frac{1}{K}\sum_{k=1}^{K}\max(p_k, 1-p_k) \tag{4.7}$$

where $K = 4$ is the number of learning activities and $max(p_k, 1-p_k)$ denotes the relative frequency of the more frequently chosen response category in activity k. It corresponds to the participant's tendency to choose one kind of response across all trials.

Fig. 4.A.1 shows the joint distribution of response bias scores in our sample, grouped by instruction. The figure also shows the excluded participants and the exclusion criterion. The vast majority of participants were below 0.7 which corresponded to 2 standard deviations above the mean. Relative to the included participants, the excluded ones had significantly shorter reaction times to choose a category (M = 1023.89, SD = 720.770 vs M = 1472.44, SD = 360.020; $t(23.99) = -4.484$, $p < .01$, Welch two-sample test) and significantly lower difficulty-weighted final percent-correct scores (dwfPC; M = .689, SD = .090 vs M = .704, SD = .080; Welch two-sample test, $t(19.93) = -2.361$, $p = .029$), suggesting that they responded in a stereotyped fashion without being engaged in the task.

4.B SELF-REPORTED RATINGS

We collected self-reported ratings about all 4 activities at two different points of the task. Immediately after the familiarization stage (see Methods in the main article), participants were asked to report a single judgment of prospective learnability for each task:

- *Prospective learnability*: Before continuing, please rate each monster family based on how much you think you can learn about its food preferences during the rest of the task ([1] Definitely cannot learn more – [10] Definitely can learn more)

After responding to the first post-familiarization question, participants proceeded to play out the free-choice stage, after which we collected 6 additional ratings:

- *Interest*: Rate each monster family based on how much you were interested in discovering what they preferred eating ([1] Less interested – [10] More interested)

Table 4.A.1: Results of quadratic-regression fits of average SC on activity preference for each pairwise preference of a harder over easier activity.

		coef	t	p
A2 - A1	intercept	0.452	69.079	$< .01$
	pref	0.033	5.636	$< .01$
	pref^2	-0.060	-17.615	$< .01$
A3 - A1	intercept	0.426	59.860	$< .01$
	pref	0.078	12.060	$< .01$
	pref^2	-0.034	-8.470	$< .01$
A3 - A2	intercept	0.408	48.276	$< .01$
	pref	0.048	6.163	$< .01$
	pref^2	-0.016	-4.383	$< .01$
A4 - A1	intercept	0.397	62.827	$< .01$
	pref	0.126	23.322	$< .01$
	pref^2	-0.004	-1.096	$= .274$
A4 - A2	intercept	0.383	50.161	$< .01$
	pref	0.100	14.971	$< .01$
	pref^2	0.009	2.373	$= .019$
A4 - A3	intercept	0.372	44.348	$< .01$
	pref	0.058	7.736	$< .01$
	pref^2	0.02	5.232	$< .01$

Note: t-tests compare coefficient values against 0 (df = 362)

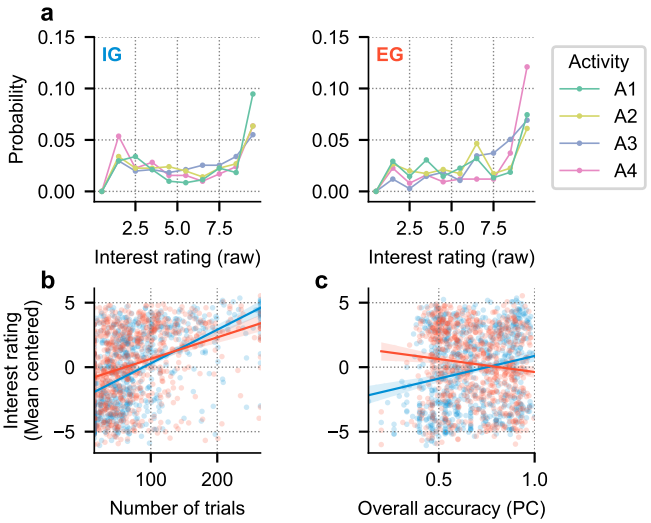

Figure 4.B.1: **Analyses of interest ratings. a**, Histograms of the raw retrospective interest ratings (1 to 10; collected after the free-play stage) for each activity in the IG (left; $N = 186$) and EG (right; $N = 196$) groups show that both groups had modes for the highest rating (10). **b**, Relationship between self-reported interest (y-axis) and number of trials for which an activity was chosen (x-axis). **c**, Relationship between self-reported interest (y-axis) and overall activity accuracy (x-axis). Note, in **b** and **c**, raw data points are presented for each instruction group (red for EG and blue for IG); regression lines are fitted separately for each group; error bands correspond to 95% confidence intervals for the linear predictions.

- *Complexity*: Rate each monster family based on how complex you thought they were ([1] Less complex – [10] More complex)
- *Rule*: Rate each monster family based on how likely you think it had a rule for food preferences ([1] Definitely no rule – [10] Definitely a rule)
- *Potential future learning*: Rate each monster family based on how much more you think you could learn if you had more time to play with it ([1] Definitely could not learn more – [10] Definitely could learn more)
- *Time spent*: Rate each monster family based on how much time you spent on them ([1] Less time – [10] More time)
- *Progress made*: Rate each monster family based on how much progress you felt you made for learning their food preferences ([1] Less progress – [10] More progress)

The subjective reports enabled us to assess how participants felt about various aspects of our task. Specifically, we were interested in two questions: and (2)

- How well did participants track their performance and choices during free exploration? Analyses of the progress ratings showed that participants had some awareness of their performance in both the EG and IG groups. Across participants and activities, self-reported Progress made was significantly correlated with true progress made (the difference between PC on the last 15 and first 15 trials on each activity; EG: $r(750) = .286,\ p < .001$; IG: $r(706) = .427,\ p < .001$). Similarly, self-reported Time spent was highly correlated with the true number of trials played (Pearson correlations, EG: $r(750) = .336,\ p < .001$; IG: $r(706) = .476,\ p < .001$). Thus, participants in both EG and IG groups accurately evaluated the relative time allocation and the progress they made across learning activities.

- How interested were they in the activities? Although participants dutifully completed the requested 250 trials of the task, they could have, in principle, reported that they were not at all interested in the activities. Contrary to this view, the distribution of Interest ratings showed a strong peak at the highest rating (10) and the average ratings were above 5 even for the activities with the lowest average ratings in each group (A4 in IG: M = 5.371, SD = 3.432, and A1 in EG: M = 5.934, SD = 3.118; Fig. 4.B.1, **a**). Importantly, interest ratings scaled with the number of trials spent on each activity above and beyond the success rates (Fig. 4.B.1, **b**). A linear regression model of mean-centered interest rating as a function of the total time spent on an activity (controlling for overall accuracy (PC over 250 trials) and the instruction received, IG vs EG), showed that ratings were reliably predicted by the actual time spent in both the IG and EG groups (slope for IG group = 7.966, $t(1454) = 15.204,\ p < .001$; interaction slope = -2.941, $t(1454) = -3.957,\ p < 0.001$). Importantly, this relation was independent of any effect of PC, suggesting that interest reflected more than mere success rates. Moreover, the correlation between PC and interest ratings was not significant in the IG group (slope = 0.379, $t(1454) = 0.646,\ p = .518$; Fig. 4.B.1, **b**), and negative in the EG group (interaction slope = -2.941, $t(1454) = -3.957,\ p < .001$; Fig. 4.B.1, **b**), suggesting that participants had an interest in the task that was above and beyond maximizing correct feedback.

4.C MASTERY POINTS

The learning criterion we present in the main text was based on people achieving 13 of 15 consecutive correct trials on an activity, which is equivalent to PC = 86.7% (probability $p = 0.0037$ of occurring by chance, assuming a binomial distribution with $n = 15$). To ensure that our conclusions were robust to choice of criterion, we repeated

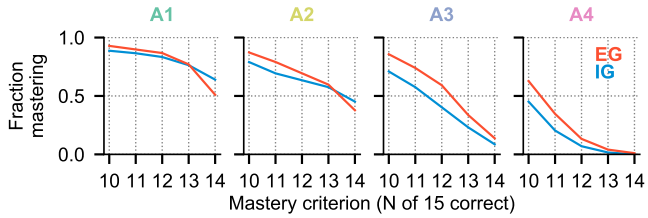

Figure 4.C.1: **Mastery criteria** Fractions of participants mastering each activity as a function of mastery criterion and group. Changing the criterion does not change the relative proportions of participants mastering each activity.

the analyses with criteria of 10, 11, 12, 13, and 14 correct trials out of 15. As expected, the fraction of people mastering each task declined as the criterion increased but, critically, the relative frequencies of the NAM designations between EG and IG groups do not change (Fig. 4.C.1). To test this, we performed a logistic regression of reaching the criterion (0 or 1) as a function of criterion and group (EG/IG). We performed a separate regression for each learning activity. We used repeated contrasts for the criterion factor to compare the fractions of participants mastering a task between the adjacent levels of the factor (i.e., comparing 10 vs 11, 11 vs 12, and so on), and regular treatment contrasts to compare fractions between groups.

The regressions produced no significant interactions between group and criterion (all $p > .05$) with only one exception: the mastery criterion of 14/15 correct was significantly less likely to be reached compared to 13/15 in the IG group (slope = -1.035, $Z(1909) = -4.155$, $p < .001$) and even less likely in the EG group (interaction slope = -0.802, $Z(1909) = -2.252$, $p < .024$). These results show that the differences between instruction groups were mostly stable over a range of criteria (as shown in Fig. 4.C.1). The positive result for the 14/15 vs 13/15 contrast shows that exceptionally high performance (PC = 94%) was more likely to be reached on the easiest task and that EG participants seemed less interested in reaching this level of accuracy. Despite this effect, these mostly nonsignificant results show that an important observation holds across a range of mastery criteria: a significant fraction of the IG group achieved mastery without being instructed to maximize learning.

4.D THE SELF-CHALLENGE INDEX

We conducted several analyses that established that our SC measure captured the tendency to choose more difficult activities (Fig. 4.D.1, **a**), did not bias our conclusions (Fig. 4.D.1, **a**), and showed the expected correlations with the model coefficients (Fig. 4.D.1, **c**).

As shown in Fig. 4.D.1 (**a**), the SC index showed a positive correlation with all possible pairwise measures of the preference for the harder activities, confirming that it measured self-challenge. However, several of these relationships were non-linear, indicating that pairwise differences do not fully capture the choices in our 4-alternative task (see Table 4.A.1 for full details on the regression fits). Specifically, the preferences for activities with moderate difficulty (A2 or A3) had an inverted U-shape trend indicating that, if these preferences were too strong, they implied lower SC by virtue of withdrawal from the most difficult activity. Similarly, the contrast of A4 vs A3 showed an upright U-shaped profile, indicating that a lower preference for A4 can correspond with higher SC if people strongly prefer A3 over A1 and A2. Thus, in our 4-alternative choice experiment, the SC index is a more parsimonious measure of the preference for challenging tasks relative to measures of preference between specific pairs of tasks.

As additional confirmation, we verified that the inverted-U relationships between SC and dwfPC shown in the main text (4) was replicated if we replaced SC with the preference for A3 or A4 (Fig. 4.D.1, **b**). In

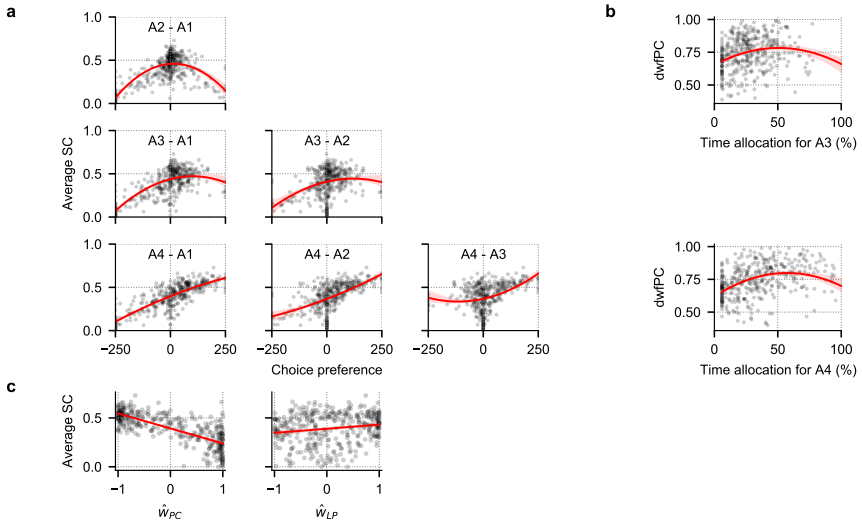

Figure 4.D.1: **Understanding the self-challenge index. a**, Correlation between behavioral preferences for harder activities (x-axes) and average SC (y-axis). Each point indicates one participant (pooled across groups: EG, $N = 188$ and IG, $N = 177$). In each panel, the x-axis is constructed so that positive values show preference for the more difficult of the two contrasted activities. **b**, Correlation between time allocation scores (x-axis) and difficulty-weighted final performance (dwfPC; y-axis). **c**, Correlations between the normalized fitted coefficients (x-axes) and average SC (y-axis). In all the panels, red lines show fits of linear-quadratic regressions with error-bands (shaded regions) indicating 95% confidence intervals (details for the fits in **a** are provided in Table 4.A.1).

case of A3's time allocation, the linear-quadratic model was better than its non-quadratic counterpart (ΔAIC = 22.238) and showcased a significant negative coefficient for the quadratic term (slope = -0.016, $t(361) = -4.979$, $p < .001$). A similar linear-quadratic model featuring time allocation for activity A4 was also better than the corresponding non-quadratic model (ΔAIC = 26.178) and likewise had a significant coefficient for its quadratic term (slope = -0.026, $t(361) = -5.383$, $p < .001$). Together, the results from Fig. 4.D.1, (**a**, **b**), demonstrate that SC served as a parsimonious measure of activity preferences and did not bias the results we report.

Finally, we examined how SC was related to the fitted (bivariate) computational-model coefficients (Fig. 4.D.1, **c**). SC was negatively correlated with w_{PC} (slope = -0.153, $t(361) = -21.999$, $p < .001$), consistent with our intuitions that choosing activities with lower PC corresponds to self-challenging choices. The regression also showed a positive correlation with w_{LP} (slope = 0.042, $t(361) = 4.954$, $p < .001$), consistent with the prediction that a sensitivity to LP guides learners to venture beyond what's easy and familiar, and choose moderately challenging activities.

4.E INDIVIDUAL MODEL FIT: A CASE STUDY

Fig. 4.E.1 demonstrates our model fitting procedure and model-based simulations for one participant's data. Panels **a** and **b** show, respectively, the participant's values of percent correct (PC) and learning progress (LP) values over time. PC and LP remained constant if the participant did not choose a task, explaining the long horizontal lines on the plots. Panel **c** shows the dynamic utility for each task based on the participant's fitted coefficients (given in the equation). Panel **d** shows the probabilities of choosing each task, simulated using the corresponding utility and the softmax function with the best-fit temperature parameter.

In this particular model, the participant's choices were characterized by a preference towards activities with high LP and low PC, which results in a model that predicted high probabilities of choosing A3 and A4 activities. Activities A1 and A2, which had consistently high PC values generated low utility and were infrequently chosen. This model captures well the transition from A4 to A3 around trial 100, where the utility of A4 started dropping as a result of a low LP signal and a corresponding plateauing of the PC signal (Fig. 4.E.1, **a**, **b**, and **d**).

4.F FAMILIARITY COMPONENT

The model comparisons from the main text show that on average, the bivariate utility function (PC + LP) explains participant's choices

better than the univariate models. The measures of PC and LP capture different aspects of competence that were hypothesized to function as intrinsic reward signals for a freely exploring learner. Different analogs of these measures have been widely used in the computational literature on intrinsically motivated learning, e.g., (Bougie and Ichise, 2019; Colas et al., 2019). These measures can be characterized as competence-based measures, because they track the information about one's competence in performing a task. There are other important families of approaches which we did not include in our study. For

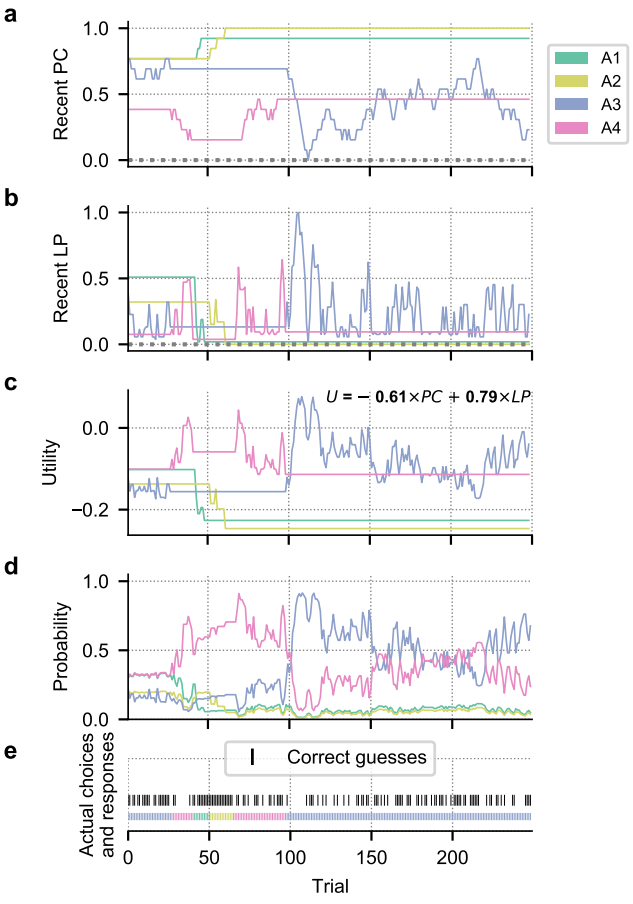

Figure 4.E.1: **Individual model fit.** The x-axis shows 250 trials of free play. Each of the first four subfigures shows the choice features of each activity through time. **a**, normalized recent percent correct (PC); b, normalized recent learning progress (LP); **c**, utility computed as a linear combination of PC and LP. The utility equation shows coefficients normalized by the Euclidean norm of the w_{PC} and w_{LP} coefficients; **d**, choice probabilities given by a softmax function at the fitted temperature parameter, $\tau = 84.98$; **e**, empirical data that the model was fitted to. The colored bar represents the observed sequences of activity choices (A3→A4→A1→A2→A4→A3) and the black vertical sticks show correct responses.

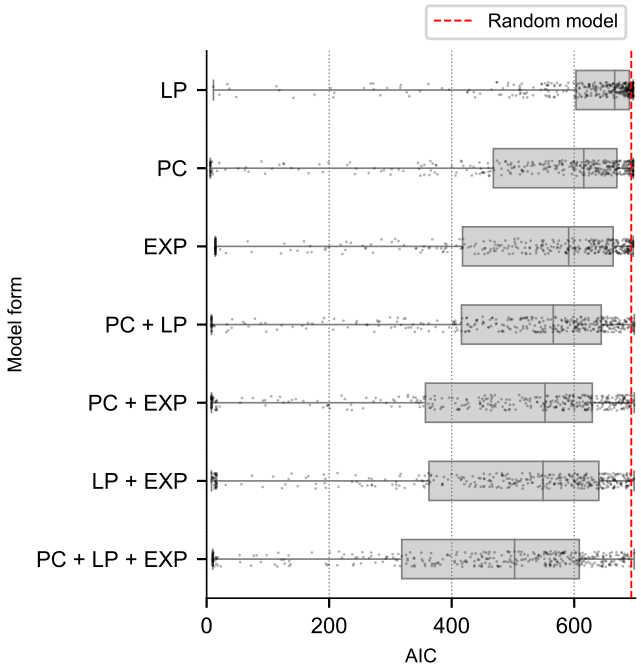

Figure 4.F.1: **Model comparisons with familiarity component.** Distributions of AIC scores for all model subsets of the full trivariate model PC + LP + exposure (EXP), where EXP represents count-based task familiarity. The box-plot box boundaries represent the 1st and the 3rd quartiles; middle bars represent sample medians; whiskers show sample minima and maxima. The random (baseline) model has AIC = 693.147 and no variance. Individual data points ($N = 365$ per model form) are shown in the overlaid strip-plots.

instance, we did not include any predictive knowledge-based measures, which would require to explicitly model the participants' beliefs about food preferences. This is an interesting direction for future work, but these kinds of models entail considerable additional complexity that is outside the scope of our investigation. However, we could test another kind of knowledge-based curiosity measure, which does not require an internal predictive model. In computational literature, this approach is referred to as a count-based, because it relies on state visitation counts (Bellemare et al., 2016). State visitation counts can be interpreted as state familiarity (the opposite of novelty).

Due to the reasoning laid out below, we did not include the measure of familiarity in the main report of model comparisons, even though it is a central idea behind some approaches to intrinsically motivated exploration. The rationale for omitting this component from the reported analyses was our focus on the roles of learning-based heuristics in the self-determined selective engagement in one of several learning activities. The familiarity measure, as defined below, is based directly on the participant's choices and as such is completely orthogonal to the

dynamics of a learner's competence. Since a single episodic sampling of a learning activity contributes to the measured familiarity of all activities equally, familiarity measured this way does not merely correlate with activity choices, it is completely determined by them. Such a measure of familiarity is a good predictor of the choice of activity, but it does not explain the choice very well. Thus, even if familiarity was an important component for the utility-based prediction (which is indeed the case), it would not be a good explanatory variable, because it itself is determined by choices.

Here we discuss a more extensive model comparisons exercise which included the additional familiarity component on top of those reported in the main text. We operationalized familiarity as exposure (EXP) to a learning activity, defined as a min-max normalized count of choice of activity. Specifically, we simply counted the number of times an activity was chosen by a participant on each trial of free play, and then re-scaled the counts to be between 0 and 1, using min-max normalization (this normalization was also applied to all PC and LP before fitting the models):

$$\text{norm}(\text{count}_{t,i}) = \frac{\text{count}_{t,i} - \min(\text{counts})}{\max(\text{counts}) - \min(\text{counts})} \tag{4.8}$$

Where $\text{count}_{t,i}$ is the number of times on which activity i was selected prior to trial t, and $\max(counts)$ and $\min(counts)$ denote, respectively, the maximum and minimum counts across all tasks and trials.

The EXP measure was added to the set of potential utility function components for all-subsets model comparisons. Fig Fig. 4.F.1. presents the distributions of AIC scores of each subset of variables included in the model. The full-form trivariate model (EXP + PC + LP) had the lowest AIC on average (M = 438.068, SD = 212.092). Thus, even when familiarity was included in the mix, both PC and LP were still important factors in increasing model likelihood. These results further support the importance of learning-based heuristics for the self-determined choice of activity. At the individual level, when compared to all of the other 6 model forms, the trivariate model (EXP + PC + LP) had the lowest AIC score in only 49.59% of participants (Fig. 4.D.1) and was at least 2 points less than any other model in only 38.90% of individuals. Moreover, the median AIC scores of the trivariate model and the next best fitting model among individuals was nonsignificant ($Z(365) = 181,\ p = .917$). These results show that although the EXP component provided some further improvement in likelihood over other models, this improvement was not very substantial: the PC + LP bivariate were, on average, significantly better than univariate models, as reported in the main text.

On the other hand, models that included both PC and LP components (i.e., EXP + PC + LP and PC + LP) had the lowest AIC in 73.42% of cases.

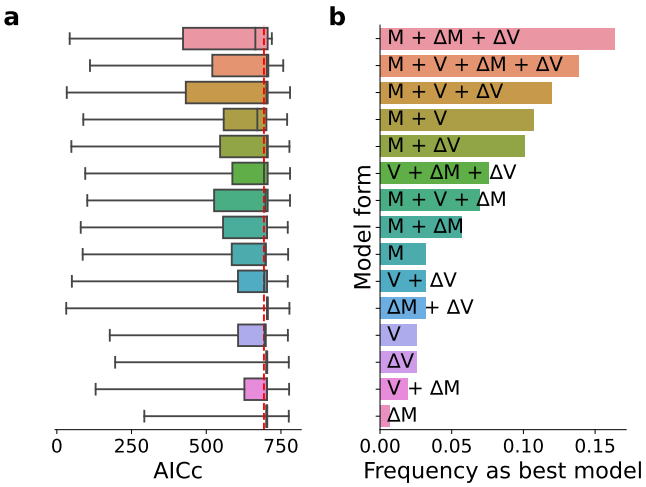

Figure 4.G.1: **Extended model comparisons.** Different colors identify model forms across the subpanels. **a, Distributions of AICc scores.** The boxes show the interquartile range, with lines inside them representing the median values. The whiskers include the entire range of each distribution. The vertical red line shows the AIC score of the random-choice model (AICc = 693.15). **b, Frequency of the best models**. The plot shows the frequencies of finding each model form as the highest-ranking model within participants.

Furthermore, compared to any univariate model – including the EXP-only model – the bivariate (PC + LP) model showed reliably better AIC scores ($Z(365) = 365$, $p = .021$, Wilcoxon signed-rank test). Notwithstanding the predictive power of the EXP component alone, PC and LP components remain important predictors of self-determined activity choices.

4.G EXTENDED MODEL EXPLORATION

The computational modeling section in the main part (Chapter 4) considers particular forms of PC and LP based on fixed-size memory of recent accuracy (see Section 4.4.2.4, METHODS/*Computational Modeling*). However, it is not obvious how much of past history humans consider when estimating their competence and progress in activities like the ones used in our study. Moreover, it is possible that other aspects of performance history (e.g., performance variability) are also relevant to choice utility. Therefore, we conducted additional analyses that explored a larger space of models. Our model space was based on a set of four variables corresponding to different types of intrinsic rewards:

- **Competence** ($M_{t|\alpha}$) defined as the exponentially weighted *mean* of correct responses over time.

- ***Change* in competence** ($\Delta M_{t|\alpha,c}$) defined as the *difference* between competence over a recent versus distant past (one form of learning progress).
- **Uncertainty** ($V_{t|\alpha}$) defined as the exponentially weighted *variance* of received feedback over time.
- ***Change* in uncertainty** ($\Delta V_{t|\alpha,c}$) defined as the *difference* between uncertainty over a recent versus distant past (another form of learning progress).

We defined competence (M) and uncertainty (V) as exponentially-weighted mean (Eq. (4.9)) and variance (Eq. (4.10)) of the incoming feedback, respectively (see **finch2009incremental**). In our context, exponential-weighting is recency-weighting: allows us to parameterize not only the amount of past history that influences an estimate, but also the extent to which older observations are discredited. Formally, we defined M and V as follows:

$$M_{t|\alpha} = M_{t-1} + \alpha(x_t - M_{t-1}) \tag{4.9}$$

$$V_{t|\alpha} = (1-\alpha)(V_{t-1} + \alpha(x_t - M_{t-1})^2) \tag{4.10}$$

Here, $\alpha \in [0,1]$ is a recency-weight parameter that controls the extent to which the latest feedback, y_t, influences the current estimates M and V. We use the "$t|\alpha$" notation to indicate that the corresponding quantity is parameterized by α. The change in each of these estimates is computed by taking the absolute difference between the current estimate (respectively, $M_{t|\alpha}$ and $V_{t|\alpha}$) and the estimates $M_{t|c\cdot\alpha}$ and $V_{t|c\cdot\alpha}$ computed with a smaller recency-weight $c \cdot \alpha$, where $c \in [0,1]$. Scaling down the α parameter by c produces estimates that are relatively more representative of the more distant past. The contrast between two estimates representing different time scales provides an estimate of the temporal derivative (or a slope) of the recent estimate. Taking the absolute value of the contrasts causes the utility model to be attracted to positive and negative changes in performance. Therefore, we can compute changes in competence (ΔM) and uncertainty (ΔV) as follows:

$$\Delta M_{t|\alpha,c} = |M_{t|\alpha} - M_{t|c\cdot\alpha}| \tag{4.11}$$

$$\Delta V_{t|\alpha,c} = |V_{t|\alpha} - V_{t|c\cdot\alpha}| \tag{4.12}$$

.

Like in the previous models, each of these four variables represents a separate utility component. These components greatly extend the

space of hypotheses about how a learner might compute a given aspect of his or her performance, not only because we consider more of them, but also because each component features one or two continuous free parameters. These parameters, of course, are not the only hypotheses we are seeking to evaluate. Our focus is on assessing what combination(s) of the four features – however they are parameterized – best explains how people self-organize their activities. As before, we define a set of linear utility functions:

$$U_{i,t} = \beta_1 M_{i,t|\alpha} + \beta_2 V_{i,t|\alpha} + \beta_3 \Delta M_{i,t|\alpha,c} + \beta_4 \Delta V_{i,t|\alpha,c} \tag{4.13}$$

where i indexes a learning activity. Equation (4.13) represents many different utility models, some of which exclude some or all of the features by settng the corresponding β coefficient to a constant value of 0. Thus, the set of 4 task-performance features gives us 16 hypothesis spaces of variable dimensionality; each hypothesis space is spanned by combinations of free parameters β as well as the recency-weighting parameter α and a scaling parameter c where it is applicable.

Since we were chiefly interested in unconstrained behavior for this set of analyses, we modeled only the IG group's choices.

We fitted each of the 16 models forms of the variable set to each participant's data (note that the null model that excludes all the utility components corresponds to a uniform-choice model with 0 parameters). The fitted models can be compared within each participant (but not across participants) using the AICc[1] scores. Figure 4.G.1 presents the distributions of AICc scores of all models, grouped by model form (Fig. 4.G.1, a) and the frequencies with which various model forms had minimum AICc scores within participants. Across participants, the most frequently encountered best model was a trivariate-utility model that included the M, ΔM and ΔV components. It was also the model with the lowest median AICc. However, other models forms were also frequently found to be the best models and in general, multivariate models with 3 to 4 components explained participants' choices better than models with fewer components.

We also compared 4 bivariate model forms with the fixed-size memory bivariate model form (PC + LP) in order to examine whether fitting the time horizons of competence and LP computations lets us fit the choices better. The 4 bivariate models from the present study included the following utility compositions: $M + \Delta M$; $M + \Delta V$; $V + \Delta M$; and $V + \Delta V$. One of these model forms had lower AICc than a fixed-size memory model in 65.41% of participants. Compared to all model forms from the extended model space (i.e. univariate, bivariate, and larger), the bivariate fixed-size memory model provided worse fit in 91.82% of participants.

1 We use AICc instead of AIC because we introduced additional parameters whilst using the same amount of data. This is a conservative precaution rather than a categorically necessary correction.

4.H MODEL COEFFICIENTS AND LEARNABILITY PREFERENCES

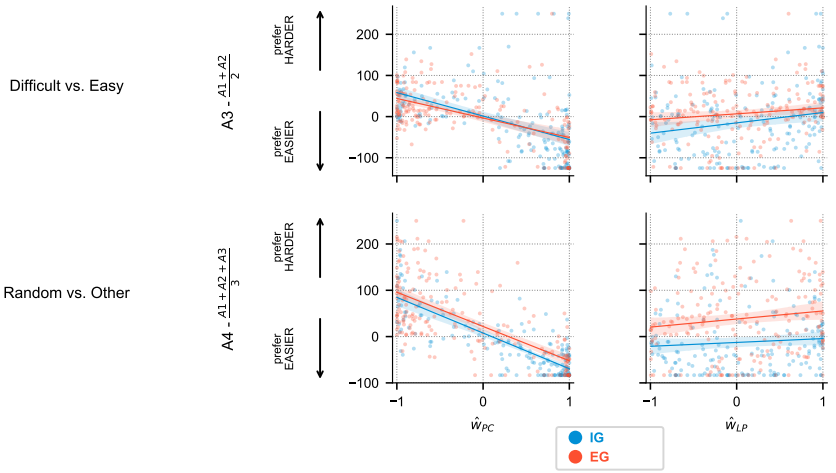

Figure 4.H.1: **Model coefficients and learnability preferences.** Each point is one participant in the IG (blue, $N = 177$) and EG (red, $N = 188$) group. The y-axis shows the difference between the number of trials a participant chose A3 minus the average number of trials spent on A1 and A2; the bottom row compares the random activity to all other activities. The x-axis shows the normalized w_{PC} and w_{LP} coefficients from the bivariate models. The lines represent linear models of activity difficulty preference as a function of normalized coefficients (each line pair fitted separately for the corresponding subplot; shaded regions represent 95% confidence intervals). w_{PC} coefficients (negative values indicating the tolerance for errors) were associated with choices of harder activities regardless of their learnability (left column). In contrast, w_{LP} coefficients (indicating more sensitivity to performance derivatives) were positively related to a preference for a harder activity only when that activity was learnable (top right), but not when it was unlearnable (bottom right).

As we discuss in the main text, PC and LP may play distinct roles in self-regulated learning. While PC can help learners identify challenging activities, LP can be used to avoid unlearnable activities. To examine this idea further, we analyzed how the w_{PC} and w_{LP} coefficients (normalized to reflect relative preferences as explained in the text) correlated with individual preferences for challenging over easier activities when the more challenging activity was, respectively, learnable or unlearnable. As a simple measure of the tendency to choose more challenging learnable activities, we computed the difference between the number of trials a participant devoted to activity A3 relative to the average amount of trials spent on easier activities (A1 and A2, Fig. 4.H.1, top row). As a measure of the tendency to choose the more challenging random activity, we computed the difference between the number of trials devoted to activity A4 relative to the average amount

of time spent on other activities (A1, A2, and A3; Fig. 4.H.1, bottom row).

The w_{PC} coefficients showed negative correlations with both measures, suggesting that they captured the participants' tendency to choose more difficult tasks regardless of learnability (Fig. 4.H.1, left). Both the preference for A3 and the preference for A4 showed negative correlations with w_{PC} (A3 vs A1&A2; IG slope = 56.969, $t(361) = -8.255$, $p < .001$; A4 vs A1-A3 slope = 86.684, $t(361) = -13.822$, $p < .001$).

In contrast, in the IG group, the w_{LP} coefficients showed a positive relationship with the preference for A3 (slope = 25.060, $t(361) = 2.817$, $p = .006$), but no relationship with the preference for A4 ($p = .356$; Fig. 4.H.1, right), suggesting that people with higher w_{LP} coefficients tended to prefer the more difficult activity only if that activity was learnable. The EG group showed no significant relationship between w_{LP} coefficients and either measure of preference.

Given the predictions of the LP hypothesis, one might expect to actually find a negative relationship between LP and a preference for an unlearnable task, not just a lack of a relationship. Indeed, one of the appeals of the LP heuristic is that it protects the learner from fixating on low-performance activities when it is not worth it. While we did not find evidence strongly supporting or refuting this prediction, we identify two ways in which it can be obtained. First, it is possible that our rather restricted operationalization of LP was not optimal for differentiating between activities A3 and A4, which were indeed very similar in terms of their recent-feedback patterns. The LP signal was especially noisy compared to the relatively clear PC signal. Investigating a wider scope of models with alternative formulations of LP could be useful for testing the predicted preference for learnable vs unlearnable tasks. Another approach would be to implement an experimental setting similar to ours, but with a larger amount of difficult learnable and unlearnable tasks. Such a setting would be more effective in showing whether sensitivity to LP helps to avoid activities that are impossible to learn.

5

METACOGNITIVE MECHANISMS OF PROGRESS JUDGMENTS

5.1 INTRODUCTION

In Chapter 2, we reviewed the basic elements and the functional diversity of computational mechanisms of intrinsically motivated exploration. We saw that exploratory sampling of learning situations through actions, goals, or social interactions enables artificial agents to efficiently diversify their experiences, learn internal models, develop rich skill-sets, and even contribute to cultural evolution. Later, in Chapter 3, we discussed why information can be inherently valuable for biological organisms and how uncertainty can motivate information-seeking. This discussion has led us to a proposition that some form of learning progress (LP, i.e., a knowledge-acquisition signal) can potentially account for many intrinsically-motivated information-seeking behaviors. Finally, in Chapter 4, we presented experimental evidence for this idea by comparing different models of trial-by-trial utility of multiple learning activities. Our results revealed great individual differences, especially when participants were not prescribed a concrete goal to pursue. Generally, activity choices were best explained by a bivariate utility model combining recent competence and change in competence. Amidst this behavioral diversity, we found that a non-trivial proportion of participants tended to engage in challenging activities and that those who spent less time "laboring in vain" (i.e. showed sensitivity to LP) achieved better results.

There are several related studies that support the idea that LP serves to motivate task engagement. A common approach (that we also adopt in our study in Chapter 4) is to measure/manipulate task difficulty and examine how it relates to task engagement. For example, Son and Metcalfe (2000) studied how self-reported ease-of-learning judgments for various texts predicted how much time and in what order participants would study them. In a later study, Gerken, Balcomb, and Minton (2011) measured objective complexity of experimental stimuli and tested its relationship with attention. Poli et al. (2020) took a step further by proposing how several candidate mechanisms processed stimuli of different complexity and examining how their outputs related to looking time. Leonard et al. (2021) manipulated performance feedback to control how children perceived their task progression and observed whether they were willing to persevere. While these studies may suggest hints for how LP can be measured objectively, they do not make explicit claims about the algorithmic implementation of the

underlying subjective computation. However, to understand human self-directed learning, we need to infer how humans represent and compute LP. As the computational literature suggests (Graves et al., 2017; Linke et al., 2020; Oudeyer and Kaplan, 2007), there are many ways LP can be computed. Not only do these computational models give rise to different patterns of exploration when simulated in their respective environments, they also make different assumptions about the underlying processes and representations (see the discussion in Chapter 6). A mechanistic understanding of LP computation implies knowing how these representations form and interact to produce LP judgments.

If we want to leverage the relationship between LP and motivation in, for example, educational or organizational contexts, understanding the (possibly biased) formation of progress judgments might be very useful. Of course, one does not absolutely need to understand how something works in order to control it, but a mechanistic understanding of a system allows making profound improvements in its operation (more efficiently). That is, we could help pupils, students, and employees get motivated by relying on heuristics gained through experience, but we could also understand more precisely how LP judgments form and how they relate to motivation. That would enable designing interventions targeting specific aspects of the process.

LP is essentially a metacognitive computation: it requires reflexive inferences about one's own knowledge/ability. Human metacognition is notoriously faulty (Fischhoff, Slovic, and Lichtenstein, 1977; Kruger and Dunning, 1999) and subject to systematic biases (Kornell and Hausman, 2017; Rozenblit and Keil, 2002; Yan, Bjork, and Bjork, 2016). Very little research has been done to investigate whether subjective LP estimation is also prone to imperfect metacognition. The only studies (to the best of our knowledge) that did investigate this problem directly (Townsend and Heit, 2011a,b) report the lack of a relationship between objective improvement and subjective improvement judgments. A related study (Kornell and Hausman, 2017) showed that participants failed to use explicit past-improvement information to inform their prediction of future performance given additional practice. Faulty metacognition challenges the LP hypothesis introduced in Chapter 3. On the one hand, there are good reasons to believe that LP-based self-regulation is optimal for learning multiple tasks (Lopes and Oudeyer, 2012; also see Dubey and Griffiths, 2020; Son and Sethi, 2006) and that humans *are* sensitive to LP Poli et al., 2020; Ten et al., 2021. On the other hand, as metacognition research suggests, LP judgments could be biased and thus ineffective or even counterproductive for efficient learning. This dissonance provides further motivation for inferring how LP gets computed.

In the next sections, we describe a pilot study that is part of a bigger research project aiming to elucidate how humans compute and

represent LP. Our approach combines two central elements. First, we attempt to emulate the naturalistic learning process unfolding during the practice of a video-game task. Video games are real-life activities that typically require time-extended practice and can be closely monitored even outside the lab. Second, using verbal questionnaires, we repeatedly query different kinds of subjective LP judgments and different aspects of participants' motivation.

Before proceeding further, we need to highlight the distinction between the so-called *retrospective* and *prospective* judgments of progress. Retrospective judgments refer to inferences about past performance, while prospective judgments reflect expectations about future performance. Presumably, these two types of LP judgments are related: retrospective assessments of a learning trajectories should contribute to future expectations.

Prospective LP judgments are likely based on a diverse set of interacting factors and processes. This reasoning is based on the theory of a closely related concept of *self-efficacy* (Bandura, 1977), defined as a belief that "one can successfully execute the behavior required to produce [certain] outcomes" (Bandura, 1977, p. 193). Thus, both self-efficacy and prospective LP reflect future expectations about task achievement. The difference is that unlike self-efficacy beliefs, prospective LP judgments involve a comparison between current and predicted task competence. This implies that prospective LP judgments could be based on self-efficacy beliefs. Bandura proposed several qualitatively different sources that contribute to one's predictive beliefs about being able to perform a task in the future, including (1) emotional arousal, (2) verbal persuasion, (3) vicarious experiences, and (4) performance accomplishments. Emotional arousal refers to feelings of stress or anxiety associated with a task. Such aversive feelings can decrease expectations of success and promote task avoidance. Verbal persuasion refers to communicative influences by social partners. For example, one can become convinced in being able to accomplish something by receiving enthusiastic support from others. Vicarious experiences refer to observations of task attempts by others. This is a highly complex category of factors because it involves social modeling and social comparisons. Finally, performance accomplishments refer to one's first-hand achievements (or failures) in a task. While Bandura's theory does not mention retrospective LP, others have proposed that it may elevate self-efficacy (Blain and Sharot, 2021).

5.2 EMPIRICAL (PILOT) STUDY OF IMPROVEMENT JUDGMENTS

One approach to study the computational tenets of metacognitive judgments of improvement is to observe how subjective self-evaluations change over the course of practicing a skill. Most of the research on metacognition, as the name might suggest, involves cognitive or

perceptual tasks and much less attention is given to self-monitoring during time-extended learning. Yet in the real world, we encounter learning activities and set goals that require relatively long-term engagement. To increase the ecological validity of the results, we want to emulate a learning process which is extended in time, interrupted by other daily activities, and does not include advanced exogenous information on one's performance. To meet these demands, we have designed a sensorimotor learning activity presented to participants as a video game, called *Lunar Lander*. The goal of the game is to guide a spacecraft (the "lander") onto a landing platform in a controlled manner, so that it does not crash on impact with the ground or go off-screen. To probe participants' subjective judgments about their performance and motivation, we solicited the corresponding verbal reports.

The goals of the pilot run of our study were (1) to assess the effects of game initialization parameters on task achievement, (2) to explore the relationships between several performance measures and improvement judgments, and (3) to explore the relationships between improvement judgments and motivation.

5.2.1 *Lunar Lander Task*

The task is based on a famous arcade video game called "Lunar Lander". Using the Box2D physics engine in JavaScript, we implemented a custom version of the game (see Fig. 5.1) in order to control the game difficulty and to be able to record the game play. Like the original, our version features a controllable spacecraft and a randomly generated uneven terrain (Fig. 5.1). The game is played across multiple trials. Within a single trial, there is a constant gravity vector that can point directly downward or be slanted to the side in order to create an impression of constant wind. We used a variable time differential to simulate the game physics. This ensured that the animation adapts to the user display's frame rate in order to keep the gameplay consistent across users (Fiedler, 2004). Thus, game states were sampled at a constant rate of 80 Hertz.

Each trial of the game ends in one of three outcomes. The spacecraft can go off-screen, in which case, the player is informed that the lander has been lost. The body of the spacecraft can make contact with the ground, in which case is informed that the lander has been crashed. Finally, if the spacecraft can be landed by being carefully placed onto the landing platform (landing pods down) and being kept there for 3 seconds, in which case the player is informed that the lander has successfully landed. The legs of the lander to which the landing pods attach are implemented as spring joints that can be compressed under a force. Thus, the momentum of the spacecraft must be controlled

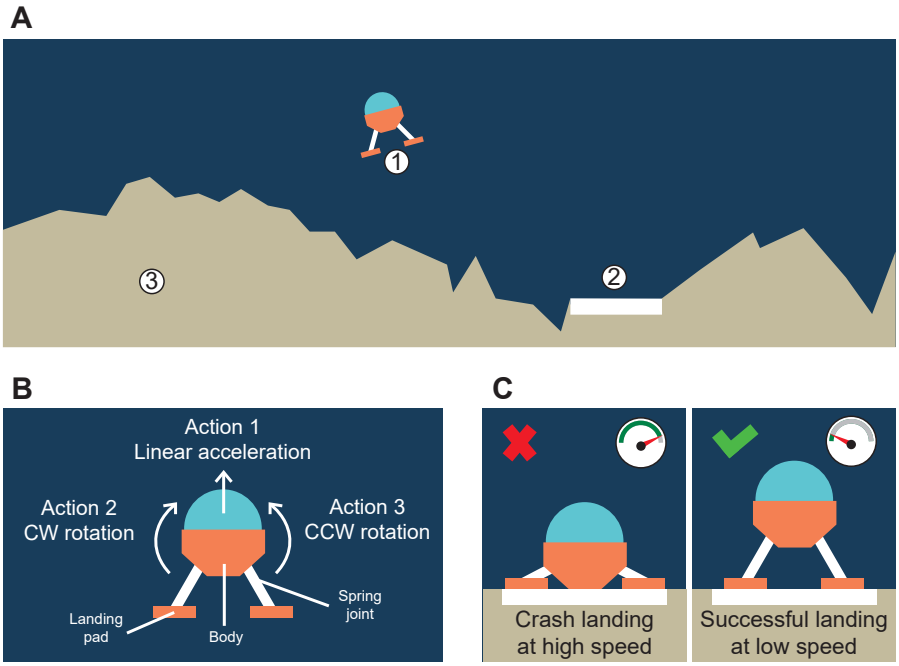

Figure 5.1: **The Lunar Lander task description. a**, A single frame from a game trial. The spacecraft (1) is controlled by the player to land it onto the platform (2) and avoid crashing into the terrain (3). The crashing event is triggered whenever the body of the lander (see **b**) collides with any other object in the environment, including the spacecraft's own landing pads. **b**, The spacecraft (consisting of the body, two spring joints, and two landing pads) can be controlled by 3 actions: linear acceleration, and clockwise/counterclockwise rotation. **c**, Successful landing requires placing the spacecraft (landing pads down) at a sufficiently low speed. Even if a player successfully drives the spacecraft to the landing platform, exceedingly high speed causes the spring joints to compress, resulting in a crash.

upon landing, even if the spacecraft descends in an upright angle; otherwise the legs will over-compress and the lander will crash.

The spacecraft can be controlled by 3 actions (4, if we count doing nothing as an action). Players can rotate the lander clockwise or counter-clockwise and propel it linearly in the direction of the longitudinal axis. Under the hood, actions apply impulse to different points on the spacecraft body. Since there is no friction, stopping or slowing down the angular motion of the body in one direction requires applying an impulse in the opposite direction. Pressing and holding an action key amounts to applying the corresponding impulse on each cycle of the physics simulation, making the spacecraft gain momentum very quickly. Mastering the game requires learning to control the spacecraft, which entails understanding the effects of actions in various contexts. Specifically, learning the game physics requires

improving predictions about the velocity of the spacecraft, given the applied actions and the existing momentum.

There are many parameters that determine the difficulty of the game. In this pilot study, we experimented with only two of them. First, we randomized the initialization distance between the lander and the platform. Traversing more space was assumed to be more challenging. We also randomized the gravity of the environment, parameterized by constant vertical and horizontal forces on the lander. By increasing the horizontal force, we created a wind effect that drags the lander to one side of the screen and not just downwards.

5.2.2 *Subjective Improvement and Motivation*

The task was practiced across three sessions, each on a different day. To measure subjective LP, we asked participants to report their judgments related to improvement after each session. Participants provided their judgments on a visually presented numerical scale. We used 1-item questionnaires to measure 4 different kinds of performance improvement. First, we asked participants to provide a subjective comparison between their current level of performance (after a session was finished) and their level of performance at the beginning of each session. Additionally, we collected prospective LP judgments by instructing participants to report how much they thought they would improve after practicing an additional session in the future. After some of the sessions, we also asked participants to compare how well they thought they did on the session that has just concluded against the session that came before it. Lastly, on the very last practice session, we asked participants to compare their performance on this final session against their performance during the very 1st session. We provide the actual questionnaire items used in our pilot study in Section 5.2.5.

It is possible that improvement judgments are based on performance indicators that can be directly observed by the experimenters and participants alike (e.g. success rate). However, it is also feasible that LP judgments are based on more privately accessible information, such as subjective feelings of competence or effort. Thus, we also tracked subjective judgments of task load, including competence, effort, and task demands. Thus, we administered the NASA task-load index (NASA-TLX) at the end of each practice session. NASA-TLX (Hart and Staveland, 1988) measures 7 components of the task load, including:

1. Mental demand
2. Physical demand
3. Temporal demand
4. Performance
5. Effort

INTRINSIC MOTIVATION MEASURES In order to assess the potential effects of LP judgments and other subjective/objective measures of performance on motivation, we used the Situated Motivated Strategies (SIMS) instrument (Guay, Vallerand, and Blanchard, 2000), which measures 4 distinct components of motivation:

1. Intrinsic Motivation (IM)
2. Identified Regulation (IR)
3. External Regulation (ER)
4. Amotivation (Am)

The SIMS-IM component assesses the extent to which an activity is evaluated as interesting, pleasant, and fun. The SIMS-IR subscale measures the importance of the activity to the individual. The score on the SIMS-ER scale indicates the extent to which the individual feels coerced into doing the activity (by external forces). Finally, SIMS-Am measures the individual's unwillingness to participate in an activity. SIMS was proposed as an alternative to the Intrinsic Motivation Inventory (IMI) (McAuley, Duncan, and Tammen, 1989), criticized for certain conceptual issues (Guay, Vallerand, and Blanchard, 2000). Despite these issues IMI components have good psychometric characteristics and can provide useful measurements beyond intrinsic motivation. While the pilot study reported below does not rely on IMI, it certainly has utility for future work.

Another tool for measuring various aspects of motivation in learning is Motivated Strategies for Learning Questionnaire (MSLQ; Duncan et al., 2015). The full version has many subscales that were not relevant for our study, but it is possible to selectively use only a subset of all scales. We were interested in measuring the following:

1. Extrinsic Goal Orientation (EGO)
2. Task Value (TV)
3. Control of Learning Beliefs (CLB)
4. Self-Efficacy for Learning and Performance (SELP)

The Motivated Strategies for Learning Questionnaire (MSLQ)-EGO subscale measures how much learners desire to accomplish goals that are separable from mastering the activity per se (e.g., demonstrating competence to others). MSLQ-TV captures the extent to which learners believe the learning or accomplishing task to be somehow beneficial for them. Task value is sometimes considered a component of the intrinsic motivation construct (e.g. McAuley, Duncan, and Tammen, 1989). By reporting MSLQ-CLB, learners indicate how much they believe that effort in learning results in mastery. Finally, MSLQ-SELP reflects the belief that a task can be mastered eventually, time constraints aside. One challenge of adopting MSLQ subscales is that they were originally designed for classroom settings. Many questions ask the respondents about courses/classes, their contents, assignments, teachings etc. Therefore, we had to adapt some of the questions to our context. Full lists of SIMS, MSLQ, and NASA-TLX questions can be found in Section 5.2.5.

In addition to the self-reported measures of motivation, we used a behavioral measure of intrinsic motivation using the free-choice technique. After finishing the "main" task (in our case, the game practice and questionnaires) participants are offered a free choice between concluding the session or engaging in additional game practice. Crucially, optional practice is not required which is communicated to the participants. Voluntary engagement in additional practice indicates intrinsic motivation.

5.2.3 *Procedure*

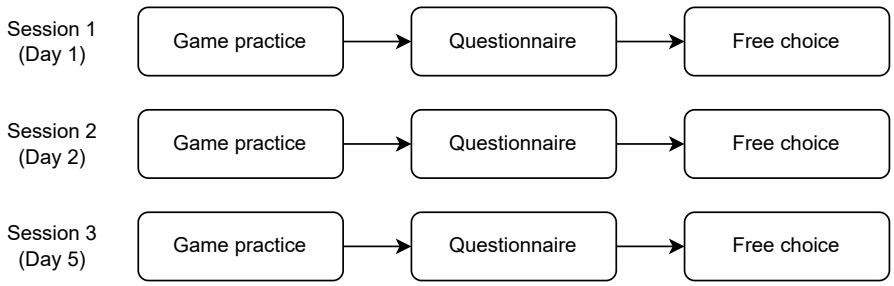

Figure 5.2: **Pilot experiment procedure.** Participants completed 3 sessions, each on a different day (top row). Each session consisted of game practice, followed by a questionnaire, and a free choice task (bottom row).

We asked participants ($N = 54$) to practice the game over 3 sessions that spanned 5 days. We used a fixed schedule for all participants: session 1 was completed on day 1, session 2 on day 2, and session 3 on day 5. Thus, there was 1 day between sessions 1 and 2; and 3 days between sessions 2 and 3. Each session consisted of 3 phases: task practice, questionnaire, and free-choice task (Fig. 5.2). The study was approved by Inria's Operational Committee for the Evaluation of Legal and Ethical Risks (OCELER).

Participants played multiple trials of Lunar Lander during the task practice phase. As a precaution from a potential floor/ceiling effect, we asked some participants to practice for 10 minutes, and others for 20 minutes, in each session. In the beginning of each session, participants read through the same instructions about the goal, the rules, and the controls of the game.

After finishing the practice phase, participants were asked to fill in a questionnaire consisting of performance-improvement, NASA-TLX, SIMS, and MSLQ questions. After finishing the questionnaire phase, participants were informed that they have finished the session and that they could practice more if they wanted or move on from our task.

5.2.4 *Findings*

5.2.4.1 *Task achievement*

To get a general sense for the difficulty of the game, we analyzed the success rates across the three practice sessions for the two session-duration conditions (10 minutes and 20 minutes). We fitted a logistic regression of binary trial outcome (success vs. crash/off-screen) as a function of session duration, session number, and their interaction (setting the 10-minute session 2 as the reference group). As shown in Fig. 5.3, success rates in 20-minute sessions were higher compared to 10-minute sessions. The logistic model predicted the odds of success in a 20-minute (second) session to be 3.924 times higher compared to a 10-minute (second) session (odds ratio (OR) $= 3.924$, 95% CI $= [3.017, 5.105]$, $z(4654) = 10.189$, $p < .001$). Success odds were also significantly different across sessions: compared to the 2nd (10-minute) session, participants were less likely to land during the 1st (10-minute) session (OR $= 0.244$, 95% CI $= [0.174, 0.343]$, $z(4654) = -8.137$, $p < .001$)

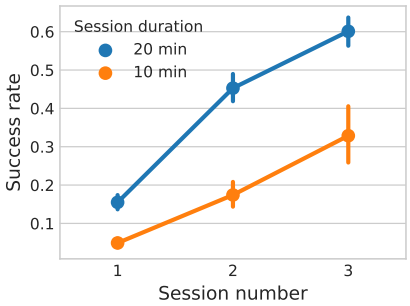

Figure 5.3: **Success rates across sessions and session durations.** The points represent success rates in the 10-minute (blue) and the 20-minute (orange) condition. Error bars show 95% CI (the increasing intervals for later sessions are due to participant dropout). Success rates increased steadily in both 10-minute and 20-minute conditions. More time per session allowed participants to perform better, but performance increased almost linearly with each successive practice session, regardless of the duration. Low levels of success in the 10-minute group suggests a floor effect.

and more likely to land in the 3rd (10-minute) session (OR $= 2.32$, 95% CI $= [1.538, 3.505]$, $z(4654) = 4.008$, $p < .001$). There was no interaction between session duration and session number, suggesting that participants improved consistently across sessions within each session-duration group. The analysis also advises against restricting the practice to 10 minutes per session, especially for the 1st session. where only about 5% of all trials were successful. This floor effect might complicate assessing the relationship between performance improvement and subjective judgments of LP.

Next, we looked at the effects of game initialization parameters on success (see Fig. 5.4). We fitted a logistic regression of trial outcome as a function of two initialization variables: the absolute wind speed and the initialization distance to the platform; we also included the trial number (cumulative across sessions) as a control variable. To compare the effects, we standardized the regression coefficients by z-scoring the covariates. We only included trials from participants who had at least one successful attempt across all sessions played. All three predictors had coefficients significantly different from zero. Thus, accounting for the increasing odds of success over time (OR $= 1.627$, 95% CI$= [1.518, 1.743]$, $z(3, 3623) = 13.871$, $p = 001$), participants were less likely to land under a stronger wind (OR $= 0.821$, 95% CI$= [0.762, 0.884]$, $z(3623) = -5.249$, $p < .001$) and when initialized farther from the target (OR $= 0.882$, 95% CI$= [0.819, 0.949]$, $z(3623) = -3.339$, $p < .001$).

The results reported in this section provide empirical validation for the predicted relationship between the two game-initialization parameters and task difficulty. Thus, it is viable to manipulate these parameters in the future, if we want to control how people learn the task. Figures 5.3 and 5.4 (C) also confirm that participants were able to improve over time (at least on a group level), making it viable to study subjective improvement judgments.

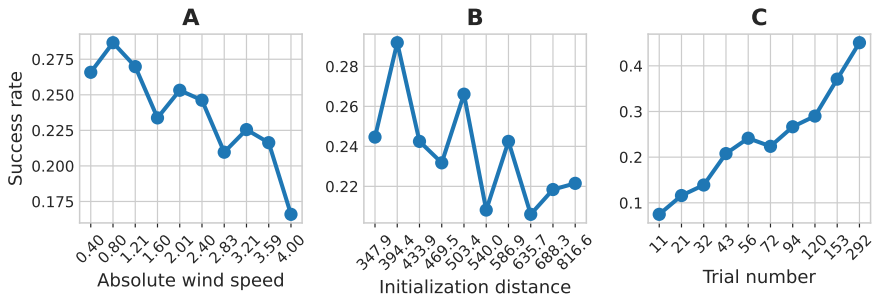

Figure 5.4: **Success rate covariates.** The three panels show the proportion of successful trials across respective covariates quantized into 10 bins (values on the x-axes show quantile upper bounds). **a** shows the negative correspondence between success rate and the absolute value of the wind parameter. **b** shows the negative relationship between success rate and initialization distance between the lander and the platform. **c** shows how average success rates increased over time across sessions. In the range of the first 300 trials, the population-level success rate appears to be increasing linearly with the number of attempts.

5.2.4.2 *Judgments of improvement*

At the end of each game-practice session, we asked participants several questions about their subjective improvement. One question probed the retrospective improvement judgment: "Rate how much your current level of performance has changed *compared to the beginning of today's session*". Participants responded by moving an interactive slider along a discrete 11-point semantic differential scale ranging between two polar response categories: "Much worse" (-5) and "Much better" (5); putting the slider at the center of the scale was assumed to indicate the reporting of no perceived change in performance. The same response scale was used to yield a prospective improvement judgment, prompted by "Rate how much you expect to improve over the next session". We explored how self-reported feelings of retrospective and prospective improvement related to different aspects of performance.

Fig. 5.5 shows the joint sample distribution between retrospective and prospective improvement judgments. There were a total of 63 observations (22 out of 85 prospective-improvement judgments were lost due to data collection error). The marginal histograms show that improvement judgments were mostly positive. However, participants judged their future improvement more variably, compared to how they thought they had previously improved, as indicated by a longer tail into the negative range of the prospective judgment scale. Spearman's correlation coefficient between prospective and retrospective judgments was moderate ($r_{\text{Spearman}}(61) = .491$, $p < .001$). Thus, retrospective feelings of improvement seem to explain some of the variance in prospective progress judgments.

Next, we investigated how the dynamics of task-achievement feedback related to the subjective judgments of improvement. For each subject and for each session, we calculated the success rate in the first and the second half of the session and then subtracted the latter from the former (a positive difference indicates improvement). Fig. 5.6 depicts relationships between

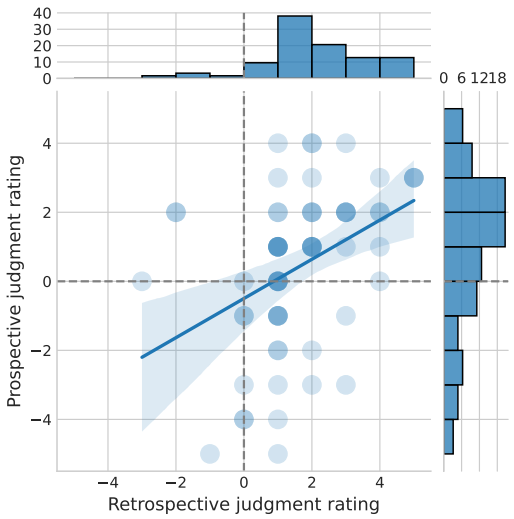

Figure 5.5: **Retrospective and prospective judgments.** The central panel shows the joint distribution of retrospective and prospective improvement judgments. Marker saturation indicates the amount of overlapping data points. The line represents a fitted linear regression of prospective judgments on retrospective judgments (the shaded area shows the 95% CI of the model). Marginal histograms on the top and right panels show relative frequencies of self-reported scores (in percentage units).

retrospective/prospective improvement judgments and changes in success rates in the corresponding session. Increasing rates of success predicted the retrospective improvement judgments, but not the prospective improvement judgments (see explanation for Fig. 5.6).

In addition to asking participants to rate their improvement within a single session, we asked them to provide more judgments about a more extended period of time. After sessions 2 and 3, we asked participants to compare their performance on the most recent session with their performance on a session just before it (improvement over consecutive sessions). After the 3rd (and last) practice session, we asked participants to also report how their subjective performance improved relative to "the start of the experiment". Due to participant dropout on sessions 2 and 3, we did not collect as much data on these judgments, yet the smaller sample sizes ($N = 39$ for session 2, and $N = 13$ for session 3) were sufficient to reveal some significant relationships between changes in objective and subjective self-assessments of competence and their corresponding subjective evaluations (Fig. 5.7, top row). Objective improvement between consecutive sessions was defined as the difference between the success rate in the session that had just concluded and the session before it (positive values indicate improvement; e.g., on session 2, this measure compares the success rate on session 2 vs session 1). Objective improvement between session 1 and 3 was defined similarly by subtracting the success rate on session 1 from the success rate on session 3. As the top row of Fig. 5.7 illustrates, participant's subjective judgments corresponded relatively well with their objective improvement. We can also see a ceiling effect for the subjective ratings, which could be obscuring the true effect size. At this point, it is difficult to say whether these longer-term

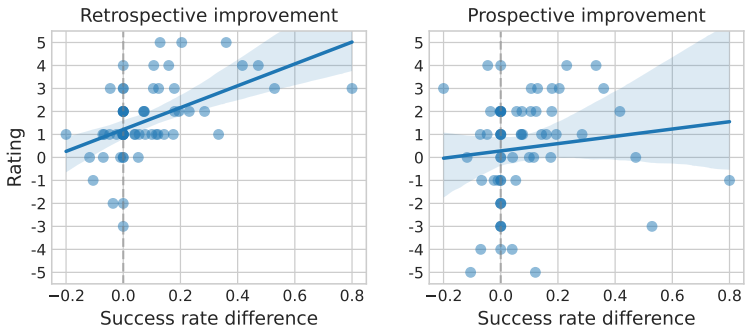

Figure 5.6: **Increasing success rate predicts self-reported past improvement, but not the expected future improvement.** The circles represent individual data points. The lines represent linear regression models of retrospective and prospective improvement ratings, respectively. The shaded regions represent 95% CI of the models' predictions. Objective performance improvement, indexed by increasing success rates, reliably predicted the retrospective judgments (slope $= 4.728$, 95% CI $= [2.298, 7.158]$, $t(83) = 3.870$, $p < .001$) but not the prospective judgments (slope $= 0.811$, 95% CI $= [-2.723, 4.345]$, $t(61) = 0.459$, $p = .648$).

judgments are better correlated with objective performance than judgments about within-session learning, but we can say that they are *at least* as well correlated. To determine the time-extent of improvement judgments that are most accurate in relation to success rate, future iterations of the study might benefit from a better-designed instrumentation and a more appropriate scale for measuring subjective improvement.

Since we did not collect subjective competence ratings *during* the self-paced practice, we could not assess the relationship between the "micro-dynamics" of subjective competence judgments within sessions and the improvement judgments at the end of each session. However, we did ask people to reflect on their perceived competence once *after* each practice session using an item from the NASA-TLX instrument. Like with success rates, we calculated two kinds of contrasts of self-reported competence scores: one comparing scores from consecutive sessions (Fig. 5.7, bottom left), and one comparing sessions 1 and 3 (Fig. 5.7, bottom right). We then explored how improvement judgments related to differences in objective and subjective measures of competence. As shown in Tab. 5.1, when considered separately, subjective and objective competence differences predicted the corresponding improvement judgments. However, when included in the same model, the results were different for the shorter term improvement judgments (consecutive sessions) compared to the longer-term ones (session 1 vs session 3). Specifically, when modeling self-reported improvement between consecutive sessions, objective difference in success rates appears to be a better predictor ($\beta = 1.003$, 95% CI = [0.282, 1.724], $t(36) = 2.821$, $p = .008$) than the difference in subjective competence judgments ($\beta = 0.384$, 95% CI = [-0.337, 1.105], $t(36) = 1.081$, $p = .287$). Neither predictor was significantly different from 0 when regressing the improvement judgments for sessions 1 and 3, but the very limited sample size for this analysis prevents us from drawing any conclusions. These results suggest that people's retrospective judgments of session-to-session improvement are better calibrated to the objective improvement than

Judgment timescale	Difference in	β	95% CI	t	p
Consecutive sessions	Success rate	1.288	[0.805, 1.771]	5.406	<.001
	Subjective competence	1.129	[0.604, 1.655]	4.359	<.001
Sessions 1 and 3	Success rate	0.989	[0.318, 1.660]	3.244	.008
	Subjective competence	0.901	[0.177, 1.624]	2.740	.019

Table 5.1: Standardized coefficient values from linear regressions predicting subjective improvement judgments (separately for 2 timescales) from differences in success rates or self-reported competence judgments. For the timescale of consecutive sessions (improvement/differences regarding performance on two consecutive sessions), the degrees of freedom for the t test are (1, 37). For the longer timescale (sessions 1 and 3) the degrees of freedom are (1, 11).

to the change in subjective competence. Interestingly, self-reported prospective improvement (predicted for the next session) was significantly correlated with the session-to-session change in subjective competence (Spearman's $\rho(23) = .445$, 95% CI = [.041, .724], $p = .026$), but not with the corresponding change in objective success rates ($p = .155$). It is not obvious why prospective judgments should be better calibrated with the change in subjective competence ratings, while retrospective judgments should be better aligned with objective improvement. This pattern of findings calls for replication and further research, especially considering past research (Townsend and Heit, 2011b) that presents results conflicting with ours.

Before proceeding to the next set of results, we would like to note that out of 46 participants, 14 (30.43%) reported positive improvement despite failing to succeed even once[1]. Some participants reported positive improvement while showing a negative change in their success rate. This could be interpreted in at least two different ways. One possibility is that while people base recent improvement judgments on success-rate dynamics, their self-reported ratings of improvement are unreliable due to metacognitive miscalibration. Another explanation is that improvement self-reports can be accounted for by something other than task-achievement feedback. Either way, there is a need to investigate the residual variation in improvement judgments that is beyond what can be explained by changing success rates or subjective performance evaluations. For if we want to assess metacognitive accuracy – we need to identify the set of valid objective measures that subjective improvement is based on, and there are no a priori reasons to assume that this set includes only the task-achievement feedback. Investigating how LP judgments arise, especially prior to witnessing task achievement, is an important direction for future research.

1 We report the results from 46 (not 54) participants, because 8 out of 54 participants dropped out immediately after completing the 1st game practice without filling in the questionnaires.

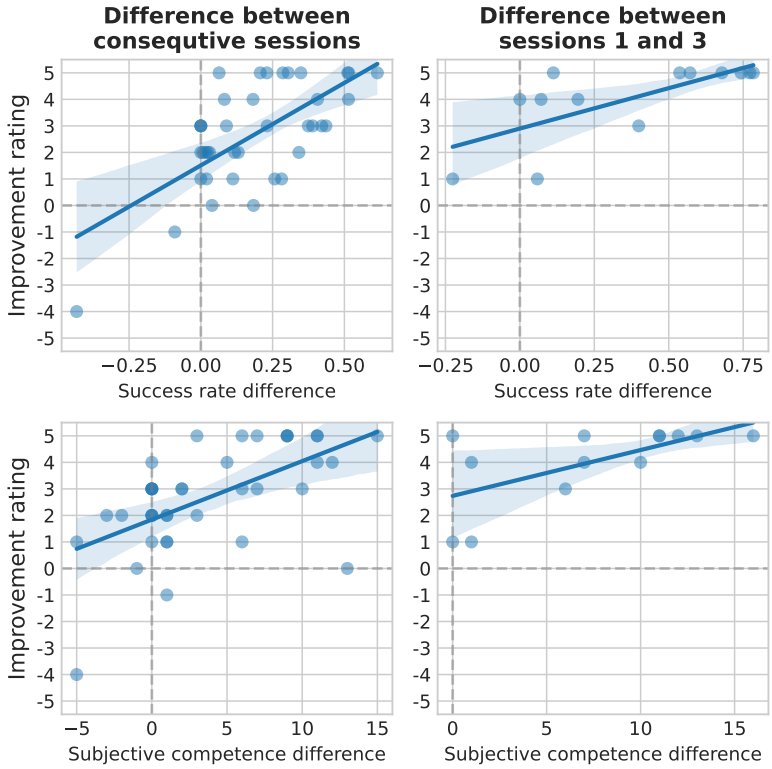

Figure 5.7: **Subjective and objective changes in competence between sessions predict self-rated improvement.** The circles represent individual data points.The lines represent linear regression models of retrospective shorter-term and longer-term improvement ratings, respectively. The shaded regions represent 95% CI of the models' predictions. The top row plots differences in success rates against the corresponding improvement ratings. The bottom row shows how difference in subjective competence evaluations relate to judgments of improvement. When considered separately, objective success-rate difference and subjective-competence difference are significantly correlated with improvement judgments, regardless of whether the difference/improvement judgment concerns consecutive sessions or sessions 1 and 3. However, when factored in together, objective success rate appears to be a better predictor for the shorter term judgment.

Here, we experimented with several performance-relevant variables. For each trial, we computed the weighted-average[2] distance, vertical speed, and horizontal speed, and then regressed these variables on the trial number in order to obtain slope coefficients describing how each variable changed throughout sessions. For instance, a negative slope for the weighted-average distance variable would indicate that the average-distance to the platform decreased over the course of the session, potentially signaling gradual improvement in performance. When considered together with success-rate difference scores, weighted-average distance and vertical/horizontal speeds were not reliably associated with judgments of improvement. This negative

2 More weight was assigned to values closer to the end of a trial

result motivates a search for performance indicators that learners might use in order to self-assess beyond success rates. An intriguing lead to pursue in the future is the subjective evaluation of sensorimotor control – people might report improvement when they feel more control.

5.2.4.3 *Motivation and Learning Beliefs*

Next, we explored how different operationalizations of LP, such as changes in objective/subjective competence and subjective improvement judgments, related to various attitudes regarding our sensorimotor game-task. Specifically, we computed Spearman's correlation coefficients between each measure of LP and different motivational and attitudinal variables measured by SIMS and MSLQ (see Fig. 5.8). As it is impossible to infer causality from these correlations, future iterations of the study will benefit from measuring motivation and learning beliefs before and after task practice. Below, we provide only speculative interpretations of the results presented by the figure.

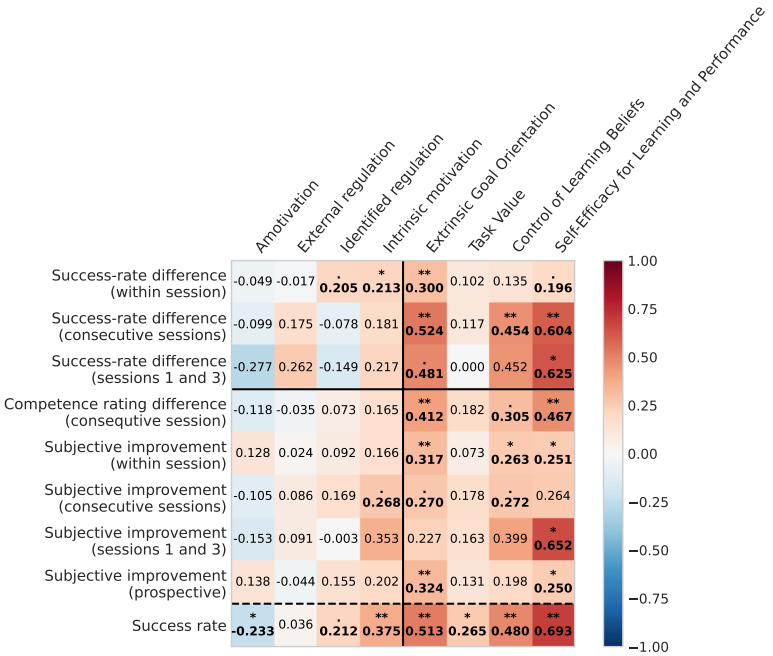

Figure 5.8: **Correlations between different operationalizations of LP and motivation.** Each cell shows the value of a Spearman's correlation coefficient. '.' indicates that the p-value for the coefficient is between .05 and .10; '*' indicates a p-value of less than .05; '**' indicates a p-value of less than .01. The solid black vertical line separates SIMS items (1st 4 columns) from MSLQ items (last 4 columns); the solid black horizontal line separates objective behavioral measures of LP from measures based on subjective ratings. The dashed horizontal line separates the plain success-rate measure from the rest of the variables since it is not a measure of LP.

We were mainly interested in the relationship between LP and intrinsic motivation. Our data did not show many significant correlations between the SIMS's Intrinsic Motivation (SIMS-IM) measure and any of our measures

of LP. Despite the lack of statistical significance, SIMS-IM scores yielded consistently positive correlation coefficients across different measures of improvement. However, even if the lack of statistical significance is due to the limited sample size, the correspondence between intrinsic motivation and subjective/objective improvement appears to be moderate at best. This result suggests that the relationship between LP and (intrinsic) motivation – if anything – is not as straightforward as sometimes conceived (e.g., Oudeyer, Gottlieb, and Lopes, 2016a).

Our measures of LP correlated more strongly and consistently with 3 out of 4 subscales from MSLQ. There were several positive correlations with the Extrinsic Goal Orientation (MSLQ-EGO) subscale, which measures the extent to which task performance is motivated by extrinsic ends (e.g. impressing or surpassing others, getting a high score). These correlations suggest that participants who wanted to do well in the game showed greater objective improvement and were largely aware of it.

Several measures of LP also predicted two distinct kinds of beliefs about learning: Control of Learning Beliefs (MSLQ-CLB) and Self-Efficacy for Learning and Performance (SELP). The first measure reflects a learner's beliefs that being successful in the task depends on his or her efforts. It could be that objective/subjective LP fuels such beliefs, but the reverse causality is also possible: believing that one is in control of one's learning could enhance learning. The same is true for the relationship between the self-efficacy and LP: improving on a task might reinforce the belief that one can eventually perform the task well, but such a belief can also invigorate the learning process. Of course, it is also possible that beliefs about the control of learning and self-efficacy have reciprocal relationships with LP whereby LP strengthens learning beliefs, which boost motivation, which contributes to LP (see Zimmerman, Schunk, and DiBenedetto, 2017, for a similar view).

Additional analyses lend some support to the reciprocity between LP on self-efficacy. First, the MSLQ-SELP rating given at the end of session 1 was a reliable predictor of the objective improvement from session 1 to session 2 (slope $= 0.070$, 95% CI $= [0.032, 0.108]$, $t(2,23) = 3.827$, $p = .001$). Second, the objective improvement from session 1 to session 2 was a reliable predictor of the MSLQ-SELP rating given after the 2nd practice session (slope $= 0.907$, 95% CI $= [0.981, 4.049]$, $t(23) = 3.391$, $p = .003$), even when controlling for the same rating from the previous session. Thus, while self-efficacy for learning beliefs predicted future improvement, improvement also predicted the updated self-efficacy beliefs, even when accounting for the old beliefs. Similar analyses did not provide the same support for the reciprocity between LP and beliefs about learning control. While the MSLQ-CLB ratings from session 1 predicted improvement from session 2 to session 1, this improvement measure failed to predict the 2nd session's ratings beyond what the 1st session's ratings predicted.

Most measures from the two motivational questionnaires were highly correlated (see Fig. 5.9). Interestingly, MSLQ-EGO did not correlate with SIMS-Am or SIMS-ER. In fact, the highest correlation coefficient between MSLQ-EGO and any SIMS measure was with SIMS-IM. MSLQ-TV showed an expected pattern of correlations (negative with SIMS-Am, positive with SIMS-IR and SIMS-IM), given that task value is sometimes considered a component of intrinsic motivation.

Perhaps most intriguingly, each of the learning-belief measures correlated *strongly* with the SIMS-IM scores (MSLQ-CLB: Spearman's $\rho(83) = 0.401$, 95%

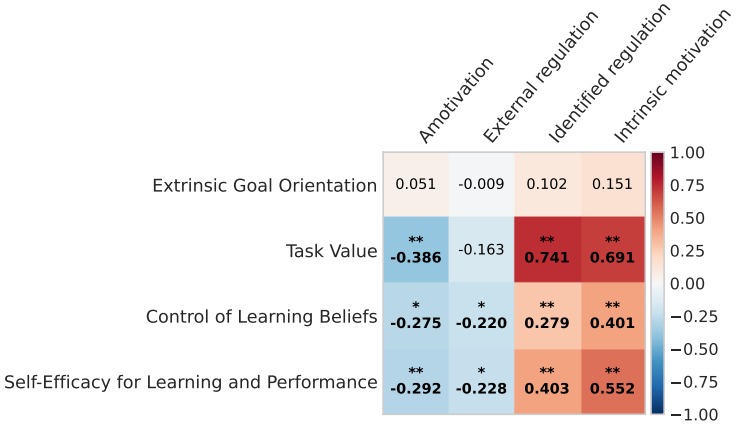

Figure 5.9: **Correlations between SIMS and MSLQ measures.** Each cell shows the value of a Spearman's correlation coefficient. '*' indicates a p-value of less than .05; '**' indicates a p-value of less than .01

CI = $[0.197, 0.571]$, $p < .001$; MSLQ-SELP: Spearman's $\rho(83) = 0.552$, 95% CI = $[0.370, 0.693]$, $p < .001$; Fig. 5.9), suggesting that learning beliefs might mediate the effect of LP on intrinsic motivation. Specifically, LP might only increase the intrinsic motivation for performing/learning a task if it feeds the belief that the task can and eventually will be learned. This could potentially explain why we failed to observe strong and reliable correlations between our measures of LP and SIMS-IM.

The last set of results concerns a strong predictor of motivation and learning beliefs that is not a measure of LP. Our data showed that plain success rates, measured as the proportion of successful trials in a session, predicted most of SIMS and MSLQ components (see last row of Fig. 5.8). Specifically, success rates were negatively correlated with amotivation (SIMS-Am; Spearman's $\rho(83) = -0.233$, 95% CI = $[-0.428, -0.018]$, $p = .031$), and positively correlated with SIMS-IM (Spearman's $\rho(83) = 0.374$, 95% CI = $[0.168, 0.550]$, $p < .001$), extrinsic goal orientation (MSLQ-EGO; Spearman's $\rho(83) = 0.513$, 95% CI = $[0.324, 0.662]$, $p < .001$), task value (MSLQ-TV; Spearman's $\rho(83) = 0.256$, 95% CI = $[0.052, 0.456]$, $p = .014$), MSLQ-CLB (Spearman's $\rho(83) = 0.480$, 95% CI = $[0.286, 0.636]$, $p < .001$), and MSLQ-SELP (Spearman's $\rho(83) = 0.693$, 95% CI = $[0.547, 0.799]$, $p < .001$).

Session-wise success rates also predicted whether a participant would accept the free choice of additional practice after finishing the main practice and the questionnaire. This was shown by a logistic regression of the event of accepting optional practice as a function of success rate (OR = 0.033, 95% CI = $[0.003, 0.395]$, $z(1, 83) = -2.689$, $p = .007$). Thus, higher success rate decreased the odds of accepting optional practice. This seems puzzling, given that success rate was positively related to motivation and learning attitudes. How could higher success rate be associated with a lesser tendency to engage in optional practice, yet predict higher subjective motivation and learning beliefs? Notably, none of the SIMS, MSLQ, or LP measures reliably predicted the acceptance of optional practice on their own. However, when considered in a multiple logistic regression *together with success rate*, several measures obtained significant coefficients, including SIMS-Am, SIMS-IR, SIMS-IM, MSLQ-TV, and MSLQ-SELP (summarized in Table 5.2). This pattern indicates while

Measure	b_{SR}	p_{SR}	b_{MA}	p_{MA}
SIMS-Am	-4.151	.003	-0.480	.027
SIMS-IR	-4.322	.004	0.562	.012
SIMS-IM	-4.837	.002	0.486	.016
MSLQ-TV	-4.567	.003	0.549	.016
MSLQ-SELP	-6.555	.000	0.631	.003

Table 5.2: Models of optional additional practice acceptance. Each row presents the coefficient estimates (and their p-values) of two predictors. One predictor is always success rate (SR) while the other predictor is a motivational/attitudinal measure (MA) indicated by the rows of the 'Measure' column.

motivational/attitudinal variables fail to account for variation in the odds of accepting optional practice, they do account for the residual variation after regressing the odds of optional practice on success rate. One plausible interpretation of this is that success rate and motivation/attitudes jointly influence behavior: if the learner performs poorly on the task, they will be compelled to engage in it if they identify with (SIMS-IR) or value (MSLQ-TV) the task, if they think the task is fun (SIMS-IM), or if they believe they can eventually perform well (MSLQ-SELP). Conversely, if a learner believes in the eventual mastery, or find the activity fun or otherwise personally important, their re-engagement will be facilitated by poor performance.

5.2.5 *Discussion*

Our pilot study pursued a number of different goals. One of them was to test our study procedure. Additionally, we aimed to assess the effects of game initialization parameters on task achievement. Task validation aside, we wanted to explore the relationships between possible performance measures and subjective improvement judgments, and subjective and objective improvement and motivation. The results provide important lessons and pose intriguing questions for future work.

TASK VALIDATION We have obtained some understanding of how several game parameters affect task-achievement rates. Manipulating difficulty is important for testing hypotheses about the causal relationships between learning dynamics and motivation/attitudes. Measuring difficulty objectively entails observing how representative groups of people perform a task with given parameters. Our results provide useful approximations of the effect sizes of distance and wind parameters on task achievement. We also gained a sense of the general learning trajectory and individual variability for the task. This knowledge can be used for manipulating group-level learning profiles in independent-group designs. For example, it would be informative to observe subjective improvement judgments and motivation to persist in groups where landing is made virtually impossible, or conversely, where the task can be mastered in very few trials.

DETERMINANTS OF IMPROVEMENT JUDGMENTS Our exploration of the determinants of subjective LP judgments showed that people might rely on the dynamics of objective success rate and/or subjective competence when verbally reporting their improvement. We hypothesized and tested several other propositions as to what information participants might use to come up with LP judgments, but failed to identify variables that would explain variation in subjective LP above and beyond what is explained by success-rate dynamics. Furthermore, changes in success rates correlated with improvement judgments better, compared to changes in subjective evaluations of self-competence. This is not fully consistent with some of the existing research (Townsend and Heit, 2011a,b) that reports the absence of correlation between changes in objective performance scores and subjective improvement judgments. There are, of course, many potentially important differences between our study and Townsend and Heit's work that can potentially account for the divergence. Most pertinently, Townsend and Heit measured objective performance (or competence) as the percentage of items recalled in the learning set, which might be far from how learners subjectively represent competence in a list-memorization task. In our task, on the other hand, success rate might be the most intuitive indicator of competence because the task requires acquiring procedural knowledge. Measures of LP based on the researcher's performance standards may differ from what people naturally consider when inferring how well they are doing and if they are improving. This points to the importance of understanding how learners evaluate their performance subjectively. Understanding the mechanistic processes behind subjective representation of competence is not only an interesting and relatively unexplored territory, but it is also key to controlling subjective judgments of improvement in applied contexts.

As we briefly mentioned earlier, self-evaluation of sensorimotor control is a viable way to gauge one's competence continuously on a sensorimotor task when binary feedback is highly skewed. We have not explored this idea in our pilot study, but it is a promising direction to follow. There are several ways to measure control. For example, we could interleave practice trials with control-testing trials where we ask participants to follow specific trajectories as closely as possible using the same game physics and controls. Alternatively, if we want to keep the learning process naturalist, we can closely track participants actions (key presses and key-press durations) together with game states (velocity vectors, lander positions) in order to measure, offline, if participants took appropriate actions in certain situations.

TEMPORAL EXTENT OF IMPROVEMENT JUDGMENTS In addition to exploring what information supports inferences about one's progress, we examined how competence changes over different temporal intervals correlated with the corresponding self-reported judgments of improvement. The original intention behind examining different time intervals was to evaluate whether judgments of some duration(s) would be better calibrated with reality than others, but our analyses failed to reveal such differences. The reported improvement judgments of different temporal sizes were similarly correlated with the corresponding objective improvement measures. Thus, when asked to describe their own learning in the past, participants seem equally good at characterizing their improvement over arbitrary time periods. It remains to be shown if there is a "basic" temporal interval which people tend to use naturally to gauge improvement for self-regulated learning (also,

if this temporal interval is flexible) or if multiple temporal intervals jointly determine an overall estimate of LP. To evaluate the credibility of these hypotheses, we might benefit from controlling the participants' learning trajectories, so that improvements computed on different temporal intervals do not agree with each other (e.g., improving within session, but getting worse compared to previous session).

SUBJECTIVE COMPETENCE, LEARNABILITY, AND TASK-ENGAGEMENT Some interesting implications followed our analyses of the interplay between metacognition, learnability beliefs, and motivation. Specifically, success rate and self-efficacy beliefs jointly predicted whether optional practice would be taken or foregone. This finding resonates with the interpretation of the bivariate choice-utility model from our previous study (see Chapter 4), where we proposed that learning progress (LP) and percent correct (PC) could have distinct roles in self-regulated learning. Judgments of competence (e.g., success rate, or PC) could serve to guide learners toward challenging tasks; expectations of improvement (e.g., self-efficacy beliefs partially determined by LP) could be used to gauge task learnability and prevent learners from laboring in vain. In line with the previous study, our pilot study suggests that the tendency to engage in optional practice correlates negatively with success rates and positively with learnability beliefs (MSLQ-SELP). That is, participants were more likely to voluntarily re-engage in the task when their performance was poor and when they thought the task could be eventually learned. This metacognitive regulation based on two kinds of signals (competence and progress) also resonates with the "Region of Proximal Learning" theory (Metcalfe and Kornell, 2005; Metcalfe, Schwartz, and Eich, 2020), which holds that information-seeking is predicated on two judgments: of whether the answer is known and whether there is enough subjective evidence for the eventual acquisition of the answer. While this theory describes the decision-making process in semantic information-seeking, it is similar to our two-factor models. In both cases, the (intrinsic) motivation to engage in the task depends on (1) whether it is believed to be achieved and (2) whether it is believed to be achievable.

SUBJECTIVE IMPROVEMENT, LEARNABILITY, AND MOTIVATION Finally, we explored how different operationalizations of LP, including objective and subjective variables, relate to motivational and attitudinal measures from SIMS and MSLQ. While LP seems to correlate rather weakly with intrinsic motivation, we found it to be a good predictor of beliefs about learning control and self-efficacy. Research in motor-learning regulation (Lewthwaite and Wulf, 2017; Wulf and Lewthwaite, 2016) convincingly shows that increasing self-efficacy and intrinsic motivation has positive effects not only on sensorimotor performance, but also on sensorimotor learning and retention. For example, selectively providing feedback information about the best versus the worst practice attempts enhances self-efficacy and intrinsic motivation during skill acquisition and results in better learning and retention of the skill (Abbas and North, 2018). Our findings are complementary to this literature. First, they suggest that better learning might strengthen the beliefs that the task can and will be learned, both of which correlate with intrinsic motivation. Second, while the cited literature studies the effects of motivation/attitudes on sensorimotor acquisition/retention proper, we

demonstrate how motivation/attitudes contribute to learning by guiding decision-making for task engagement.

Our findings about the relationships between LP, beliefs about learning control, self-efficacy, and intrinsic motivation fit well within rich theoretical frameworks of Bandura's Self-Efficacy Expectations (Bandura, 1977), Dweck's Growth Mindset (Yeager and Dweck, 2012), and Ryan and Deci's Self-Determination Theory (Ryan and Deci, 2017). We have already discussed how retrospective LP judgments relate to their prospective counterparts and self-efficacy. A similar integration has been also suggested by others Blain and Sharot, 2021. However, the link with the growth mindset theory is an original contribution of this thesis. This link was revealed by the association between LP and learning-control beliefs. Growth mindset refers to a set of beliefs that portray intellect and ability as malleable, especially with hard work and perseverance (Yeager and Dweck, 2012). One glance at the questionnaire items of MSLQ-CLB (see Section 5.B) makes it clear that this subscale measures "beliefs about effort", an important aspect of the growth mindset. Thus, LP can be conceived as a basis of learnability beliefs such as self-efficacy (for learning) expectations and growth mindset. Learnability beliefs might elevate intrinsic motivation by promising to satisfy the psychological need for competence (see Deci and Moller, 2005). That is, the belief in the effectiveness of one's toils and self-belief in the eventual task achievement imply that one can eventually become competent on that task.

In summary, in addition to validating the experimental procedure and the behavioral task, our pilot study has challenged past research, and inspired novel hypotheses about the self-feeding process of intrinsically motivated learning, which we summarize in Fig. 5.10. Pursuing future research suggested by the present discussion seems to be a promising direction toward a mechanistic understanding of intrinsically motivated learning.

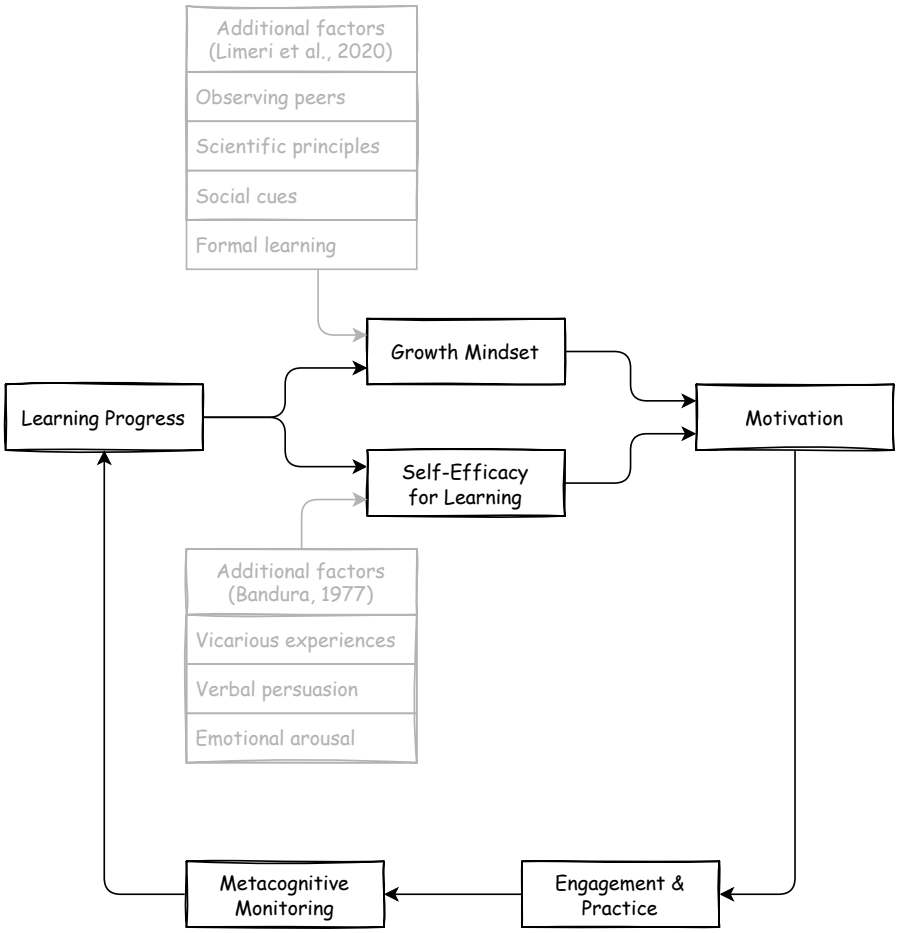

Figure 5.10: **Sketch of a cognitive-mechanistic process of competence-based intrinsically motivated learning.** Task practice is cognitively monitored to produce judgments of LP. These improvement judgments feed beliefs about the task learnability. Notably, metacognitive evaluation of the first-hand learning experience is only one of many factors (de-emphasized in gray) that can contribute to learnability beliefs (as proposed, for example, by Bandura, 1977; Limeri et al., 2020).

APPENDIX TO CHAPTER 5

Below are the questionnaire items used in the pilot study.

5.A SITUATED INTRINSIC MOTIVATION SCALE (SIMS)

Responses were recorded on a 7-point Likert scale (1="Not at all", 2="Very little", 3="A little", 4="Moderately", 5="Enough", 6="A lot", 7="Exactly") indicating the extent to which the respondent agrees with the reason for engaging in the learning activity. Specifically, the prompt for every item read: "Read each item carefully. Using the scale below, please indicate *how much each item describes the reason why you are engaged in this activity* (i.e., Lunar Lander game)".

- Amotivation
  - I do this activity but I am not sure if it is worth my time
  - I don't know; I don't see what this game brings me
  - There may be good reasons for practicing this game, but personally I don't see any
  - I keep practicing, but I am not sure I should continue
- External Regulation
  - Because I feel that I have to do it
  - Because I am supposed to do it
  - Because I don't have any choice
  - Because it is something that I have to do
- Identified Regulation
  - Because I believe that this game is important for me
  - It is for my own good
  - Because I think that this activity is good for me
  - Because I feel like playing this game
- Intrinsic Motivation
  - Because I feel good when playing this game
  - Because this game is fun
  - Because I think that this activity is pleasant
  - Because I think that this game is interesting

5.B MOTIVATED STRATEGIES FOR LEARNING QUESTIONNAIRE (MSLQ)

Responses were recorded on a 7-point semantic differential scale (1="Not at all true for me", 7="Very true for me") indicating how much the respondents agreed with a given statement. The instructions for responding to this questionnaire's items read: "The following questions ask about your **motivation for and attitudes about** practicing the Lunar Lander game. Remember *there are no right or wrong answers*, just answer as accurately as possible. Use the scale below to answer the questions. If you think the statement is very true of you, place the slider at the rightmost position (Very true for me); if a statement is not at all true of you, place the slider at the leftmost position

(Not at all true for me). If the statement is more or less true of you, place the slider somewhere in-between to best indicate how you feel."

- Extrinsic Goal Orientation
  - I want to do well in this game because it is important to show my ability to others
  - If I can, I want to get better scores in this game than most of the participants.
  - The most important thing for me is improving my overall score point average, so my main concern in this game is getting a good score
  - Getting a good score in this game is the most satisfying thing for me
- Task Value
  - Understanding the purpose of this learning activity is important to me
  - I like this kind of game
  - I think learning to play this game is useful for me
  - I am very interested in this kind of game
  - It's important for me to learn to play this game
  - I think I will be able to use what I learn in this game in other situations
- Control of Learning Beliefs
  - If I don't understand how to succeed in this game, it is because I didn't try hard enough
  - If I try hard enough, then I will understand how to succeed in the game
  - It is my own fault if I don't learn how to succeed in the game
  - If I learn in appropriate ways, then I will be able to succeed in the game
- Self-Efficacy for Learning and Performance
  - I'm certain I can master skills this game teaches
  - I expect to do well in this game
  - I'm confident I can do an excellent job in this game
  - I'm confident I can master the most complex version of this game
  - I'm confident I can learn the basic skills this game requires
  - I'm certain I can play the most difficult mode in the game
  - I believe I will achieve good results in this game

5.C NASA TASK LOAD INDEX (TLX)

Responses were recorded on a 20-point semantic differentiation scale (1="Very low", 2="Very high").

- Frustration
  - How insecure, discouraged, irritated, stressed, and annoyed were you while playing the game?
- Effort
  - How hard did you have to work to perform at your level of performance in the game?

- Performance
  - How successful were you in the game?
- Temporal demand
  - How hurried or rushed was the pace of the game?
- Physical demand
  - How physically demanding was the game?
- Mental demand
  - How mentally demanding was the game?

5.D IMPROVEMENT JUDGMENTS

Responses were recorded on an 11-point semantic differentiation scale (1="Much worse", 11="Much better") and translated into scores ranging between -5 and 5 (0 indicating no improvement).

- Within-session improvement
  - Rate how much you current level of performance has changed *compared to the beginning of today's session*
- Improvement between consecutive sessions
  - Rate how much you current level of performance has changed *compared to the previous session*
- Improvement between sessions 1 and 3
  - Rate how much you current level of performance has changed *compared to the very first session of the experiment*
- Prospective improvement
  - Rate how much you expect to improve over the next session

Part III

DISCUSSION

6

DISCUSSION

6.1 THESIS SUMMARY

The general objective of this thesis was to investigate the cognitive and motivational mechanisms enabling humans to actively explore the world efficiently. To familiarize ourselves with the scope of existing computational models, we first looked at the problem of autonomous learning in artificial agents. Not only do they need to decide how to act, and thus what to experience, in order to accomplish the task at hand, but they also need to generate, select, and pursue their own tasks. This is because genuinely autonomous machine learning systems do not have pre-specified sets of tasks granted to them. Artificial agents approach the problem by following intrinsic rewards. These rewards, in a sense, are task independent – they serve to reinforce behaviors that expand the agent's general intellectual capacity, for example, to predict, to plan, and to accomplish more tasks. We have reviewed a broad variety of intrinsically motivated AI systems and, in doing so, identified the main dimensions of variability among the existing architectures and provided a unifying framework to make sense of their diversity.

In the following chapter (Chapter 3), we provided an in-depth discussion on the functional significance of non-instrumental information-seeking in biological organisms, specifically in humans. There, we presented the idea that the evolutionary function of intrinsically motivated information-seeking is to facilitate the accumulation of (declarative and procedural) knowledge. Better knowledge allows organisms to plan actions for various tasks. Moreover, the pursuit of knowledge pushes individuals to explore novel tasks and acquire diverse repertoires of skills. Further, we identified the gap in our understanding of the ontogenetic mechanisms underlying this knowledge accumulation process. We advanced the proposition that affective/motivational states, colloquially referred to as curiosity and interest, arise in response to different kinds of uncertainty, and how non-instrumental information-seeking behaviors can be reinforced by uncertainty reduction events that signal learning progress (LP) (the Learning Progress Hypothesis, or LPH).

The experimental contributions of this thesis reported in Part ii presented original empirical evidence for the LPH (Chapter 4) and explored the effects of LP on motivation via metacognition (Chapter 5). In the former, we demonstrated that time allocation during self-regulated learning of multiple activities is best explained by a combination of factors corresponding to the current level of competence (percentage of correct responses) and recent LP. This was shown by AIC-based analyses of models fitted to trial-by-trial choices at the participant level. Our bivariate model suggested that percent correct (PC) and LP could play distinct roles in self-directed exploration (driving learners towards challenging activities and informing learners about learnability, respectively).

The (pilot) study reported in Chapter 5 was motivated by noting various shortcomings of metacognition revealed by metacognition research. Presenting LP computation as a metacognitive evaluation of one's own learning

dynamics, we wondered how can people follow this theoretically important quantity, if they might not be able to reliably represent it in the first place? Thus, we set out to study whether people are consciously aware of their LP and if so, how do they actually compute it? We looked at several potential indicators of performance, out of which the change in success rate was most strongly related to subjective improvement judgments. This finding suggests that (in a specific learning setting) people are indeed aware of their performance improvement (or deterioration) and that they might base their representation of it on success rate dynamics. Being a largely exploratory effort, this study proposed multiple novel directions, which we are excited to pursue in the future. For instance, our inclusion of measures of learning and self-efficacy beliefs helped us to hypothesize a detailed view of the process by which direct experience from practicing a task influence the motivation for the continued engagement in that task. Namely, competence progress could be a factor contributing to the learner's beliefs about his or her ability to eventually solve the task at hand. Such beliefs can be feasibly formalized, through existing belief-representation frameworks (e.g., Bayesian models), into a unifying concept of "self-model". This opens up an exciting prospect of bridging together fabulously rich and insightful strands of psychological research on self-efficacy (Bandura, 1977), self-concept (Markus and Wurf, 1987), cognitive evaluation (Deci and Ryan, 1985), and growth mindset (Yeager and Dweck, 2012).

6.2 LIMITATIONS AND FUTURE DIRECTIONS

6.2.1 *The "Free Exploration" Paradigm*

A notable contribution of this thesis is the introduction of a flexible experimental "free exploration" paradigm defined by a number of key features. First, it is crucial to enable participants to freely choose among multiple learning activities in order to capture their exploratory behavior. Second, it is important to be able to systematically measure performance in order to track the participants' learning trajectories. Finally, it is critical to manipulate both difficulty and learnability of the activities in order to examine relationships between the relevant learning dynamics and task engagement. Other parameters in the experimental setup are flexible. These degrees of freedom inspire interesting questions for future investigations. We believe that such investigations should be pursued, not only to replicate our results, but also to fill in some residual gaps.

While our task takes a step towards a more naturalistic lab setting by giving people the freedom to choose their own learning activities, it lacks several important facets of autonomous learning in real life. For example, we supply a rather small set of unique learning activities. Outside the lab, learners may need to choose between a much larger set of activities, some of which may be similar in difficulty and learnability (e.g., Baranes, Oudeyer, and Gottlieb, 2014). Future studies will benefit from showing preferences for activities associated with certain features (e.g., PC or LP) and a simultaneous indifference to activities that are similar in terms of those features. Another interesting direction that can be easily implemented in our paradigm is to allow people to learn about individual families more autonomously. People learn differently when they are able to select their own training data (Markant and Gureckis, 2013), so it is intriguing to see how self-regulated learning in

our paradigm will unfold under conditions of more autonomy. Finally, it would be very informative for the developmental literature to try out our paradigm on populations of both much younger and much older age groups.

Notably, our "free exploration" paradigm provides the means for modeling the evolution of both aleatoric and epistemic uncertainty[1]. Measures of PC and LP featuring in our choice utility model, however, provide only indirect indicators of the latent uncertainty (and its dynamics) that could underlie activity choices. All participants in our study were informed that food preferences could depend on the monsters' appearance. This detail justifies the assumption that they represented a set of hypotheses about the potential rules determining correct responses. Further assuming a particular representation of the hypothesis space (e.g., Markant and Gureckis, 2013; Tenenbaum, 1999), it is possible to model the trajectories of epistemic (which hypothesis is more likely?) and aleatoric (which food item is more likely?) uncertainties directly. Thus, a detailed modeling of the knowledge-acquisition process itself might enable us to find a stronger support for the LPH laid out in Chapter 3 (Section 3.5). Additionally, it will enable us to investigate the possible knowledge-transfer process by which insights obtained in one activity influence the hypothesis distribution in other activities. Our task can be easily extended to permit such detailed modeling of the learning process. For example, by having participants provide graded confidence ratings on their guesses (e.g., Martí et al., 2018), we can infer the latent states of the assumed hypothesis spaces.

An important criticism can be raised regarding our interpretation of the observation that many people identified the random-feedback activity as the most complex, yet rated it as interesting and spent a lot of time engaging it. We inferred that this behavior reflected a preference towards activities that are the most challenging (in terms of percent correct, or PC), rather than those on which they might be learning. However, our approach does let us rule out another possibility. Since determining whether an activity is unlearnable is a learning activity of its own, the sampling of the random-feedback activity could reflect progress-based motivation for reducing epistemic, rather than aleatoric uncertainty. That is, people might have been interested to know if the random-feedback activity is in fact random. One objection to this alternative explanation is that the random activity was sampled a lot more in the group that received an explicit instruction to learn as much as possible *about the food preferences* of the monsters, not about whether activities were solvable. Presumably, this instruction (as well as the fact that 3 out of 4 activities were rule-based) constraints the hypothesis generation process to output hypotheses about food preferences and not about whether such preferences exist. Still, we cannot be certain as to how participants interpreted our instruction; nor can we completely discredit the possibility that people were distracted from the assigned task in favor of their epistemic goal. Besides, many participants in the group without an externally prescribed goal showed a similar pattern of task-engagement. Furthermore, it seems rather plausible that had we informed these challenge-seeking participants that their favorite activity was random and unlearnable, they would lose

1 Some authors argue for the importance of further distinguishing epistemic uncertainty into parametric uncertainty (the same sense in which we use the term) and uncertainty of parametric volatility (how parameters themselves change; Payzan-LeNestour and Bossaerts, 2011). While we do not invoke the same distinction, the "free exploration" paradigm can easily incorporate activities on which the rules change over time.

their interest. This is an important limitation that should be addressed in the future.

These considerations call for a more detailed view of different kinds of learning processes (different kinds of epistemic uncertainty) involved in our task, and, perhaps, in free exploration in general (see Haber et al., 2018). The obvious learning process is learning about the food preferences. While guessing food preferences of individual monsters, participants were also forming generalizations about monster families in the form of hypothesized categorization rules. In parallel with this epistemic process, it is possible that participants were also maintaining a self-model of competence in each activity. Arguably, our model of choice utility reflects this self-learning process better than the food-preference learning process, because PC and LP are feedback-based measures that directly indicative of the level of competence, rather than the fidelity of hypotheses about food preferences per se (see our discussion below). Someone with a perfect knowledge of food preferences could, in principle, have minimal PC by always choosing the knowingly wrong option. Adopting an inclusive definition of LP as a knowledge improvement signal (as we did in Chapter 3) compels us to consider the possibility that any or both of these concurrent learning processes could supply LP and motivate activity engagement. A recent model of self-model prediction-errors can potentially explain the peculiar *Aha!* moments (Dubey et al., 2021) that, at least anecdotally, can energize epistemic pursuits. It is easy to see that the two learning processes are related: better understanding of the latent rule allows for more accurate responses. Future studies dissociating these processes will be useful in understanding their respective effects on motivation and task engagement. One approach is to directly manipulate the feedback that participants receive on the random task (e.g., increase the positive feedback rate independently of the sampled monsters). Alternatively, it is possible to provide deceptive verbal feedback independent of the accuracy feedback (e.g., "You are doing much better", when in reality the success rate stays the same).

6.2.2 *The "Lunar Lander" Experiment*

The experimental setup from Chapter 5 – let us call it the "time-extended skill acquisition" paradigm – also offers exciting opportunities for future research. The most important step would be to run another iteration of this experiment. The pilot study recruited an adequate number of participants, but many have dropped out due to the longitudinal nature of the experiment. This was partly due to the malfunctioning of the automatic reminder system in the beginning of data collection. To explain, in order to encourage full participation, we have set up a system to send email reminders to participants prior to when they would be required to join another session. For the pilot study, we decided to prevent participants, who failed to comply with the schedule, from logging into our platform, because we wanted to control the inter-session intervals. In the hindsight, this decision resulted in discarding potentially useful data which could be used in exploratory analyses such as our pilot study. Clearly, the future studies using the "time-extended skill acquisition" paradigm should sample more participants and plan ahead for the potential participant dropout.

Another important limitation of this study is in the use of novel questionnaire items for soliciting progress judgments. Because no previous work

has attempted to create a reliable psychometric instrument for measuring progress judgments, we had to use our own questions and scales. While our results support the validity of our scales (people's objective and subjective improvement correlated), we do not know the extent to which these questions/scales are reliable. This raises the need for developing a more precise psychometric tool for gauging progress judgments through verbal report. This could be done in a future study dedicated to instrument development, where many alternative formulations of questions/scales measuring the same construct are administered. Luckily, the field of psychometric measurement offers practical guidelines (Irwing, Booth, and Hughes, 2018).

6.2.3 *Mechanisms of Progress Computation*

In Chapter 5, we introduced an experimental setup that holds the potential to enable modeling subjective beliefs that form on the basis of objective performance dynamics. While the pilot study trying out this paradigm pursued its own objectives of validating the task/procedure and exploring hypotheses, the ultimate motivation is to investigate the following related questions: what information are LP judgments based on and how does the brain access and represent this information?

Computational literature offers a host of algorithms that could potentially account for the actual psychological mechanism (Graves et al., 2017; Linke et al., 2020; Oudeyer and Kaplan, 2007; Twomey and Westermann, 2018). All of these algorithms assume an interaction between two modules. One module – let's call it the *task module* – is a mechanism that learns to perform a task at hand. It serves to convert sensory/mnemonic inputs into responses (e.g., motor action, perceptual inference, categorization, recall) that satisfy certain ends. The second module – the *meta-module* – evaluates the task module in order to inform decisions pertaining to active learning and/or task selection. The meta-module is not concerned with reaching specific goal-states like task modules are, but it can be crucial for goal selection and planning.

Before considering different algorithmic approaches for metacognitive LP computation, we would like to briefly discuss what functions LP representation might serve. As we have mentioned throughout this thesis, LP can be useful as an intrinsic reward to guide the autonomous enrichment of declarative and procedural knowledge. While declarative and procedural knowledge are inextricably related (Berge and van Hezewijk, 1999), they can be distinguished in terms of what they do. Declarative knowledge describes how the world is, was, or will/would be, so it is used primarily to infer states of the world (including the body). In contrast, procedural knowledge provides imperatives for actions, so it is used to control behavior (including mental actions). Different forms of LP can be conceived (see Oudeyer and Kaplan, 2007) to specifically promote the development of either declarative or procedural knowledge. For instance, if a system is set up to value and pursue improvements in prediction, it will tend to enhance its capacity to model the world (i.e., serve an *epistemic* function), without necessarily acquiring procedural knowledge; conversely, if a system is set up to value and pursue improvements in control (i.e., serve a *pragmatic* function), it will tend to accumulate skills, without necessarily enriching its declarative knowledge as much as it could (Mirolli and Baldassarre, 2013). The following presentation of different computational mechanisms of progress estimation should be read with the aforementioned functional distinctions in mind.

There are numerous ways how the interactivity between the task-module and the meta-module can be set up for LP computation. We can identify two distinct families of mechanisms. The first family assumes that the meta-module computes LP based on the task module's performance. We will refer to this family of mechanisms as *feedback-based* mechanisms. Here, as far as the meta-module is concerned, the task module is essentially a black box, so it is the meta-module's job to infer how this black box learns by observing its achievements and failures. The second family – *introspective* mechanisms – assumes that LP is computed by observing the structural changes in the task module itself. Here, the meta-module has elevated access to the task module's "innards", so it is no longer considered a black box. This privileged access allows the meta-module to observe and quantify changes in the task module's structure as it is learning. The next section reviews a few examples representing feedback-based and introspective mechanisms, respectively.

6.2.3.1 *Introspective Mechanisms*

Introspective approaches to computing LP are based on the model of the learning process of the task module. In contrast to the feedback-based mechanisms, where the meta-module observes consequences of the task module's behavior, the introspective accounts require specifying how the task module itself adapts to its task demands. How LP is computed depends entirely on the details of the learning mechanism specified for the task module.

A commonly used approach to estimating LP in AI, is to compute a temporal derivative of the task module's error trajectory. It should be noted immediately that whereas error could be considered to be feedback, it is not the same kind of feedback used by feedback-based mechanisms. Here, the error is an essential aspect of the learning algorithm (i.e., its loss or cost) used by the task module to update its parameters. For example, in an early algorithm by Schmidhuber (1991) estimates LP as:

$$\mathrm{LP(t)} = o_C(t) - o'_C(t) \tag{6.1}$$

where $o_C(t)$ is an estimated reliability of the task module – or the "confidence" at time t of the meta-module in the task module's ability to predict; the term $o'_C(t)$ denotes the estimate of the task-module's reliability. In a nutshell, this algorithm drives the agent to explore states where the meta-module "thinks" prediction error reliability changes (Schmidhuber, 1991b).

In a similar algorithm, by Oudeyer, Kaplan, and Hafner (2007), LP is defined as:

$$\mathrm{LP(t)} = e_R(t) - e_R(t-\tau) \tag{6.2}$$

where $e_R(t)$ is the average prediction error of the task module prior to time t; the parameter τ controls the temporal reference point to which $e_R(t)$ is compared. The original algorithm also parameterizes the computation of the prediction-error averages to control their smoothness. Importantly, LP is computed separately for different regions, indexed by R, of the sensorimotor space to prevent the agent from "fabricating" progress by alternating between attempting unrelated low- and high-error tasks in an undifferentiated space. Like in Schmidhuber's algorithm, the mean error term $e_R(t)$ can also be construed as the meta-module's confidence in the task module.

Another example is a connectionist model of infant categorization from Twomey and Westermann (2018). The model features an autoencoder neural

network for compressing stimulus representations into a relatively small set of features. These compressed feature representations are not given, but have to be learned from observing stimuli – here, instances of a latent structure. When an autoencoder "understands" the latent structure well, it can encode the full representation into a compact format and then decode it back into the full representation. Latent representations are learned by adjusting connection weights in the network via backward propagation of error – a mismatch between the encoded and the decoded representation. Based on this architecture, Twomey and Westerman compared several intrinsic-motivation signals that could inform how the network should choose stimuli to learn from. One of the signals was the total amount of weight adaptation in the network. For a single weight connecting an input neuron to an output neuron, the update is given by:

$$\Delta w = (i - o)o(1 - o) \tag{6.3}$$

where i is the target activation and o is the actual activation of the output neuron. Total weight adaptation can be obtained by summing over the absolute values of individual weight updates. Since knowledge in connectionist systems resides in connection weights, this measure can be regarded as a version of LP because weights are adjusted to minimize the network's error (i.e., improve the network's knowledge). Thus:

$$\text{LP} = \sum_{w \in \boldsymbol{w}} |\Delta w| \tag{6.4}$$

where $\boldsymbol{w}$ is a vector of all weights in the network.

Yet another example of an introspective mechanism comes from a study by Graves and colleagues (Graves et al., 2017). The authors investigated the effects of different LP measures on automated curriculum learning – a process by which a meta-module autonomously selects data to train the task module. The task module was a Bayesian neural network with probabilistic weight parameters (Blundell et al., 2015). In contrast to traditional neural networks, Bayesian neural networks feature a multivariate parametric distribution for their connection weights. Instead of optimizing the weights themselves, Bayesian neural networks learn by optimizing distributional parameters. In (Graves et al., 2017), network parameters were optimized with respect to the minimum description length objective (see Graves, 2011). Based on these specifications, one version of LP – called *variational complexity gain* – was defined as a decrease in model complexity:

$$\text{LP} = D_{\text{KL}}(P_{\phi'} || Q_{\psi'}) - D_{\text{KL}}(P_{\phi} || Q_{\psi}) \tag{6.5}$$

where $D_{\text{KL}}(P_{\phi} || Q_{\psi})$ is the Kullback-Leibler divergence between the variational posterior distribution P_{ϕ} and the prior distribution Q_{ψ}. The terms $P_{\phi'}$ and $Q_{\phi'}$ refer to the posterior and prior after the projected update from the data. In the context of variational optimization of the minimum description length objective, the LP definition from above is interpreted as model complexity gain, which occurs only when data is compressed by a greater amount (Graves et al., 2017). In practice, when minimizing the variational free energy loss function, the model complexity term, $D_{\text{KL}}(P_{\phi} || Q_{\psi})$, increases in proportion to the model's ability to generalize from a learning example. That is, model complexity tends to increase when the network is able to derive generalizable knowledge from an observation.

Poli et al. (2020) studied a probabilistic model of infant attention in a task where a visual cue stochastically appeared in one of four locations

on the screen according to a fixed discrete probability distribution. Infants could learn this distribution by observing multiple cue presentations, and looked away once the distribution was learned. The authors modeled the task module as a probabilistic predictor of the cue location that optimally changed its predictions on each bout of cue presentation via sequentially updated Bayesian inference. They defined LP as the KL divergence between the prior distribution of cue locations before cue presentation and the posterior distribution after cue presentation:

$$\text{LP} = D_{\text{KL}}(p^j || p^{j-1}) \tag{6.6}$$

where p^j is the posterior distribution of the parameters determining the prediction on trial j, and p^{j-1} is the prior. This form of LP can be interpreted as the degree to which predictions of the task module change. The mechanism may seem less introspective compared to the previous two because Bayesian cognitive models are typically characterized as computational-level models – i.e. black boxes with known functions but unspecified mechanisms. However, we still classify this model as introspective, because the meta-module observes changes in the parameters of the task module, rather than the dynamics of the evaluative feedback on these parameters. What makes this LP mechanism work in practice is the fact that posterior updating never worsens the inferences of the task module (at least in the setup of the study).

6.2.3.2 *Feedback-based Mechanisms*

The temporal derivation approach has also been used to compute competence progress – an estimation of change in the agent's ability to reach its goals in a specific task space. Goal-achievement feedback is different from error feedback from the previous section, as it is not essential for learning a model for prediction. For instance, a model-free RL agent Colas et al. (2019) learned several tasks modeling environmental dynamics. The authors defined LP as follows:

$$\text{LP(n)} = |c_R(n) - c_R(n-\tau)| \tag{6.7}$$

where $c_R(n)$ is the subjectively estimated competence of the agent in a discrete task space, indexed by R; n is the number of self-evaluations performed to estimate the competence score. Subjective competence is evaluated by weighting binary goal-achievement outcomes in a task space by recency and taking the average of the weighted scores[2]. Competence is computed for all n self-evaluation trials and again for a more recent $n - \tau$ portion of these trials, and the two estimates are compared. Note that this formulation takes the absolute value of the derivative. While not essential to the definition of LP, this implementation raises the question of whether improvement and deterioration in performance are equivalent for motivation and what their differences might be. In Colas et al. (2019), taking the absolute value of the competence differential allowed the agent to actively practice tasks on which it was getting worse over time (e.g. due to forgetting), which ensured that the overall competence was maximized.

In the above approach, the agent has to infer its own competence. The resulting estimate can be interpreted as the agent's subjective belief about

2 Colas et al. (Colas et al., 2019) used a queue-based implementation, but the effect of the computation is the same as taking a recency-weighted average of a binary vector.

its competence. This interpretation is compatible with the notion that LP is a subjective belief-updating process. The above mechanism relies on point-estimate representations that do not account for belief uncertainty. On the other hand, psychological literature suggests that beliefs can have varying degrees of uncertainty. If the computation of LP involves belief comparison, then belief uncertainty should have a considerable footprint on the process.

Probabilistic beliefs (that one can accomplish a task) can be characterized by more or less confidence and change as a result of self-monitoring. The evolution of such beliefs can be expressed in terms of posterior probability. For example, suppose that an agent represents a state space, a subset of which is a state-achievement event A that has a probability of occurring, $P(A)$. This probability can be framed as the subjective belief that the event A can be reached by the agent; we might as well call it the agent's confidence in achieving A. Confidence can be updated by observing a history of state-achievement events from the past D:

$$P(A|D) = \frac{P(D|A)P(A)}{P(D|A)P(A) + P(D|A^c)P(A^c)} \tag{6.8}$$

where A^c denotes the complement of A. This equation prescribes an optimal way to update a binary state-achievement belief by combining the prior expectation $P(A)$ with the normalized likelihood of that belief $P(D|A)$. Both of these components reflect the agent's uncertain knowledge about its abilities to predict the future or reach specific goal states.

The prior confidence $P(A)$ biases how the observed data influences the posterior. For example, imagine that while estimating confidence, all that the agent observes is external binary feedback on some goal-achievement task. Now, consider the following priors and likelihood values:

Belief (B)	$P(B)$	$P(D = \text{success}\|B)$	$P(D = \text{fail}\|B)$
A	.99	.90	.10
A^c	.01	.05	.95

The posterior from a 'success' outcome will be .9994 (an increase of .0094), while a 'fail' outcome will give us a posterior of .9124 (a decrease of .0776). Thus, high prior confidence in accomplishing the task results in asymmetric updates for different outcomes. Incidentally, the surprise from observing a failure while strongly expecting success is much higher than the surprise from observing a success. The strong-expectation prior can be contrasted with the maximally uncertain prior, $P(A) = .5$: the posteriors will change by +.45 and -.40, for 'success' and 'fail' outcomes, respectively. In this case, the update is larger and relatively less asymmetric. Such dynamics are not captured by point-estimate heuristic methods.

Note that to compute the posterior $P(A|D)$, the agent needs to represent a contingency table of its past attempts. This entails storing the counts of successes and failures across beliefs that assume that A can and cannot be reached. An alternative approach is to represent subjective competence as the parameter q of the Bernoulli distribution, therefore, $P(A|q) = q$. This way, competence is an uncertain quantity that changes according to observations:

$$p(q|D) = \frac{P(D|q)p(q)}{\int_0^1 P(D|q)p(q)dq} \tag{6.9}$$

To update the prior $p(q)$, it suffices to remember the binary outcome of the most recent attempt, since the likelihood can be computed from q itself:

$P(D|q) = q^D(1-q)^{(1-D)}$; there is no need to tally up the outcomes and store the entire history of task attempts. Assuming that $p(q)$ is given by the Beta distribution, we get a well-known Beta-Binomial Bayesian model that can be readily applied to empirical data to test assumptions about prior expectations and confidence updating.

In the simple examples above, the agent only considers binary feedback data to update its beliefs. While performance feedback affects subjective confidence judgments (Martí et al., 2018; Rouault, Dayan, and Fleming, 2019), other factors relating to task performance might be at play, especially when external feedback is sparse (e.g., Holm, Wadenholt, and Schrater, 2019; Locke, Mamassian, and Landy, 2020; Rouault, Dayan, and Fleming, 2019) or heavily skewed (e.g. receiving only negative feedback). For instance, when trying to answer a question, one might consider the utility of self-generated candidate answers or the amount of question-cued information in order to judge whether one is getting close to the answer (see Coenen, Nelson, and Gureckis, 2019). Feelings of knowing (Koriat, 1993), tip-of-the-tongue states (Schwartz and Metcalfe, 2011), and *Aha!* moments (Dubey et al., 2021) are good examples of people estimating how close they are to accomplishing a task before it is accomplished. Other examples include complex sensorimotor skills (e.g., juggling), in which it is useful to be able to track one's proximity to the desired behavior. In a recent visuomotor task, participants tracked the invisible center of a flickering dot-cloud (Locke, Mamassian, and Landy, 2020). The authors showed that participants monitored the distance between the target and the cursor to make judgments about their performance. Such continuous evaluations are useful for assessing one's progress when binary feedback is not available or skewed. This implies that feelings of progress may be supported not only by monitoring the success rate, but also the proximity to success.

6.2.3.3 *Open Challenges*

Given the diversity of computational mechanisms of LP, which approaches can we expect to be better suited for cognitive modeling? While not necessarily incompatible with introspective mechanisms, feedback-based approaches seem to present a stronger case for warranting further investigation for a number of reasons. First, as we hope our examples demonstrate, models based on introspective mechanisms require making non-trivial assumptions about the learning mechanism of the task module. This is not an inherent flaw, but it makes these models more difficult to justify and interpret. Second, we know that people rely on exogenous achievement feedback during competence evaluation (Martí et al., 2018, also, our study in Chapter 4). While self-evaluation without feedback is possible and often a reality, receiving it makes us more confident in our evaluations (Rouault, Dayan, and Fleming, 2019). Accordingly, people often actively seek out performance feedback, even when it is costly (FitzGibbon, Komiya, and Murayama, 2021; Holm, Wadenholt, and Schrater, 2019). Moreover, when clear binary achievement feedback is not available, people use performance correlates that are available to infer their competence (Locke, Mamassian, and Landy, 2020). Finally, feedback-based metacognition explains how procedural knowledge – which is tacit, and thus difficult to introspect directly – is evaluated.

The consideration of feedback-based mechanisms above raises two important questions. First, how do people determine task-achievement parameters and set competence standards? None of the models discussed so far specify

how to select performance parameters for progress monitoring, yet it is crucial to understand this process if we want to unravel how people evaluate their performance and LP, when there is no useful normative feedback. The question is especially poignant in the context of complex tasks encountered outside psychology labs and attempted very few times in a lifetime (earning a Ph.D. is a perfect example). Self-evaluation and progress estimation is not necessarily easy (Kornell and Hausman, 2017; Raaijmakers et al., 2019; Townsend and Heit, 2011a,b), yet these abilities seem crucial for self-regulated learning, particularly at a young age (Oudeyer, 2018). Understanding how representations of competence standards form could help us explain the mismatch between the theorized importance and the apparent difficulty of accurate self-assessment for self-regulation. Considering the potentially idiosyncratic nature of self-assessment across individuals and situations (Boekaerts, 1991), it is feasible that some of the past metacognition work has operationalized LP differently from how it might be represented by individuals, and thus failed to find an association between objective measurement of progress and subjective judgments.

The problem of lacking reliable feedback is also at the heart of intrinsically motivated, machine learning (ML) (see Chapter 2; see also Oudeyer, 2018) where the proposed solution is to provide the agent with intrinsic reward functions that support learning in the absence of primary rewards. Such intrinsic-reward functions, however, are designed to be task-independent, for they are intended to enhance the agent's competence in a general way. On the other hand, evaluation of the proximity to task achievement is tied to the task itself. Understanding how task-specific goal-proximity evaluation can be flexibly deployed across different tasks can be especially useful in autotelic agents that generate and pursue their own tasks (Colas et al., 2021a).

The second question concerns the temporal extent of progress judgments. To illustrate, consider the process of learning a complex skill. As discussed in the introduction, one might expect to improve based on many considerations, including the retrospective feelings of progress – a comparison between the present level of performance (i.e., competence) and a reference point in the past. The question raised here, is how far back does the reference point go? Setting the reference point too far back may bias the progress estimate: it will signal positive progress even if performance stagnates; setting the reference point too close to the current estimate may produce a noisy and unreliable LP signal (see Fig. 6.1). We have seen examples where LP on a given task is computed across a fixed window of time, however, other approaches are possible.

Fixed time-window computation might be too restrictive to account for the diversity of learning trajectories across different tasks. Shorter time windows are more useful for easier tasks where learning progresses rapidly, while longer time windows are more appropriate for more slowly developing skills. Fixed time-window computation also requires ad hoc assumptions to handle situations in which the reference point extends beyond what is available. For example, given a time window of size τ, the learner would require at least τ performance evaluations to compute LP, unless the parameter is allowed to vary in the beginning. A more flexible approach would be to reset the reference point to whenever the task is switched to. Such an approach raises the question of how task-disengagement is decided. It turns the relationship between LP and its temporal extent upside down: instead of LP depending on the fixed time window, the temporal extent of LP judgments would depend

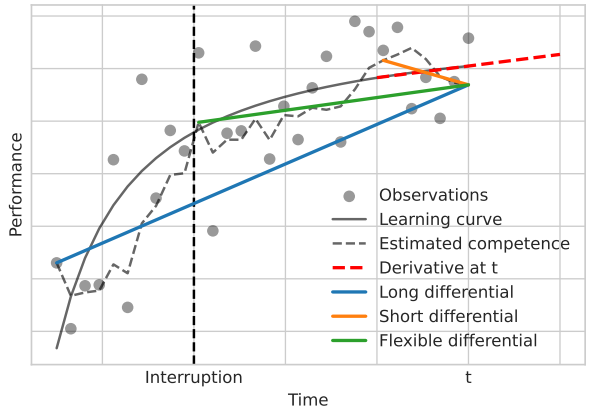

Figure 6.1: **Temporal extent of LP estimation**. The plot depicts a hypothetical learning curve on which performance increases with diminishing returns over time. The derivative of this curve represents the true rate of improvement. The learner does not know neither the true learning curve nor its derivative. However, learners compare the estimated performance values at different points in time. Here, the dashed curve represents the box-car estimated learning curve derived from noisy observations (black dots). Note that learning is interrupted for an unspecified period of time (dashed vertical line). Comparing current performance with a very distant reference point is likely to overestimate the derivative (blue line). Comparing it to reference points that are too close in time is unreliable, given the noise in observations (orange line). Resetting the reference point to the beginning of the current episode is more likely to be accurate, if enough attempts are made during the episode (green line).

on the rate of learning (assuming that low LP signals the need to disengage from the current task). This temporally flexible LP estimation approach is assumed by the psychological *Region of Proximal Learning* theory (Metcalfe and Kornell, 2005), which proposes that once a task is chosen, the amount of time spent on the task will depend on subjective LP defined as the difference in competence at the beginning and the end of a learning episode.

6.3 CONCLUDING REMARKS

Having discussed our results and the residual unknowns, it is evident that we have only started to scratch the surface of what is going on during intrinsically motivated information-seeking. We are still far removed from a complete understanding of the interaction between learning and motivation. We hope that this closing chapter has highlighted some promising directions for future work. An overarching theme of is the subjectivity (or individuality) surrounding the process of belief formation during self-regulated learning. Going forward, we need to pay close attention to multiple learning processes that learners might track as they are learning; we need to be aware of the idiosyncratic goals they might pursue; and we need to consider different

kinds of information they might use to make inferences about their learning dynamics.

Abbas, Z. A. and J. S. North (2018). "Good-vs. poor-trial feedback in motor learning: The role of self-efficacy and intrinsic motivation across levels of task difficulty." en. In: *Learning and Instruction* 55, pp. 105–112. DOI: 10.1016/j.learninstruc.2017.09.009. URL: https://www.sciencedirect.com/science/article/pii/S0959475216300950 (visited on 01/11/2022).

Adriaans, Pieter and Johan van Benthem (2008). "Introduction: Information is what information does." en. In: *Philosophy of Information*. Elsevier, pp. 3–26. DOI: 10.1016/B978-0-444-51726-5.50006-6. URL: https://linkinghub.elsevier.com/retrieve/pii/B9780444517265500066 (visited on 10/04/2021).

Anderson, Eric C. et al. (2019). "The Relationship Between Uncertainty and Affect." In: *Frontiers in Psychology* 10, p. 2504. DOI: 10.3389/fpsyg.2019.02504. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6861361/ (visited on 10/18/2021).

Andreae, John Hugh (1977). *Thinking with the Teachable Machine*. en. Google-Books-ID: 4cdfAAAAMAAJ. Academic Press.

Andrychowicz, Marcin et al. (2018). "Hindsight Experience Replay." In: *arXiv:1707.01495 [cs]*. arXiv: 1707.01495. URL: http://arxiv.org/abs/1707.01495 (visited on 11/24/2021).

Aron, A. R. et al. (2004). "Human Midbrain Sensitivity to Cognitive Feedback and Uncertainty During Classification Learning." en. In: *Journal of Neurophysiology* 92.2, pp. 1144–1152. DOI: 10.1152/jn.01209.2003. URL: https://www.physiology.org/doi/10.1152/jn.01209.2003 (visited on 10/25/2021).

Aubret, Arthur, Laetitia Matignon, and Salima Hassas (2019). "A survey on intrinsic motivation in reinforcement learning." In: *arXiv:1908.06976 [cs]*. arXiv: 1908.06976. URL: http://arxiv.org/abs/1908.06976 (visited on 11/24/2021).

Averbeck, Bruno B. (2015). "Theory of Choice in Bandit, Information Sampling and Foraging Tasks." en. In: *PLOS Computational Biology* 11.3. Publisher: Public Library of Science, e1004164. DOI: 10.1371/journal.pcbi.1004164. URL: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004164 (visited on 11/03/2021).

Baker, Bowen et al. (2020). "Emergent Tool Use From Multi-Agent Autocurricula." In: tex.ids: Baker2019 arXiv: 1909.07528. URL: https://openreview.net/forum?id=SkxpxJBKwS (visited on 05/18/2020).

Bandura, Albert (1977). "Self-efficacy: Toward a unifying theory of behavioral change." In: *Psychological Review* 84.2. Place: US Publisher: American Psychological Association, pp. 191–215. DOI: 10.1037/0033-295X.84.2.191.

Baranes, Adrien F., Pierre-Yves Oudeyer, and Jacqueline Gottlieb (2014). "The effects of task difficulty, novelty and the size of the search space on intrinsically motivated exploration." en. In: *Frontiers in Neuroscience* 8. DOI: 10.3389/fnins.2014.00317. URL: http://journal.frontiersin.org/article/10.3389/fnins.2014.00317/abstract (visited on 04/09/2021).

Baranes, Adrien (2013). "Active learning of inverse models with intrinsically motivated goal exploration in robots." en. In: *Robotics and Autonomous Systems*, p. 25.

Baranes, Adrien and Pierre-Yves Oudeyer (2009). "R-IAC: Robust Intrinsically Motivated Exploration and Active Learning." In: *IEEE Transactions on Autonomous Mental Development* 1.3. Conference Name: IEEE Transactions on Autonomous Mental Development, pp. 155–169. DOI: 10.1109/TAMD.2009.2037513.

Baranes, Adrien, Pierre-Yves Oudeyer, and Jacqueline Gottlieb (2015). "Eye movements reveal epistemic curiosity in human observers." en. In: *Vision Research* 117, pp. 81–90. DOI: 10.1016/j.visres.2015.10.009. URL: https://linkinghub.elsevier.com/retrieve/pii/S0042698915003430 (visited on 01/22/2021).

Barron, Andrew B. et al. (2015). "Embracing multiple definitions of learning." en. In: *Trends in Neurosciences* 38.7, pp. 405–407. DOI: 10.1016/j.tins.2015.04.008. URL: https://linkinghub.elsevier.com/retrieve/pii/S0166223615000983 (visited on 11/03/2021).

Barto, Andrew, Marco Mirolli, and Gianluca Baldassarre (2013). "Novelty or Surprise?" In: *Frontiers in Psychology* 4, p. 907. DOI: 10.3389/fpsyg.2013.00907. URL: https://www.frontiersin.org/article/10.3389/fpsyg.2013.00907 (visited on 10/01/2021).

Bazhydai, Marina, Katherine Twomey, and Gert Westermann (2020). "Curiosity and Exploration." en. In: *Encyclopedia of Infant and Early Childhood Development*. Elsevier, pp. 370–378. DOI: 10.1016/B978-0-12-809324-5.05804-1. URL: https://linkinghub.elsevier.com/retrieve/pii/B9780128093245058041 (visited on 03/12/2021).

Bellemare, Marc et al. (2016). "Unifying count-based exploration and intrinsic motivation." In: *Advances in Neural Information Processing Systems*, pp. 1471–1479.

Benureau, Fabien C. Y. and Pierre-Yves Oudeyer (2016). "Behavioral Diversity Generation in Autonomous Exploration through Reuse of Past Experience." In: *Frontiers in Robotics and AI* 3, p. 8. DOI: 10.3389/frobt.2016.00008. URL: https://www.frontiersin.org/article/10.3389/frobt.2016.00008 (visited on 11/24/2021).

Berge, Timon ten and Rene van Hezewijk (1999). "Procedural and Declarative Knowledge: An Evolutionary Perspective." en. In: *Theory & Psychology* 9.5. Publisher: SAGE Publications Ltd, pp. 605–624. DOI: 10.1177/0959354399095002. URL: https://doi.org/10.1177/0959354399095002 (visited on 10/18/2021).

Berlyne, D. E. (1958). "The influence of complexity and novelty in visual figures on orienting responses." In: *Journal of Experimental Psychology* 55.3. Place: US Publisher: American Psychological Association, pp. 289–296. DOI: 10.1037/h0043555.

Berlyne, D E (1966). "Curiosity and Exploration." en. In: 153, p. 9.

Berlyne, Daniel E. (1954a). "A theory of human curiosity." en. In: *British Journal of Psychology. General Section* 45.3, pp. 180–191. DOI: 10.1111/j.2044-8295.1954.tb01243.x. URL: https://onlinelibrary.wiley.com/doi/10.1111/j.2044-8295.1954.tb01243.x (visited on 09/06/2021).

– (1954b). "An experimental study of human curiosity." en. In: *British Journal of Psychology. General Section* 45.4, pp. 256–265. DOI: 10.1111/j.2044-8295.1954.tb01253.x. URL: https://onlinelibrary.wiley.com/doi/10.1111/j.2044-8295.1954.tb01253.x (visited on 10/22/2021).

– (1957). "Uncertainty and conflict: A point of contact between information-theory and behavior-theory concepts." en. In: *Psychological Review* 64.6, Pt.1, pp. 329–339. DOI: 10.1037/h0041135. URL: http://doi.apa.org/getdoi.cfm?doi=10.1037/h0041135 (visited on 10/22/2021).

Berlyne, Daniel E (1960). "Conflict, arousal, and curiosity." In.

Berridge, Kent C. (2007). "The debate over dopamine's role in reward: the case for incentive salience." eng. In: *Psychopharmacology* 191.3, pp. 391–431. DOI: 10.1007/s00213-006-0578-x.

– (2009). "Wanting and Liking: Observations from the Neuroscience and Psychology Laboratory." In: *Inquiry (Oslo, Norway)* 52.4, p. 378. DOI: 10.1080/00201740903087359. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2813042/ (visited on 10/24/2021).

– (2018). "Evolving Concepts of Emotion and Motivation." In: *Frontiers in Psychology* 9, p. 1647. DOI: 10.3389/fpsyg.2018.01647. URL: https://www.frontiersin.org/article/10.3389/fpsyg.2018.01647 (visited on 10/25/2021).

Berridge, Kent C., Terry E. Robinson, and J. Wayne Aldridge (2009). "Dissecting components of reward: 'liking', 'wanting', and learning." In: *Current opinion in pharmacology* 9.1, pp. 65–73. DOI: 10.1016/j.coph.2008.12.014. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2756052/ (visited on 10/25/2021).

Berridge, Kent C. and Morten L. Kringelbach (2015). "Pleasure Systems in the Brain." en. In: *Neuron* 86.3, pp. 646–664. DOI: 10.1016/j.neuron.2015.02.018. URL: https://linkinghub.elsevier.com/retrieve/pii/S0896627315001336 (visited on 01/22/2021).

Berseth, Glen et al. (2021). "SMiRL: Surprise Minimizing Reinforcement Learning in Unstable Environments." In: *arXiv:1912.05510 [cs, stat]*. arXiv: 1912.05510. URL: http://arxiv.org/abs/1912.05510 (visited on 11/24/2021).

Blain, Bastien and Tali Sharot (2021). "Intrinsic reward: potential cognitive and neural mechanisms." en. In: *Current Opinion in Behavioral*

Sciences 39, pp. 113–118. DOI: 10.1016/j.cobeha.2021.03.008. URL: https://linkinghub.elsevier.com/retrieve/pii/S2352154621000565 (visited on 05/13/2021).

Blanchard, Tommy C (2015). "Orbitofrontal Cortex Uses Distinct Codes for Different Choice Attributes in Decisions Motivated by Curiosity." en. In: *Neuron* 85, pp. 602–614.

Blundell, Charles et al. (2015). "Weight Uncertainty in Neural Networks." In: *arXiv:1505.05424 [cs, stat]*. arXiv: 1505.05424. URL: http://arxiv.org/abs/1505.05424 (visited on 12/08/2021).

Boekaerts, Monique (1991). "Subjective competence, appraisals and self-assessment." en. In: *Learning and Instruction* 1.1, pp. 1–17. DOI: 10.1016/0959-4752(91)90016-2. URL: https://linkinghub.elsevier.com/retrieve/pii/0959475291900162 (visited on 06/28/2021).

Bougie, Nicolas and R. Ichise (2019). "Skill-based curiosity for intrinsically motivated reinforcement learning." In: *Machine Learning*. DOI: 10.1007/s10994-019-05845-8.

Bougie, Nicolas and Ryutaro Ichise (2020a). "Fast and slow curiosity for high-level exploration in reinforcement learning." en. In: *Applied Intelligence*. DOI: 10.1007/s10489-020-01849-3. URL: http://link.springer.com/10.1007/s10489-020-01849-3 (visited on 12/02/2020).

– (2020b). "Fast and slow curiosity for high-level exploration in reinforcement learning." In: *Applied Intelligence*, pp. 1–22.

Bourret, Robert B. (2006). "Census of Prokaryotic Senses." en. In: *Journal of Bacteriology* 188.12, pp. 4165–4168. DOI: 10.1128/JB.00311-06. URL: https://journals.asm.org/doi/10.1128/JB.00311-06 (visited on 10/04/2021).

Braunsdorf, Christina, Daniela Mailänder-Sánchez, and Martin Schaller (2016). "Fungal sensing of host environment: Fungal sensing." en. In: *Cellular Microbiology* 18.9, pp. 1188–1200. DOI: 10.1111/cmi.12610. URL: https://onlinelibrary.wiley.com/doi/10.1111/cmi.12610 (visited on 10/04/2021).

Breton-Provencher, Vincent, Gabrielle T. Drummond, and Mriganka Sur (2021). "Locus Coeruleus Norepinephrine in Learned Behavior: Anatomical Modularity and Spatiotemporal Integration in Targets." In: *Frontiers in Neural Circuits* 15, p. 46. DOI: 10.3389/fncir.2021.638007. URL: https://www.frontiersin.org/article/10.3389/fncir.2021.638007 (visited on 10/19/2021).

Brod, Garvin, Markus Werkle-Bergner, and Yee Lee Shing (2013). "The Influence of Prior Knowledge on Memory: A Developmental Cognitive Neuroscience Perspective." en. In: *Frontiers in Behavioral Neuroscience* 7. DOI: 10.3389/fnbeh.2013.00139. URL: http://journal.frontiersin.org/article/10.3389/fnbeh.2013.00139/abstract (visited on 08/10/2021).

Bromberg-Martin, Ethan S. (2011). "Lateral habenula neurons signal errors in the prediction of reward information." en. In: *nature NEUROSCIENCE* 14.9, p. 11.

Bromberg-Martin, Ethan S. and Okihide Hikosaka (2009). "Midbrain Dopamine Neurons Signal Preference for Advance Information about Upcoming Rewards." en. In: 63, pp. 119–126.

Bromberg-Martin, Ethan S. and Tali Sharot (2020). "The Value of Beliefs." en. In: *Neuron* 106.4, pp. 561–565. DOI: 10.1016/j.neuron.2020.05.001. URL: https://linkinghub.elsevier.com/retrieve/pii/S0896627320303470 (visited on 08/10/2021).

Burda, Yuri et al. (2018). "Large-Scale Study of Curiosity-Driven Learning." en. In: *arXiv:1808.04355 [cs, stat]*. arXiv: 1808.04355. URL: http://arxiv.org/abs/1808.04355 (visited on 01/22/2021).

Byrd, Richard H et al. (1995). "A limited memory algorithm for bound constrained optimization." In: *SIAM Journal on scientific computing* 16.5, pp. 1190–1208.

Cacioppo, John T. and Richard E. Petty (1982). "The need for cognition." en. In: *Journal of Personality and Social Psychology* 42.1, pp. 116–131. DOI: 10.1037/0022-3514.42.1.116. URL: http://doi.apa.org/getdoi.cfm?doi=10.1037/0022-3514.42.1.116 (visited on 10/26/2021).

Caligiore, D. et al. (2008). *Using Motor Babbling and Hebb Rules for Modeling the Development of Reaching with Obstacles and Grasping*. en. URL: https://www.semanticscholar.org/paper/Using-Motor-Babbling-and-Hebb-Rules-for-Modeling-of-Caligiore-Ferrauto/f7d6d07627720039c216a21adcc25c502f636dbe (visited on 11/24/2021).

Carleton, R. Nicholas (2016). "Fear of the unknown: One fear to rule them all?" eng. In: *Journal of Anxiety Disorders* 41, pp. 5–21. DOI: 10.1016/j.janxdis.2016.03.011.

Cervera, Roberto Lopez, Maya Zhe Wang, and Benjamin Y Hayden (2020). "Systems neuroscience of curiosity." en. In: *Current Opinion in Behavioral Sciences* 35, pp. 48–55. DOI: 10.1016/j.cobeha.2020.06.011. URL: https://linkinghub.elsevier.com/retrieve/pii/S2352154620300991 (visited on 09/18/2021).

Chu, Junyi and Laura E. Schulz (2020a). "Play, Curiosity, and Cognition." en. In: *Annual Review of Developmental Psychology* 2.1, pp. 317–343. DOI: 10.1146/annurev-devpsych-070120-014806. URL: https://www.annualreviews.org/doi/10.1146/annurev-devpsych-070120-014806 (visited on 03/04/2021).

Chu, Junyi and Laura Schulz (2020b). *Exploratory play, rational action, and efficient search*. en. preprint. PsyArXiv. DOI: 10.31234/osf.io/9yra2. URL: https://osf.io/9yra2 (visited on 03/04/2021).

Clément, Benjamin et al. (2015). "Multi-Armed Bandits for Intelligent Tutoring Systems." In: *Journal of Educational Data Mining* 7.2, pp. 20–48. URL: https://hal.inria.fr/hal-00913669.

Clément, Benjamin, Pierre-Yves Oudeyer, and Manuel Lopes (2016). "A Comparison of Automatic Teaching Strategies for Heterogeneous Student Populations." In: *EDM 16 - 9th International Conference on Educational Data Mining*. Proceedings of the 9th International Conference on Educational Data Mining. Raleigh, United States. URL: https://hal.inria.fr/hal-01360338 (visited on 11/24/2021).

Clément, Benjamin et al. (2015). "Multi-Armed Bandits for Intelligent Tutoring Systems." In: *EDM*. DOI: 10.5281/ZENODO.3554667.

Coenen, Anna, Jonathan D. Nelson, and Todd M. Gureckis (2019). "Asking the right questions about the psychology of human inquiry: Nine open challenges." en. In: *Psychonomic Bulletin & Review* 26.5, pp. 1548–1587. DOI: 10.3758/s13423-018-1470-5. URL: http://link.springer.com/10.3758/s13423-018-1470-5 (visited on 03/04/2021).

Cohen, Jonathan D, Samuel M McClure, and Angela J Yu (2007). "Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration." en. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 362.1481, pp. 933–942. DOI: 10.1098/rstb.2007.2098. URL: https://royalsocietypublishing.org/doi/10.1098/rstb.2007.2098 (visited on 01/22/2021).

Cohn, David, Zoubin Ghahramani, and Michael Jordan (1995). "Active Learning with Statistical Models." In: *Advances in Neural Information Processing Systems*. Ed. by G. Tesauro, D. Touretzky, and T. Leen. Vol. 7. MIT Press. URL: https://proceedings.neurips.cc/paper/1994/file/7f975a56c761db6506eca0b37ce6ec87-Paper.pdf.

Colas, Cédric, Olivier Sigaud, and Pierre-Yves Oudeyer (2018). "GEP-PG: Decoupling Exploration and Exploitation in Deep Reinforcement Learning Algorithms." In: *arXiv:1802.05054 [cs]*. arXiv: 1802.05054. URL: http://arxiv.org/abs/1802.05054 (visited on 11/03/2021).

Colas, Cédric et al. (2019). "CURIOUS: Intrinsically Motivated Modular Multi-Goal Reinforcement Learning." en. In: *arXiv:1810.06284 [cs]*. arXiv: 1810.06284. URL: http://arxiv.org/abs/1810.06284 (visited on 12/02/2020).

Colas, Cédric et al. (2020). "Language as a Cognitive Tool to Imagine Goals in Curiosity-Driven Exploration." en. In: *arXiv:2002.09253 [cs]*. arXiv: 2002.09253. URL: http://arxiv.org/abs/2002.09253 (visited on 12/02/2020).

Colas, Cédric et al. (2021a). "Autotelic Agents with Intrinsically Motivated Goal-Conditioned Reinforcement Learning: a Short Survey." In: *arXiv:2012.09830 [cs]*. arXiv: 2012.09830. URL: http://arxiv.org/abs/2012.09830 (visited on 03/02/2022).

– (2021b). "Intrinsically Motivated Goal-Conditioned Reinforcement Learning: a Short Survey." In: *arXiv:2012.09830 [cs]*. arXiv: 2012.09830. URL: http://arxiv.org/abs/2012.09830 (visited on 11/24/2021).

Collins, Anne GE, James F Cavanagh, and Michael J Frank (2014). "Human EEG uncovers latent generalizable rule structure during learning." In: *Journal of Neuroscience* 34.13, pp. 4677–4685.

Cook, Claire, Noah D. Goodman, and Laura E. Schulz (2011). "Where science starts: Spontaneous experiments in preschoolers' exploratory play." en. In: *Cognition* 120.3, pp. 341–349. DOI: 10.1016/j.cognition.2011.03.003. URL: https://linkinghub.elsevier.com/retrieve/pii/S0010027711000916 (visited on 07/02/2021).

Cunningham, William A and Tobias U Brosch (2012). "Motivational Salience: Amygdala Tuning From Traits, Needs, Values, and Goals." In: *Current Directions in Psychological Science* 21.1, pp. 54–59. URL: https://journals.sagepub.com/doi/abs/10.1177/0963721411430832 (visited on 10/24/2021).

Daddaoua, Nabil, Manuel Lopes, and Jacqueline Gottlieb (2016). "Intrinsically motivated oculomotor exploration guided by uncertainty reduction and conditioned reinforcement in non-human primates." en. In: *Scientific Reports* 6.1, p. 20202. DOI: 10.1038/srep20202. URL: http://www.nature.com/articles/srep20202 (visited on 01/22/2021).

Daffner, K. R. et al. (1992). "Diminished curiosity in patients with probable Alzheimer's disease as measured by exploratory eye movements." eng. In: *Neurology* 42.2, pp. 320–328. DOI: 10.1212/wnl.42.2.320.

Damasio, Antonio and Gil B. Carvalho (2013). "The nature of feelings: evolutionary and neurobiological origins." en. In: *Nature Reviews Neuroscience* 14.2, pp. 143–152. DOI: 10.1038/nrn3403. URL: http://www.nature.com/articles/nrn3403 (visited on 10/25/2021).

Daw, Nathaniel D et al. (2011). "Trial-by-trial data analysis using computational models." In: *Decision making, affect, and learning: Attention and performance XXIII* 23.1.

Day, H. I. (1982). "Curiosity and the interested explorer." In: *Performance & Instruction* 21.4. Place: US Publisher: National Society for Performance & Instruction, pp. 19–22. DOI: 10.1002/pfi.4170210410.

Day, H. I. et al. (1972). "Prior knowledge and the desire for information." en. In: *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement* 4.4, pp. 330–337. DOI: 10.1037/h0082318. URL: http://doi.apa.org/getdoi.cfm?doi=10.1037/h0082318 (visited on 11/03/2021).

Dayan, Peter and Terrence J Sejnowski (1996). "Exploration bonuses and dual control." In: *Machine Learning* 25.1, pp. 5–22.

Deci, Edward L. and Arlen C. Moller (2005). "The Concept of Competence: A Starting Place for Understanding Intrinsic Motivation and Self-Determined Extrinsic Motivation." In: *Handbook of competence and motivation*. New York, NY, US: Guilford Publications, pp. 579–597.

Deci, Edward L. and Richard M. Ryan (1985). "Cognitive Evaluation Theory." en. In: *Intrinsic Motivation and Self-Determination in Human Behavior*. Ed. by Edward L. Deci and Richard M. Ryan. Perspectives in Social Psychology. Boston, MA: Springer US, pp. 43–85. DOI: 10.1007/978-1-4899-2271-7_3. URL: https://doi.org/10.1007/978-1-4899-2271-7_3 (visited on 03/07/2022).

Delmas, Alexandra et al. (2018). "Fostering Health Education With a Serious Game in Children With Asthma: Pilot Studies for Assessing Learning Efficacy and Automatized Learning Personalization." In: *Frontiers in Education* 3, p. 99. DOI: 10.3389/feduc.2018.00099. URL: https://www.frontiersin.org/article/10.3389/feduc.2018.00099 (visited on 11/24/2021).

Dobzhansky, Theodosius (1973). "Nothing in Biology Makes Sense except in the Light of Evolution." In: *The American Biology Teacher* 35.3, pp. 125–129. DOI: 10.2307/4444260. URL: https://doi.org/10.2307/4444260 (visited on 11/24/2021).

Duan, Hongxia et al. (2020). "The effect of intrinsic and extrinsic motivation on memory formation: insight from behavioral and imaging study." In: *Brain Structure and Function* 225, pp. 1561–1574.

Dubey, Rachit and Thomas L Griffiths (2020). "Reconciling Novelty and Complexity Through a Rational Analysis of Curiosity." en. In: *Psychological Review* 127.3, pp. 455–476. DOI: http://dx.doi.org/10.1037/rev0000175.

Dubey, Rachit et al. (2021). *Aha! moments correspond to meta-cognitive prediction errors*. en-us. Tech. rep. type: article. PsyArXiv. DOI: 10.31234/osf.io/c5v42. URL: https://psyarxiv.com/c5v42/ (visited on 01/10/2022).

Duncan, Teresa et al. (2015). "Motivated Strategies for Learning Questionnaire (MSLQ) Manual." en. In: Publisher: Unpublished. DOI: 10.13140/RG.2.1.2547.6968. URL: http://rgdoi.net/10.13140/RG.2.1.2547.6968 (visited on 07/15/2021).

Dunning, David (2011). "Chapter five - The Dunning–Kruger Effect: On Being Ignorant of One's Own Ignorance." en. In: *Advances in Experimental Social Psychology*. Ed. by James M. Olson and Mark P. Zanna. Vol. 44. Academic Press, pp. 247–296. DOI: 10.1016/B978-0-12-385522-0.00005-6. URL: https://www.sciencedirect.com/science/article/pii/B9780123855220000056 (visited on 10/21/2021).

Eliot, T. S. (2014). *The Rock*. en. Google-Books-ID: Qmr8AgAAQBAJ. Houghton Mifflin Harcourt.

Ellsberg, Daniel (1961). "Risk, Ambiguity, and the Savage Axioms." In: *The Quarterly Journal of Economics* 75.4. Publisher: Oxford University Press, pp. 643–669. DOI: 10.2307/1884324. URL: https://www.jstor.org/stable/1884324 (visited on 10/19/2021).

Etcheverry, Mayalen, Clement Moulin-Frier, and Pierre-Yves Oudeyer (2021). "Hierarchically Organized Latent Modules for Exploratory Search in Morphogenetic Systems." In: *arXiv:2007.01195 [nlin, stat]*.

arXiv: 2007.01195. URL: http://arxiv.org/abs/2007.01195 (visited on 10/31/2021).

Ferreira-Vieira, Talita H. et al. (2016). "Alzheimer's Disease: Targeting the Cholinergic System." In: *Current Neuropharmacology* 14.1, pp. 101–115. DOI: 10.2174/1570159X13666150716165726. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4787279/ (visited on 10/19/2021).

Festinger, Leon (1962). *A Theory of Cognitive Dissonance*. en. Stanford University Press.

Fiedler, Glen (2004). *Fix Your Timestep!* en-us. URL: https://gafferongames.com/post/fix_your_timestep/ (visited on 12/26/2021).

Fine, Justin M. and Benjamin Y. Hayden (2022). "The whole prefrontal cortex is premotor cortex." In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 377.1844. Publisher: Royal Society, p. 20200524. DOI: 10.1098/rstb.2020.0524. URL: https://royalsocietypublishing.org/doi/10.1098/rstb.2020.0524 (visited on 02/12/2022).

Fischhoff, Baruch, Paul Slovic, and Sarah Lichtenstein (1977). "Knowing with certainty: The appropriateness of extreme confidence." In: *Journal of Experimental Psychology: Human Perception and Performance* 3.4. Place: US Publisher: American Psychological Association, pp. 552–564. DOI: 10.1037/0096-1523.3.4.552.

FitzGibbon, Lily, Asuka Komiya, and Kou Murayama (2021). "The Lure of Counterfactual Curiosity: People Incur a Cost to Experience Regret." eng. In: *Psychological Science* 32.2, pp. 241–255. DOI: 10.1177/0956797620963615.

FitzGibbon, Lily, Johnny King L. Lau, and Kou Murayama (2020). "The seductive lure of curiosity: information as a motivationally salient reward." en. In: *Current Opinion in Behavioral Sciences* 35, pp. 21–27. DOI: 10.1016/j.cobeha.2020.05.014. URL: https://linkinghub.elsevier.com/retrieve/pii/S2352154620300875 (visited on 03/04/2021).

Florensa, Carlos et al. (2018). "Automatic Goal Generation for Reinforcement Learning Agents." en. In: *arXiv:1705.06366 [cs]*. arXiv: 1705.06366. URL: http://arxiv.org/abs/1705.06366 (visited on 03/11/2021).

Fogarty, Laurel and Nicole Creanza (2017). "The niche construction of cultural complexity: interactions between innovations, population size and the environment." eng. In: *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 372.1735, p. 20160428. DOI: 10.1098/rstb.2016.0428.

Forestier, Sébastien et al. (2020). "Intrinsically Motivated Goal Exploration Processes with Automatic Curriculum Learning." en. In: *arXiv:1708.02190 [cs]*. arXiv: 1708.02190. URL: http://arxiv.org/abs/1708.02190 (visited on 12/02/2020).

Friston, Karl (2009). "The free-energy principle: a rough guide to the brain?" en. In: *Trends in Cognitive Sciences* 13.7, pp. 293–301. DOI: 10.1016/j.tics.2009.04.005. URL: https://linkinghub.elsevier.com/retrieve/pii/S136466130900117X (visited on 05/23/2021).

Geana, Andra et al. (2016). "Boredom, information-seeking, and exploration." In: *Proceedings of the 38th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society. URL: https://cogsci.mindmodeling.org/2016/papers/0307/index.html.

Gerken, LouAnn, Frances K. Balcomb, and Juliet L. Minton (2011). "Infants avoid 'labouring in vain' by attending more to learnable than unlearnable linguistic patterns: Infants attend more to learnable patterns." en. In: *Developmental Science* 14.5, pp. 972–979. DOI: 10.1111/j.1467-7687.2011.01046.x. URL: http://doi.wiley.com/10.1111/j.1467-7687.2011.01046.x (visited on 07/02/2021).

Gershman, Samuel J (2018a). "Deconstructing the human algorithms for exploration." en. In: p. 9.

Gershman, Samuel J. (2018b). "Uncertainty and Exploration." en. In: DOI: 10.1101/265504. URL: http://biorxiv.org/lookup/doi/10.1101/265504 (visited on 01/22/2021).

Gipson, Cassandra D. et al. (2009). "Preference for 50% reinforcement over 75% reinforcement by pigeons." en. In: *Learning & Behavior* 37.4, pp. 289–298. DOI: 10.3758/LB.37.4.289. URL: https://doi.org/10.3758/LB.37.4.289 (visited on 10/27/2021).

Gold, Benjamin P. et al. (2019). "Predictability and Uncertainty in the Pleasure of Music: A Reward for Learning?" en. In: *The Journal of Neuroscience* 39.47, pp. 9397–9409. DOI: 10.1523/JNEUROSCI.0428-19.2019. URL: https://www.jneurosci.org/lookup/doi/10.1523/JNEUROSCI.0428-19.2019 (visited on 07/16/2021).

Gordon, G., E. Fonio, and E. Ahissar (2014). "Emergent Exploration via Novelty Management." en. In: *Journal of Neuroscience* 34.38, pp. 12646–12661. DOI: 10.1523/JNEUROSCI.1872-14.2014. URL: https://www.jneurosci.org/lookup/doi/10.1523/JNEUROSCI.1872-14.2014 (visited on 01/31/2022).

Gordon, Goren (2020). "Infant-inspired intrinsically motivated curious robots." In: *Current Opinion in Behavioral Sciences* 35, pp. 28–34. DOI: 10.1016/j.cobeha.2020.05.010.

Gordon, Goren and Ehud Ahissar (2012). "Hierarchical curiosity loops and active sensing." en. In: *Neural Networks* 32, pp. 119–129. DOI: 10.1016/j.neunet.2012.02.024. URL: https://linkinghub.elsevier.com/retrieve/pii/S0893608012000536 (visited on 01/31/2022).

Gordon, Goren, Cynthia Breazeal, and Susan Engel (2015). "Can Children Catch Curiosity from a Social Robot?" en. In: *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*. Portland Oregon USA: ACM, pp. 91–98. DOI: 10.1145/2696454.2696469. URL: https://dl.acm.org/doi/10.1145/2696454.2696469 (visited on 01/31/2022).

Gottlieb, Jacqueline and Pierre-Yves Oudeyer (2018). "Towards a neuroscience of active sampling and curiosity." en. In: *Nature Reviews Neuroscience* 19.12, pp. 758–770. DOI: 10.1038/s41583-018-0078-0. URL: http://www.nature.com/articles/s41583-018-0078-0 (visited on 01/22/2021).

Gottlieb, Jacqueline et al. (2013). "Information-seeking, curiosity, and attention: computational and neural mechanisms." en. In: *Trends in Cognitive Sciences* 17.11, pp. 585–593. DOI: 10.1016/j.tics.2013.09.001. URL: https://linkinghub.elsevier.com/retrieve/pii/S1364661313002052 (visited on 01/22/2021).

Gottlieb, Jacqueline et al. (2020). "Curiosity, information demand and attentional priority." en. In: *Current Opinion in Behavioral Sciences* 35, pp. 83–91. DOI: 10.1016/j.cobeha.2020.07.016. URL: https://linkinghub.elsevier.com/retrieve/pii/S2352154620301182 (visited on 06/23/2021).

Graves, Alex (2011). "Practical Variational Inference for Neural Networks." In: *Advances in Neural Information Processing Systems*. Vol. 24. Curran Associates, Inc. URL: https://papers.nips.cc/paper/2011/hash/7eb3c8be3d411e8ebfab08eba5f49632-Abstract.html (visited on 12/08/2021).

Graves, Alex et al. (2017). "Automated curriculum learning for neural networks." In: *arXiv preprint arXiv:1704.03003*.

Griffiths, Thomas L. et al. (2010). "Probabilistic models of cognition: exploring representations and inductive biases." English. In: *Trends in Cognitive Sciences* 14.8. Publisher: Elsevier, pp. 357–364. DOI: 10.1016/j.tics.2010.05.004. URL: https://www.cell.com/trends/cognitive-sciences/abstract/S1364-6613(10)00112-9 (visited on 10/30/2021).

Grizou, Jonathan et al. (2020). "A curious formulation robot enables the discovery of a novel protocell behavior." In: *Science Advances* 6.5, eaay4237. DOI: 10.1126/sciadv.aay4237. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6994213/ (visited on 11/24/2021).

Gross, Madeleine E., Claire M. Zedelius, and Jonathan W. Schooler (2020). "Cultivating an understanding of curiosity as a seed for creativity." In: *Current Opinion in Behavioral Sciences* 35. Place: Netherlands Publisher: Elsevier Science, pp. 77–82. DOI: 10.1016/j.cobeha.2020.07.015.

Grossnickle, Emily M. (2016). "Disentangling Curiosity: Dimensionality, Definitions, and Distinctions from Interest in Educational Contexts." en. In: *Educational Psychology Review* 28.1, pp. 23–60. DOI: 10.1007/s10648-014-9294-y. URL: https://doi.org/10.1007/s10648-014-9294-y (visited on 10/18/2021).

Gruber, Matthias J., Bernard D. Gelman, and Charan Ranganath (2014). "States of Curiosity Modulate Hippocampus-Dependent Learning via the Dopaminergic Circuit." en. In: *Neuron* 84.2, pp. 486–496. DOI: 10.1016/j.neuron.2014.08.060. URL: https://linkingh

ub.elsevier.com/retrieve/pii/S0896627314008046 (visited on 01/22/2021).

Grupe, Dan W. and Jack B. Nitschke (2013). "Uncertainty and anticipation in anxiety: an integrated neurobiological and psychological perspective." en. In: *Nature Reviews Neuroscience* 14.7. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 7 Primary_atype: Reviews Publisher: Nature Publishing Group Subject_term: Psychiatric disorders Subject_term_id: psychiatric-disorders, pp. 488–501. DOI: 10.1038/nrn3524. URL: https://www.nature.com/articles/nrn3524 (visited on 10/24/2021).

Guay, Frédéric, Robert J. Vallerand, and Céline Blanchard (2000). "On the Assessment of Situational Intrinsic and Extrinsic Motivation: The Situational Motivation Scale (SIMS)." en. In: *Motivation and Emotion* 24.3, pp. 175–213. DOI: 10.1023/A:1005614228250. URL: http://link.springer.com/10.1023/A:1005614228250 (visited on 07/15/2021).

Haber, Nick et al. (2018). "Learning to Play with Intrinsically-Motivated Self-Aware Agents." In: *arXiv:1802.07442 [cs, stat]*. arXiv: 1802.07442. URL: http://arxiv.org/abs/1802.07442 (visited on 11/24/2021).

Harlow, Harry F., Margaret Kuenne Harlow, and Donald R. Meyer (1950). "Learning motivated by a manipulation drive." In: *Journal of Experimental Psychology* 40.2. Place: US Publisher: American Psychological Association, pp. 228–234. DOI: 10.1037/h0056906.

Hart, Sandra G. and Lowell E. Staveland (1988). "Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research." en. In: *Advances in Psychology*. Ed. by Peter A. Hancock and Najmedin Meshkati. Vol. 52. Human Mental Workload. North-Holland, pp. 139–183. DOI: 10.1016/S0166-4115(08)62386-9. URL: https://www.sciencedirect.com/science/article/pii/S0166411508623869 (visited on 12/15/2021).

Hebb, Donald O. (1955). "Drives and the C. N. S. (conceptual nervous system)." In: *Psychological Review* 62.4. Place: US Publisher: American Psychological Association, pp. 243–254. DOI: 10.1037/h0041823.

– (2002). *The organization of behavior: a neuropsychological theory*. en. Mahwah, N.J: L. Erlbaum Associates.

Hidi, Suzanne E. and K. Ann Renninger (2019). "Interest Development and Its Relation to Curiosity: Needed Neuroscientific Research." en. In: *Educational Psychology Review* 31.4, pp. 833–852. DOI: 10.1007/s10648-019-09491-3. URL: http://link.springer.com/10.1007/s10648-019-09491-3 (visited on 08/10/2021).

Hidi, Suzanne and K Ann Renninger (2006). "The Four-Phase Model of Interest Development." en. In: *Educational Psychologist* 41.1, p. 18. DOI: 10.1207/s15326985ep4102_4. URL: https://doi.org/10.1207/s15326985ep4102_4.

Hillen, Marij A. et al. (2017). "Tolerance of uncertainty: Conceptual analysis, integrative model, and implications for healthcare." en. In:

Social Science & Medicine 180, pp. 62–75. DOI: 10.1016/j.socscimed.2017.03.024. URL: https://www.sciencedirect.com/science/article/pii/S0277953617301703 (visited on 10/26/2021).

Holm, Linus, Gustaf Wadenholt, and Paul Schrater (2019). "Episodic curiosity for avoiding asteroids: Per-trial information gain for choice outcomes drive information seeking." en. In: *Scientific Reports* 9.1, p. 11265. DOI: 10.1038/s41598-019-47671-x. URL: http://www.nature.com/articles/s41598-019-47671-x (visited on 01/22/2021).

Horstmann, Gernot (2015). "The surprise-attention link: a review." eng. In: *Annals of the New York Academy of Sciences* 1339, pp. 106–115. DOI: 10.1111/nyas.12679.

Hull, C. L. (1943). *Principles of behavior: an introduction to behavior theory*. Principles of behavior: an introduction to behavior theory. Pages: x, 422. Oxford, England: Appleton-Century.

Hunt, J. McV. (1960). "Experience and the Development of Motivation: Some Reinterpretations." In: *Child Development* 31.3. Publisher: [Wiley, Society for Research in Child Development], pp. 489–504. DOI: 10.2307/1126044. URL: https://www.jstor.org/stable/1126044 (visited on 10/19/2021).

Hüllermeier, Eyke and Willem Waegeman (2021). "Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods." en. In: *Machine Learning* 110.3, pp. 457–506. DOI: 10.1007/s10994-021-05946-3. URL: http://link.springer.com/10.1007/s10994-021-05946-3 (visited on 12/02/2021).

Iigaya, Kiyohito et al. (2016). "The modulation of savouring by prediction error and its effects on choice." en. In: *eLife* 5, e13747. DOI: 10.7554/eLife.13747. URL: https://elifesciences.org/articles/13747 (visited on 01/22/2021).

Irwing, Paul, Tom Booth, and David J. Hughes (2018). *The Wiley Handbook of Psychometric Testing: A Multidisciplinary Reference on Survey, Scale and Test Development*. en. Google-Books-ID: lTVMDwAAQBAJ. John Wiley & Sons.

Itti, Laurent and Pierre Baldi (2009). "Bayesian surprise attracts human attention." en. In: *Vision Research*, p. 12.

Jaderberg, Max et al. (2016). "Reinforcement Learning with Unsupervised Auxiliary Tasks." In: *arXiv:1611.05397 [cs]*. arXiv: 1611.05397. URL: http://arxiv.org/abs/1611.05397 (visited on 11/24/2021).

Jaques, Natasha et al. (2019). "Social Influence as Intrinsic Motivation for Multi-Agent Deep Reinforcement Learning." In: *Proceedings of the 35 th International Conference on Machine Learning, Stockholm, Sweden*. arXiv: 1810.08647. URL: https://openreview.net/forum?id=B1lG42C9Km.

Jepma, Marieke (2012). "Neural mechanisms underlying the induction and relief of perceptual curiosity." en. In: *Frontiers in Behavioral Neuroscience* 6.

Jordan, M. I. and T. M. Mitchell (2015). "Machine learning: Trends, perspectives, and prospects." In: *Science* 349.6245. Publisher: American Association for the Advancement of Science, pp. 255–260. DOI: 10.1126/science.aaa8415. URL: https://www.science.org/doi/10.1126/science.aaa8415 (visited on 11/03/2021).

Jordan, Michael and David Rumelhart (1992). "Forward models: Supervised learning with a distal teacher." en. In: *Cognitive Science* 16.3, pp. 307–354. DOI: 10.1016/0364-0213(92)90036-T. URL: https://www.sciencedirect.com/science/article/pii/036402139290036T (visited on 11/24/2021).

Juechems, Keno and Christopher Summerfield (2019). "Where Does Value Come From?" en. In: *Trends in Cognitive Sciences* 23.10, pp. 836–850. DOI: 10.1016/j.tics.2019.07.012. URL: https://linkinghub.elsevier.com/retrieve/pii/S1364661319302001 (visited on 03/04/2021).

Kang, Min Jeong et al. (2009). "The wick in the candle of learning: Epistemic curiosity activates reward circuitry and enhances memory." eng. In: *Psychological Science* 20.8, pp. 963–973. DOI: 10.1111/j.1467-9280.2009.02402.x.

Kaplan, Frederic and Pierre-Yves Oudeyer (2003). "Motivational principles for visual know-how development." In.

– (2007). "In search of the neural circuits of intrinsic motivation." en. In: *Frontiers in Neuroscience* 1.1, pp. 225–236. DOI: 10.3389/neuro.01.1.1.017.2007. URL: http://journal.frontiersin.org/article/10.3389/neuro.01.1.1.017.2007/abstract (visited on 04/09/2021).

Kendall, Alex and Yarin Gal (2017). "What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?" In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2017/hash/2650d6089a6d640c5e85b2b88265dc2b-Abstract.html (visited on 12/02/2021).

Kepecs, Adam and Zachary F. Mainen (2012). "A computational framework for the study of confidence in humans and animals." en. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 367.1594, pp. 1322–1337. DOI: 10.1098/rstb.2012.0037. URL: https://royalsocietypublishing.org/doi/10.1098/rstb.2012.0037 (visited on 07/02/2021).

Kidd, Celeste and Benjamin Y. Hayden (2015). "The Psychology and Neuroscience of Curiosity." en. In: *Neuron* 88.3, pp. 449–460. DOI: 10.1016/j.neuron.2015.09.010. URL: https://linkinghub.elsevier.com/retrieve/pii/S0896627315007679 (visited on 01/22/2021).

Kidd, Celeste, Steven T. Piantadosi, and Richard N. Aslin (2012). "The Goldilocks Effect: Human Infants Allocate Attention to Visual Sequences That Are Neither Too Simple Nor Too Complex." en. In: *PLoS ONE* 7.5. Ed. by Antoni Rodriguez-Fornells, e36399. DOI: 10

.1371/journal.pone.0036399. URL: https://dx.plos.org/10.1371/journal.pone.0036399 (visited on 01/22/2021).

Kim, Kuno et al. (2020). "Active World Model Learning with Progress Curiosity." In: *arXiv:2007.07853 [cs, stat]*. arXiv: 2007.07853. URL: http://arxiv.org/abs/2007.07853 (visited on 11/24/2021).

Kobayashi, Kenji et al. (2019). "Diverse motives for human curiosity." en. In: *Nature Human Behaviour* 3.6, pp. 587–595. DOI: 10.1038/s41562-019-0589-3. URL: http://www.nature.com/articles/s41562-019-0589-3 (visited on 04/20/2021).

Kolmogorov, Andrei N (1965). "Three approaches to the quantitative definition ofinformation'." In: *Problems of information transmission* 1.1, pp. 1–7.

Koriat, A. (1993). "How do we know that we know? The accessibility model of the feeling of knowing." eng. In: *Psychological Review* 100.4, pp. 609–639. DOI: 10.1037/0033-295x.100.4.609.

Kornell, Nate and Hannah Hausman (2017). "Performance bias: Why judgments of learning are not affected by learning." en. In: *Memory & Cognition* 45.8, pp. 1270–1280. DOI: 10.3758/s13421-017-0740-1. URL: https://doi.org/10.3758/s13421-017-0740-1 (visited on 01/19/2022).

Kovač, Grgur, Adrien Laversanne-Finot, and Pierre-Yves Oudeyer (2020). "GRIMGEP: Learning Progress for Robust Goal Sampling in Visual Deep Reinforcement Learning." en. In: *arXiv:2008.04388 [cs, stat]*. arXiv: 2008.04388. URL: http://arxiv.org/abs/2008.04388 (visited on 12/02/2020).

Kruger, J. and D. Dunning (1999). "Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments." eng. In: *Journal of Personality and Social Psychology* 77.6, pp. 1121–1134. DOI: 10.1037//0022-3514.77.6.1121.

Lau, Johnny King L et al. (2020). "Shared striatal activity in decisions to satisfy curiosity and hunger at the risk of electric shocks." In: *Nature Human Behaviour* 4.5, pp. 531–543.

Laversanne-Finot, Adrien, Alexandre Péré, and Pierre-Yves Oudeyer (2018). "Curiosity Driven Exploration of Learned Disentangled Goal Spaces." en. In: *arXiv:1807.01521 [cs, stat]*. arXiv: 1807.01521. URL: http://arxiv.org/abs/1807.01521 (visited on 01/22/2021).

– (2021). "Intrinsically Motivated Exploration of Learned Goal Spaces." In: *Frontiers in Neurorobotics* 14, p. 109. DOI: 10.3389/fnbot.2020.555271. URL: https://www.frontiersin.org/article/10.3389/fnbot.2020.555271 (visited on 11/24/2021).

Lazarus, Richard S. (1991). "Progress on a cognitive-motivational-relational theory of emotion." In: *American Psychologist* 46.8. Place: US Publisher: American Psychological Association, pp. 819–834. DOI: 10.1037/0003-066X.46.8.819.

Lefort, Mathieu and Alexander Gepperth (2015). "Active learning of local predictable representations with artificial curiosity." en. In:

2015 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob). Providence, RI, USA: IEEE, pp. 228–233. DOI: 10.1109/DEVLRN.2015.7346145. URL: http://ieeexplore.ieee.org/document/7346145/ (visited on 12/02/2020).

Leibo, Joel Z. et al. (2019). "Autocurricula and the Emergence of Innovation from Social Interaction: A Manifesto for Multi-Agent Intelligence Research." In: *arXiv:1903.00742 [cs, q-bio]*. arXiv: 1903.00742. URL: http://arxiv.org/abs/1903.00742 (visited on 11/24/2021).

Leonard, Julia et al. (2021). *Young children calibrate effort based on the trajectory of their performance*. en. preprint. PsyArXiv. DOI: 10.31234/osf.io/hc62q. URL: https://osf.io/hc62q (visited on 08/10/2021).

Lewthwaite, Rebecca and Gabriele Wulf (2017). "Optimizing motivation and attention for motor performance and learning." en. In: *Current Opinion in Psychology* 16, pp. 38–42. DOI: 10.1016/j.copsyc.2017.04.005. URL: https://linkinghub.elsevier.com/retrieve/pii/S2352250X1630152X (visited on 01/15/2022).

Limeri, Lisa B. et al. (2020). "Growing a growth mindset: characterizing how and why undergraduate students' mindsets change." In: *International Journal of STEM Education* 7.1, p. 35. DOI: 10.1186/s40594-020-00227-2. URL: https://doi.org/10.1186/s40594-020-00227-2 (visited on 02/18/2022).

Linke, Cam et al. (2020). "Adapting Behaviour via Intrinsic Reward: A Survey and Empirical Study." en. In: *arXiv:1906.07865 [cs, stat]*. arXiv: 1906.07865. URL: http://arxiv.org/abs/1906.07865 (visited on 12/02/2020).

Liquin, Emily G and Tania Lombrozo (2020). "Explanation-seeking curiosity in childhood." en. In: *Current Opinion in Behavioral Sciences* 35, pp. 14–20. DOI: 10.1016/j.cobeha.2020.05.012. URL: https://linkinghub.elsevier.com/retrieve/pii/S2352154620300851 (visited on 03/04/2021).

Litman, Jordan (2019). "Curiosity." en. In: *The Cambridge Handbook of Motivation and Learning*. 1st ed. Cambridge University Press. DOI: 10.1017/9781316823279. URL: https://www.cambridge.org/core/product/identifier/9781316823279/type/book (visited on 09/18/2021).

Locke, Shannon M., Pascal Mamassian, and Michael S. Landy (2020). "Performance monitoring for sensorimotor confidence: A visuomotor tracking study." en. In: *Cognition* 205, p. 104396. DOI: 10.1016/j.cognition.2020.104396. URL: https://linkinghub.elsevier.com/retrieve/pii/S0010027720302158 (visited on 05/26/2021).

Loewenstein, G. (1987). "Anticipation and the Valuation of Delayed Consumption." In: DOI: 10.2307/2232929.

Loewenstein, George (1994). "The psychology of curiosity: A review and reinterpretation." en. In: *Psychological Bulletin* 116.1, pp. 75–98.

Lopes, Manuel and Pierre-Yves Oudeyer (2012). "The Strategic Student Approach for Life-Long Exploration and Learning." en. In: p. 8.

Lucas, Robert E (2004). "The Industrial Revolution: Past and Future." en. In: *Economic Education Bulletin* XLIV.8, p. 8.

Ma, Wei Ji and Mehrdad Jazayeri (2014). "Neural coding of uncertainty and probability." eng. In: *Annual Review of Neuroscience* 37, pp. 205–220. DOI: 10.1146/annurev-neuro-071013-014017.

Maia, Tiago V. (2009). "Reinforcement learning, conditioning, and the brain: Successes and challenges." en. In: *Cognitive, Affective, & Behavioral Neuroscience* 9.4, pp. 343–364. DOI: 10.3758/CABN.9.4.343. URL: http://link.springer.com/10.3758/CABN.9.4.343 (visited on 10/26/2021).

Markant, Douglas B. and Todd M. Gureckis (2013). "Is it better to select or to receive? Learning via active and passive hypothesis testing." en. In: *Journal of Experimental Psychology: General* 143.1, pp. 94–122. DOI: 10.1037/a0032108. URL: http://doi.apa.org/getdoi.cfm?doi=10.1037/a0032108 (visited on 03/04/2021).

Markus, Hazel and Elissa Wurf (1987). "The dynamic self-concept: A social psychological perspective." In: *Annual Review of Psychology* 38. Place: US Publisher: Annual Reviews, pp. 299–337. DOI: 10.1146/annurev.ps.38.020187.001503.

Martí, Louis et al. (2018). "Certainty Is Primarily Determined by Past Performance During Concept Learning." en. In: *Open Mind* 2.2, pp. 47–60. DOI: 10.1162/opmi_a_00017. URL: https://www.mitpressjournals.org/doi/abs/10.1162/opmi_a_00017 (visited on 03/12/2021).

Matiisen, Tambet et al. (2019). "Teacher-student curriculum learning." In: *IEEE transactions on neural networks and learning systems*.

Maturana, H. R. and F. J. Varela (1980). *Autopoiesis and Cognition: The Realization of the Living*. English. 1st edition. Dordrecht, Holland ; Boston: D. Reidel Publishing Company.

McAuley, E., T. Duncan, and V. V. Tammen (1989). "Psychometric properties of the Intrinsic Motivation Inventory in a competitive sport setting: a confirmatory factor analysis." eng. In: *Research Quarterly for Exercise and Sport* 60.1, pp. 48–58. DOI: 10.1080/02701367.1989.10607413.

McClelland, James L. (2009). "The Place of Modeling in Cognitive Science." en. In: *Topics in Cognitive Science* 1.1, pp. 11–38. DOI: 10.1111/j.1756-8765.2008.01003.x. URL: http://doi.wiley.com/10.1111/j.1756-8765.2008.01003.x (visited on 01/22/2021).

McClelland, James L. et al. (2010). "Letting structure emerge: connectionist and dynamical systems approaches to cognition." en. In: *Trends in Cognitive Sciences* 14.8, pp. 348–356. DOI: 10.1016/j.tics.2010.06.002. URL: https://linkinghub.elsevier.com/retrieve/pii/S1364661310001245 (visited on 01/22/2021).

McClelland, James L et al. (2014). "Interactive Activation and Mutual Constraint Satisfaction in Perception and Cognition." en. In: *Cognitive Science*, p. 51.

Metcalfe, Janet and Nate Kornell (2005). "A Region of Proximal Learning model of study time allocationq." en. In: *Journal of Memory and Language*, p. 16.

Metcalfe, Janet, Bennett L Schwartz, and Teal S Eich (2020). "Epistemic curiosity and the region of proximal learning." en. In: *Current Opinion in Behavioral Sciences* 35, pp. 40–47. DOI: 10.1016/j.cobeha.2020.06.007. URL: https://linkinghub.elsevier.com/retrieve/pii/S2352154620300954 (visited on 03/04/2021).

Mirolli, Marco and Gianluca Baldassarre (2013). "Functions and Mechanisms of Intrinsic Motivations." en. In: *Intrinsically Motivated Learning in Natural and Artificial Systems*. Ed. by Gianluca Baldassarre and Marco Mirolli. Berlin, Heidelberg: Springer, pp. 49–72. DOI: 10.1007/978-3-642-32375-1_3. URL: https://doi.org/10.1007/978-3-642-32375-1_3 (visited on 11/24/2021).

Mnih, Volodymyr et al. (2015). "Human-level control through deep reinforcement learning." en. In: *Nature* 518.7540. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 7540 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Computer science Subject_term_id: computer-science, pp. 529–533. DOI: 10.1038/nature14236. URL: https://www.nature.com/articles/nature14236 (visited on 11/24/2021).

Moulin-Frier, Clement and Pierre-Yves Oudeyer (2013). "Exploration strategies in developmental robotics: A unified probabilistic framework." en. In: *2013 IEEE Third Joint International Conference on Development and Learning and Epigenetic Robotics (ICDL)*. Osaka, Japan: IEEE, pp. 1–6. DOI: 10.1109/DevLrn.2013.6652535. URL: http://ieeexplore.ieee.org/document/6652535/ (visited on 12/02/2020).

Moulin-Frier, Clément, Sao Mai Nguyen, and Pierre-Yves Oudeyer (2014). "Self-organization of early vocal development in infants and machines: the role of intrinsic motivation." en. In: *Frontiers in Psychology* 4. DOI: 10.3389/fpsyg.2013.01006. URL: http://journal.frontiersin.org/article/10.3389/fpsyg.2013.01006/abstract (visited on 12/02/2020).

Mu, Tong et al. (2018). "Combining adaptivity with progression ordering for intelligent tutoring systems." In: *Proceedings of the Fifth Annual ACM Conference on Learning at Scale*, pp. 1–4.

Murayama, Kou, Lily FitzGibbon, and Michiko Sakaki (2019). "Process Account of Curiosity and Interest: A Reward-Learning Perspective." en. In: *Educational Psychology Review*, p. 21.

Murayama, Kou et al. (2019). "Motivated for near impossibility: How task type and reward modulates intrinsic motivation and the striatal activation for an extremely difficult task." In: *bioRxiv*, p. 828756.

Nair, Ashvin et al. (2018). "Visual Reinforcement Learning with Imagined Goals." In: *arXiv:1807.04742 [cs, stat]*. arXiv: 1807.04742. URL: http://arxiv.org/abs/1807.04742 (visited on 11/18/2020).

Nake, Frieder (1974). *Ästhetik als Informationsverarbeitung: Grundlagen und Anwendungen der Informatik im Bereich ästhetischer Produktion und Kritik*. Deutsch. Springer-Verlag/Wien.

Nguyen, Sao Mai and Pierre-Yves Oudeyer (2012). "Active choice of teachers, learning strategies and goals for a socially guided intrinsic motivation learner." In: *Paladyn, Journal of Behavioral Robotics* 3.3. arXiv: 1804.06819. DOI: 10.2478/s13230-013-0110-z. URL: http://arxiv.org/abs/1804.06819 (visited on 11/24/2021).

Nickerson, Raymond S. (1998). "Confirmation Bias: A Ubiquitous Phenomenon in Many Guises." en. In: *Review of General Psychology* 2.2. Publisher: SAGE Publications Inc, pp. 175–220. DOI: 10.1037/1089-2680.2.2.175. URL: https://doi.org/10.1037/1089-2680.2.2.175 (visited on 10/27/2021).

Nisioti, Eleni and Clément Moulin-Frier (2020). "Grounding Artificial Intelligence in the Origins of Human Behavior." In: *arXiv:2012.08564 [cs]*. arXiv: 2012.08564. URL: http://arxiv.org/abs/2012.08564 (visited on 11/24/2021).

Noordewier, Marret K. and Eric van Dijk (2017). "Curiosity and time: from not knowing to almost knowing." en. In: *Cognition and Emotion* 31.3, pp. 411–421. DOI: 10.1080/02699931.2015.1122577. URL: https://www.tandfonline.com/doi/full/10.1080/02699931.2015.1122577 (visited on 01/22/2021).

Nussenbaum, Kate and Catherine A. Hartley (2019). "Reinforcement learning across development: What insights can we draw from a decade of research?" en. In: *Developmental Cognitive Neuroscience* 40, p. 100733. DOI: 10.1016/j.dcn.2019.100733. URL: https://linkinghub.elsevier.com/retrieve/pii/S1878929319303202 (visited on 03/04/2021).

Oller, D. Kimbrough (2000). *The emergence of the speech capacity*. The emergence of the speech capacity. Pages: xvii, 428. Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Olshausen, Bruno A. (2013). "Perception as an Inference Problem." en. In: *The Cognitive Neurosciences*. V. MIT Press, p. 18.

Oudeyer, P.-Y., J. Gottlieb, and M. Lopes (2016a). "Intrinsic motivation, curiosity, and learning." en. In: *Progress in Brain Research*. Vol. 229. Elsevier, pp. 257–284. DOI: 10.1016/bs.pbr.2016.05.005. URL: https://linkinghub.elsevier.com/retrieve/pii/S0079612316300589 (visited on 03/07/2022).

Oudeyer, P-Y, Jacqueline Gottlieb, and Manuel Lopes (2016b). "Intrinsic motivation, curiosity, and learning: Theory and applications in educational technologies." In: *Progress in brain research*. Vol. 229. Elsevier, pp. 257–284.

Oudeyer, Pierre-Yves (2018). "Computational Theories of Curiosity-Driven Learning." en. In: *arXiv:1802.10546 [cs]*. arXiv: 1802.10546. URL: http://arxiv.org/abs/1802.10546 (visited on 01/22/2021).

Oudeyer, Pierre-Yves, Frdric Kaplan, and Verena V. Hafner (2007). "Intrinsic Motivation Systems for Autonomous Mental Development." en. In: *IEEE Transactions on Evolutionary Computation* 11.2, pp. 265–286. DOI: 10.1109/TEVC.2006.890271. URL: http://ieeexplore.ieee.org/document/4141061/ (visited on 01/22/2021).

Oudeyer, Pierre-Yves and Frederic Kaplan (2007). "What is intrinsic motivation? A typology of computational approaches." en. In: *Frontiers in Neurorobotics* 1, p. 14.

Oudeyer, Pierre-Yves and Frédéric Kaplan (2006). "Discovering communication." In: *Connection Science* 18.2. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/09540090600768567, pp. 189–206. DOI: 10.1080/09540090600768567. URL: https://doi.org/10.1080/09540090600768567 (visited on 11/24/2021).

Oudeyer, Pierre-Yves and Linda B Smith (2016). "How Evolution May Work Through Curiosity-Driven Developmental Process." en. In: *Topics in Cognitive Science*, p. 11.

Pan, Minxue et al. (2020). "Reinforcement learning based curiosity-driven testing of Android applications." In: *Proceedings of the 29th ACM SIGSOFT International Symposium on Software Testing and Analysis*. ISSTA 2020. New York, NY, USA: Association for Computing Machinery, pp. 153–164. DOI: 10.1145/3395363.3397354. URL: https://doi.org/10.1145/3395363.3397354 (visited on 11/24/2021).

Pateria, Shubham et al. (2021). "Hierarchical Reinforcement Learning: A Comprehensive Survey." en. In: *ACM Computing Surveys* 54.5, pp. 1–35. DOI: 10.1145/3453160. URL: https://dl.acm.org/doi/10.1145/3453160 (visited on 10/17/2021).

Pathak, Deepak et al. (2017). "Curiosity-driven Exploration by Self-supervised Prediction." en. In: *arXiv:1705.05363 [cs, stat]*. arXiv: 1705.05363. URL: http://arxiv.org/abs/1705.05363 (visited on 01/22/2021).

Payzan-LeNestour, Elise and Peter Bossaerts (2011). "Risk, Unexpected Uncertainty, and Estimation Uncertainty: Bayesian Learning in Unstable Settings." en. In: *PLOS Computational Biology* 7.1. Publisher: Public Library of Science, e1001048. DOI: 10.1371/journal.pcbi.1001048. URL: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1001048 (visited on 03/02/2022).

Poli, F. et al. (2020). "Infants tailor their attention to maximize learning." en. In: *Science Advances* 6.39, eabb5053. DOI: 10.1126/sciadv.abb5053. URL: https://advances.sciencemag.org/lookup/doi/10.1126/sciadv.abb5053 (visited on 03/04/2021).

Pong, Vitchyr H. et al. (2020). "Skew-Fit: State-Covering Self-Supervised Reinforcement Learning." In: *arXiv:1903.03698 [cs, stat]*. arXiv: 1903.03698. URL: http://arxiv.org/abs/1903.03698 (visited on 05/15/2020).

Portelas, Rémy et al. (2020a). "Automatic Curriculum Learning For Deep RL: A Short Survey." In: *Proceedings of IJCAI 2020*.

Portelas, Rémy et al. (2020b). "Automatic Curriculum Learning For Deep RL: A Short Survey." en. In: *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*. Yokohama, Japan: International Joint Conferences on Artificial Intelligence Organization, pp. 4819–4825. DOI: 10.24963/ijcai.2020/671. URL: https://www.ijcai.org/proceedings/2020/671 (visited on 12/02/2020).

Potts, Richard (2013). "Hominin evolution in settings of strong environmental variability." en. In: *Quaternary Science Reviews* 73, pp. 1–13. DOI: 10.1016/j.quascirev.2013.04.003. URL: https://www.sciencedirect.com/science/article/pii/S0277379113001340 (visited on 11/24/2021).

Quian Quiroga, Rodrigo (2016). "Neuronal codes for visual perception and memory." en. In: *Neuropsychologia* 83, pp. 227–241. DOI: 10.1016/j.neuropsychologia.2015.12.016. URL: https://linkinghub.elsevier.com/retrieve/pii/S0028393215302578 (visited on 03/02/2022).

Raaijmakers, Steven F. et al. (2019). "Effects of self-assessment feedback on self-assessment and task-selection accuracy." en. In: *Metacognition and Learning* 14.1, pp. 21–42. DOI: 10.1007/s11409-019-09189-5. URL: http://link.springer.com/10.1007/s11409-019-09189-5 (visited on 07/04/2021).

Ranjbar-Slamloo, Yadollah and Zeinab Fazlali (2020). "Dopamine and Noradrenaline in the Brain; Overlapping or Dissociate Functions?" In: *Frontiers in Molecular Neuroscience* 12, p. 334. DOI: 10.3389/fnmol.2019.00334. URL: https://www.frontiersin.org/article/10.3389/fnmol.2019.00334 (visited on 10/19/2021).

Reinke, Chris, Mayalen Etcheverry, and Pierre-Yves Oudeyer (2020). "Intrinsically Motivated Discovery of Diverse Patterns in Self-Organizing Systems." In: *arXiv:1908.06663 [cs, stat]*. arXiv: 1908.06663. URL: http://arxiv.org/abs/1908.06663 (visited on 11/24/2021).

Robinson, M. J. F. et al. (2016). "Roles of "Wanting" and "Liking" in Motivating Behavior: Gambling, Food, and Drug Addictions." en. In: *Behavioral Neuroscience of Motivation*. Ed. by Eleanor H. Simpson and Peter D. Balsam. Current Topics in Behavioral Neurosciences. Cham: Springer International Publishing, pp. 105–136. DOI: 10.1007/7854_2015_387. URL: https://doi.org/10.1007/7854_2015_387 (visited on 10/25/2021).

Rolf, Matthias, Jochen J. Steil, and Michael Gienger (2010). "Goal Babbling Permits Direct Learning of Inverse Kinematics." In: *IEEE Transactions on Autonomous Mental Development* 2.3. Conference Name: IEEE Transactions on Autonomous Mental Development, pp. 216–229. DOI: 10.1109/TAMD.2010.2062511.

Rouault, Marion, Peter Dayan, and Stephen M. Fleming (2019). "Forming global estimates of self-performance from local confidence." en. In: *Nature Communications* 10.1, p. 1141. DOI: 10.1038/s41467-019-

09075-3. URL: http://www.nature.com/articles/s41467-019-09075-3 (visited on 01/22/2021).

Rozenblit, Leonid and Frank Keil (2002). "The misunderstood limits of folk science: an illusion of explanatory depth." In: *Cognitive science* 26.5, pp. 521–562. DOI: 10.1207/s15516709cog2605_1. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3062901/ (visited on 02/14/2022).

Ryan, R. M. and E. L. Deci (2000). "Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being." eng. In: *The American Psychologist* 55.1, pp. 68–78. DOI: 10.1037//0003-066x.55.1.68.

Ryan, Richard M. and Edward L. Deci (2017). *Self-determination theory: Basic psychological needs in motivation, development, and wellness*. Self-determination theory: Basic psychological needs in motivation, development, and wellness. Pages: xii, 756. New York, NY, US: The Guilford Press. DOI: 10.1521/978.14625/28806.

– (2020). "Intrinsic and extrinsic motivation from a self-determination theory perspective: Definitions, theory, practices, and future directions." en. In: *Contemporary Educational Psychology* 61, p. 101860. DOI: 10.1016/j.cedpsych.2020.101860. URL: https://linkinghub.elsevier.com/retrieve/pii/S0361476X20300254 (visited on 03/04/2021).

Saegusa, Ryo et al. (2009). "Active motor babbling for sensorimotor learning." In: *2008 IEEE International Conference on Robotics and Biomimetics*, pp. 794–799. DOI: 10.1109/ROBIO.2009.4913101.

Santucci, Vieri G., Gianluca Baldassarre, and Marco Mirolli (2013). "Which is the best intrinsic motivation signal for learning multiple skills?" en. In: *Frontiers in Neurorobotics* 7. DOI: 10.3389/fnbot.2013.00022. URL: http://journal.frontiersin.org/article/10.3389/fnbot.2013.00022/abstract (visited on 01/22/2021).

Sauvé, Sarah A and Marcus T Pearce (2019). "Information-theoretic Modeling of Perceived Musical Complexity." In: *Music Perception: An Interdisciplinary Journal* 37.2, pp. 165–178.

Schaul, Tom et al. (2016). "Prioritized Experience Replay." en. In: *arXiv:1511.05952 [cs]*. arXiv: 1511.05952. URL: http://arxiv.org/abs/1511.05952 (visited on 01/22/2021).

Schmidhuber, J. (1991a). "A possibility for implementing curiosity and boredom in model-building neural controllers." In: DOI: 10.7551/mitpress/3115.003.0030.

Schmidhuber, Jürgen (2010). "Formal theory of creativity, fun, and intrinsic motivation (1990–2010)." In: *IEEE Transactions on Autonomous Mental Development* 2.3, pp. 230–247.

Schmidhuber, Jürgen (1991b). "Curious model-building control systems." In: *[Proceedings] 1991 IEEE International Joint Conference on Neural Networks*, 1458–1463 vol.2. DOI: 10.1109/IJCNN.1991.170605.

Schueller, William, Vittorio Loreto, and Pierre-Yves Oudeyer (2018). *Complexity Reduction in the Negotiation of New Lexical Conventions*. arXiv: 1805.05631 [cs.MA].

Schultz, Wolfram (2016). "Dopamine reward prediction error coding." en. In: *Dialogues in Clinical Neuroscience* 18.1, p. 10.

Schultz, Wolfram, Peter Dayan, and P Read Montague (1997). "A Neural Substrate of Prediction and Reward." en. In: 275, p. 7.

Schulz, Eric and Samuel J. Gershman (2019). "The algorithmic architecture of exploration in the human brain." en. In: *Current Opinion in Neurobiology* 55, pp. 7–14. DOI: 10.1016/j.conb.2018.11.003. URL: https://linkinghub.elsevier.com/retrieve/pii/S0959438818300904 (visited on 01/22/2021).

Schulz, Eric et al. (2019). "Structured, uncertainty-driven exploration in real-world consumer choice." en. In: *Proceedings of the National Academy of Sciences* 116.28, pp. 13903–13908. DOI: 10.1073/pnas.1821028116. URL: http://www.pnas.org/lookup/doi/10.1073/pnas.1821028116 (visited on 03/04/2021).

Schwab, Ivan R. (2018). "The evolution of eyes: major steps. The Keeler lecture 2017: centenary of Keeler Ltd." en. In: *Eye* 32.2, pp. 302–313. DOI: 10.1038/eye.2017.226. URL: http://www.nature.com/articles/eye2017226 (visited on 10/07/2021).

Schwartenbeck, Philipp et al. (2019). "Computational mechanisms of curiosity and goal-directed exploration." en. In: *eLife* 8.e41703, p. 45.

Schwartz, Bennett L. and Janet Metcalfe (2011). "Tip-of-the-tongue (TOT) states: retrieval, behavior, and experience." en. In: *Memory & Cognition* 39.5, pp. 737–749. DOI: 10.3758/s13421-010-0066-8. URL: http://link.springer.com/10.3758/s13421-010-0066-8 (visited on 01/02/2022).

Shannon, Claude E. (1948). "A mathematical theory of communication." In: *The Bell System Technical Journal* 27.3. Conference Name: The Bell System Technical Journal, pp. 379–423. DOI: 10.1002/j.1538-7305.1948.tb01338.x.

Sharma, Nikhil (2008). *(5) (PDF) The Origin of Data Information Knowledge Wisdom (DIKW) Hierarchy*. URL: https://www.researchgate.net/publication/292335202_The_Origin_of_Data_Information_Knowledge_Wisdom_DIKW_Hierarchy (visited on 10/18/2021).

Shin, Dajung Diane and Sung-il Kim (2019). "Homo Curious: Curious or Interested?" en. In: *Educational Psychology Review* 31.4, pp. 853–874. DOI: 10.1007/s10648-019-09497-x. URL: http://link.springer.com/10.1007/s10648-019-09497-x (visited on 01/22/2021).

Silvetti, Massimo et al. (2021). *A Reinforcement Meta-Learning Framework of Executive Function and Information Demand*. en. Tech. rep. Section: New Results Type: article. bioRxiv, p. 2021.07.18.452793. DOI: 10.1101/2021.07.18.452793. URL: https://www.biorxiv.org/content/10.1101/2021.07.18.452793v1 (visited on 03/02/2022).

Singh, Satinder et al. (2010). "Intrinsically Motivated Reinforcement Learning: An Evolutionary Perspective." en. In: *IEEE TRANSACTIONS ON AUTONOMOUS MENTAL DEVELOPMENT* 2.2, p. 13.

Smith, Kyle S. et al. (2009). "Ventral Pallidum Roles in Reward and Motivation." In: *Behavioural brain research* 196.2, pp. 155–167. DOI: 10.1016/j.bbr.2008.09.038. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2606924/ (visited on 10/25/2021).

Son, L. K. and J. Metcalfe (2000). "Metacognitive and control strategies in study-time allocation." eng. In: *Journal of Experimental Psychology. Learning, Memory, and Cognition* 26.1, pp. 204–221. DOI: 10.1037//0278-7393.26.1.204.

Son, Lisa K and Rajiv Sethi (2006). "Metacognitive Control and Optimal Learning." en. In: *Cognitive Science*, p. 16.

Speekenbrink, Maarten and Emmanouil Konstantinidis (2015). "Uncertainty and Exploration in a Restless Bandit Problem." en. In: *Topics in Cognitive Science* 7.2, pp. 351–367. DOI: 10.1111/tops.12145. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/tops.12145 (visited on 11/01/2021).

Spetch, M L et al. (1990). "Suboptimal choice in a percentage-reinforcement procedure: effects of signal condition and terminal-link length." In: *Journal of the Experimental Analysis of Behavior* 53.2, pp. 219–234. DOI: 10.1901/jeab.1990.53-219. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1323008/ (visited on 10/27/2021).

Stalnaker, Thomas A, Nisha K Cooch, and Geoffrey Schoenbaum (2015). "What the orbitofrontal cortex does not do." en. In: *Nature Neuroscience* 18.5, pp. 620–627. DOI: 10.1038/nn.3982. URL: http://www.nature.com/articles/nn.3982 (visited on 05/31/2021).

Sterling, Peter (2012). "Allostasis: a model of predictive regulation." eng. In: *Physiology & Behavior* 106.1, pp. 5–15. DOI: 10.1016/j.physbeh.2011.06.004.

Stout, Andrew and Andrew G. Barto (2010). "Competence progress intrinsic motivation." In: *2010 IEEE 9th International Conference on Development and Learning*. ISSN: 2161-9476, pp. 257–262. DOI: 10.1109/DEVLRN.2010.5578835.

Sutton, Richard S and Andrew G Barto (2018). *Reinforcement learning: An introduction*. MIT press.

Takahashi, Kuniyuki et al. (2017). "Dynamic motion learning for multi-DOF flexible-joint robots using active–passive motor babbling through deep learning." In: *Advanced Robotics* 31.18. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01691864.2017.1383939, pp. 1002–1015. DOI: 10.1080/01691864.2017.1383939. URL: https://doi.org/10.1080/01691864.2017.1383939 (visited on 11/24/2021).

Tang, Haoran et al. (2017). "#Exploration: A Study of Count-Based Exploration for Deep Reinforcement Learning." In: *Advances in*

Neural Information Processing Systems 30. Ed. by I Guyon et al. Curran Associates, Inc., pp. 2750–2759.

Ten, Alexandr et al. (2021). "Humans monitor learning progress in curiosity-driven exploration." en. In: *Nature Communications* 12.1. Bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Human behaviour;Psychology Subject_term_id: human-behaviour;psychology, p. 5972. DOI: 10.1038/s41467-021-26196-w. URL: https://www.nature.com/articles/s41467-021-26196-w (visited on 11/04/2021).

Tenenbaum, Joshua B (1999). "Bayesian modeling of human concept learning." en. In: *Advances in Neural Information Processing Systems* 11, p. 7.

Thrun, S. (1994). "A lifelong learning perspective for mobile robot control." In: *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'94)*. Vol. 1, 23–30 vol.1. DOI: 10.1109/IROS.1994.407413.

Todorov, Emanuel and Michael I. Jordan (2002). "Optimal feedback control as a theory of motor coordination." eng. In: *Nature Neuroscience* 5.11, pp. 1226–1235. DOI: 10.1038/nn963.

Townsend, Corinne L and Evan Heit (2011a). "Judgments of learning and improvement." en. In: p. 13.

– (2011b). "Metacognitive Judgments of Improvement are Uncorrelated with Learning Rate." en. In: p. 6.

Trewavas, Anthony (2005). "Plant intelligence." en. In: *Naturwissenschaften* 92.9, pp. 401–413. DOI: 10.1007/s00114-005-0014-9. URL: http://link.springer.com/10.1007/s00114-005-0014-9 (visited on 10/04/2021).

Tsutsui, Ako and Gentarow Ohmi (2011). "Complexity Scale and Aesthetic Judgments of Color Combinations." en. In: *Empirical Studies of the Arts* 29.1, p. 15.

Twomey, Katherine E. and Gert Westermann (2018). "Curiosity-based learning in infants: a neurocomputational approach." en. In: *Developmental Science* 21.4, e12629. DOI: 10.1111/desc.12629. URL: http://doi.wiley.com/10.1111/desc.12629 (visited on 12/02/2020).

Volkow, Nora D. et al. (2002). ""Nonhedonic" food motivation in humans involves dopamine in the dorsal striatum and methylphenidate amplifies this effect." en. In: *Synapse* 44.3. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sy pp. 175–180. DOI: 10.1002/syn.10075. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/syn.10075 (visited on 10/25/2021).

Wang, Maya Zhe and Benjamin Y. Hayden (2019). "Monkeys are curious about counterfactual outcomes." In: *Cognition* 189, pp. 1–10. DOI: 10.1016/j.cognition.2019.03.009. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8029581/ (visited on 10/27/2021).

White, J. Kael et al. (2019). "A neural network for information seeking." en. In: *Nature Communications* 10.1. Bandiera_abtest: a Cc_license_type:

cc_by Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Cognitive neuroscience;Motivation Subject_term_id: cognitive-neuroscience;motivation, p. 5168. DOI: 10.1038/s41467-019-13135-z. URL: https://www.nature.com/articles/s41467-019-13135-z (visited on 10/14/2021).

Wilson, Robert C. et al. (2014). "Orbitofrontal Cortex as a Cognitive Map of Task Space." en. In: *Neuron* 81.2, pp. 267–279. DOI: 10.1016/j.neuron.2013.11.005. URL: https://linkinghub.elsevier.com/retrieve/pii/S0896627313010398 (visited on 10/15/2021).

Wulf, Gabriele and Rebecca Lewthwaite (2016). "Optimizing performance through intrinsic motivation and attention for learning: The OPTIMAL theory of motor learning." en. In: *Psychonomic Bulletin & Review* 23.5, pp. 1382–1414. DOI: 10.3758/s13423-015-0999-9. URL: http://link.springer.com/10.3758/s13423-015-0999-9 (visited on 01/15/2022).

Yan, Veronica X., Elizabeth Ligon Bjork, and Robert A. Bjork (2016). "On the difficulty of mending metacognitive illusions: A priori theories, fluency effects, and misattributions of the interleaving benefit." en. In: *Journal of Experimental Psychology: General* 145.7, pp. 918–933. DOI: 10.1037/xge0000177. URL: http://doi.apa.org/getdoi.cfm?doi=10.1037/xge0000177 (visited on 12/16/2021).

Yeager, David Scott and Carol S. Dweck (2012). "Mindsets That Promote Resilience: When Students Believe That Personal Characteristics Can Be Developed." en. In: *Educational Psychologist* 47.4, pp. 302–314. DOI: 10.1080/00461520.2012.722805. URL: http://www.tandfonline.com/doi/abs/10.1080/00461520.2012.722805 (visited on 11/10/2021).

Yu, Angela J. and Peter Dayan (2005). "Uncertainty, Neuromodulation, and Attention." en. In: *Neuron* 46.4, pp. 681–692. DOI: 10.1016/j.neuron.2005.04.026. URL: https://linkinghub.elsevier.com/retrieve/pii/S0896627305003624 (visited on 01/22/2021).

Yu, Angela and Peter Dayan (2003). "Expected and Unexpected Uncertainty: ACh and NE in the Neocortex." en. In: *Advances in Neural Information Processing Systems*. Vol. 15. MIT Press, p. 8.

Zimmerman, Barry J., Dale H. Schunk, and Maria K. DiBenedetto (2017). "The role of self-efficacy and related beliefs in self-regulation of learning and performance." In: *Handbook of competence and motivation: Theory and application, 2nd ed*. New York, NY, US: The Guilford Press, pp. 313–333.

van Bergen, Ruben S. et al. (2015). "Sensory uncertainty decoded from visual cortex predicts behavior." en. In: *Nature Neuroscience* 18.12. Number: 12 Publisher: Nature Publishing Group, pp. 1728–1730. DOI: 10.1038/nn.4150. URL: https://www.nature.com/articles/nn.4150 (visited on 03/01/2022).

van Lieshout, Lieke L.F. et al. (2018). "Induction and Relief of Curiosity Elicit Parietal and Frontal Activity." en. In: *The Journal of Neuroscience*

38.10, pp. 2579–2588. DOI: 10.1523/JNEUROSCI.2816-17.2018. URL: http://www.jneurosci.org/lookup/doi/10.1523/JNEUROSCI.2816-17.2018 (visited on 04/20/2021).

van Lieshout, Lieke, Floris de Lange, and Roshan Cools (2020). *Curiosity: An appetitive or an aversive drive?* en. preprint. PsyArXiv. DOI: 10.31234/osf.io/s3zp4. URL: https://osf.io/s3zp4 (visited on 05/31/2021).