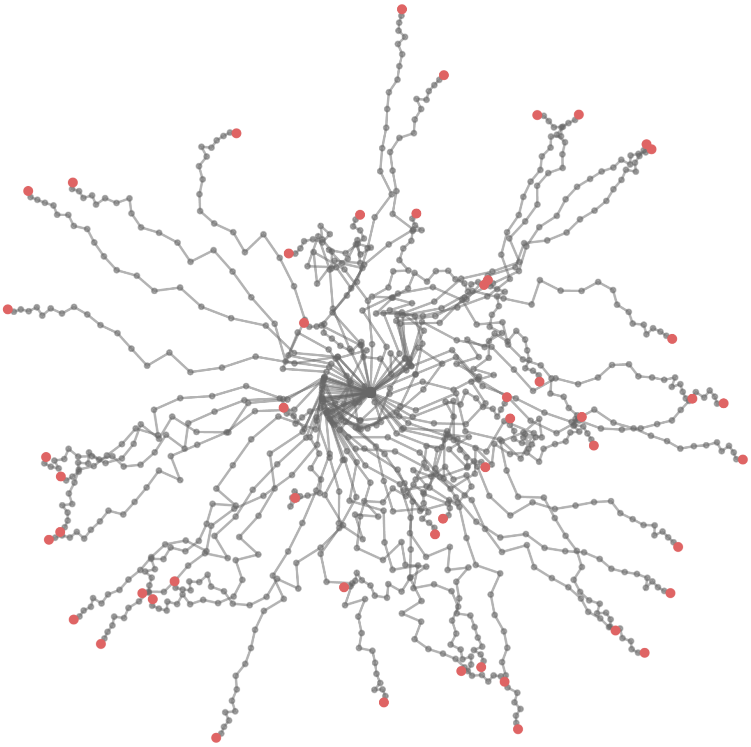


Thesis presented in order to obtain the title of
Docteur de l'Université de Bordeaux

Mathématiques et Informatique
Spécialité Informatique

by Fabien C. Y. Benureau

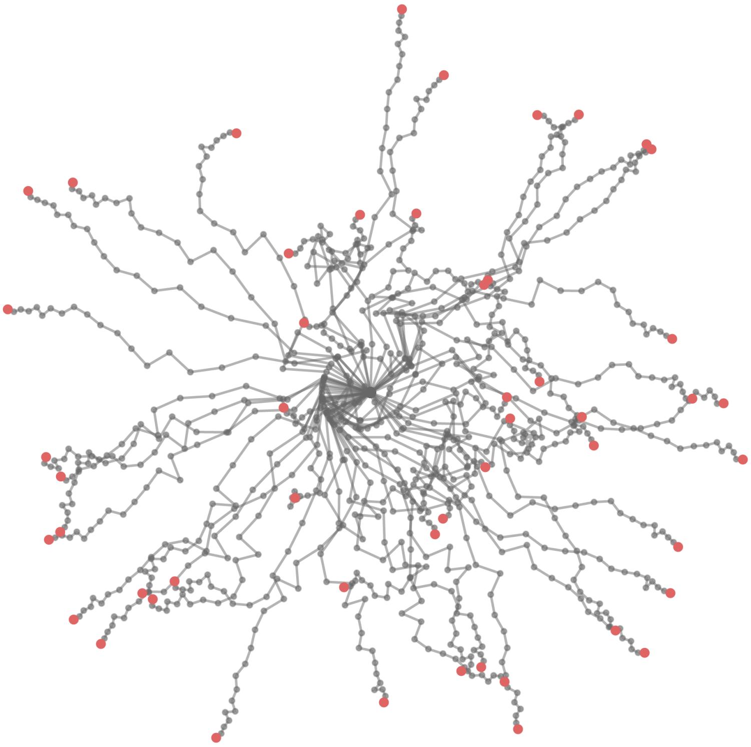
Self-Exploration of Sensorimotor Spaces in Robots

under the direction of Pierre-Yves Oudeyer, D.R.,
Flowers Team, Inria Bordeaux Sud-Ouest, ENSTA Paritech

president Pierre Bessière, D.R., *ISIR, UPMC*
reviewer Verena V. Hafner, Prof. Dr., *Institut für Informatik, Humboldt-Universität zu Berlin*
examiner Jean-Baptiste Mouret, C.R., *Larsen Team, Inria Nancy*
examiner Manuel Lopes, C.R., *Flowers Team, Inria Bordeaux Sud-Ouest, ENSTA Paritech*

Defended the 18th of May, 2015

université
de BORDEAUX



Thèse présentée pour obtenir le grade de

Docteur de l'Université de Bordeaux

Mathématiques et Informatique
Spécialité Informatique

par Fabien C. Y. Benureau

L'auto-exploration des espaces sensorimoteurs chez les robots

sous la direction de Pierre-Yves Oudeyer, D.R.,
Flowers Team, Inria Bordeaux Sud-Ouest, ENSTA Paritech

president	Pierre Bessière, D.R., *ISIR, UPMC*
rapporteur	Verena V. Hafner, Prof. Dr., *Institut für Informatik, Humboldt-Universität zu Berlin*
examinateur	Jean-Baptiste Mouret, C.R., *Larsen Team, Inria Nancy*
examinateur	Manuel Lopes, C.R., *Flowers Team, Inria Bordeaux Sud-Ouest, ENSTA Paritech*

Soutenue le 18 mai 2015

université de BORDEAUX

Abstract

Developmental robotics has begun in the last fifteen years to study robots that have a childhood—crawling before trying to run, playing before being useful—and that are basing their decisions upon a lifelong and embodied experience of the real-world.

In this context, this thesis studies sensorimotor exploration—the discovery of a robot's own body and proximal environment—during the early developmental stages, when no prior experience of the world is available. Specifically, we investigate how to generate a diversity of effects in an unknown environment. This approach distinguishes itself by its lack of user-defined reward or fitness function, making it especially suited for integration in self-sufficient platforms.

In a first part, we motivate our approach, formalize the exploration problem, define quantitative measures to assess performance, and propose an architectural framework to devise algorithms. Through the extensive examination of a multi-joint arm example, we explore some of the fundamental challenges that sensorimotor exploration faces, such as high-dimensionality and sensorimotor redundancy, in particular through a comparison between motor and goal babbling exploration strategies. We propose several algorithms and empirically study their behaviour, investigating the interactions with developmental constraints, external demonstrations and biologically-inspired motor synergies. Furthermore, because even efficient algorithms can provide disastrous performance when their learning abilities do not align with the environment's characteristics, we propose an architecture that can dynamically discriminate among a set of exploration strategies.

Even with good algorithms, sensorimotor exploration is still an expensive proposition—a problem since robots inherently face constraints on the amount of data they are able to gather; each observation takes a non-negligible time to collect.

In a second part, we propose the *reuse* algorithm that allows to exploit the exploration trajectories of a previous environment in another new, unknown one, to improve exploration, with the only constraining assumptions being that the two environments share the same motor space—which is often the case as a robot's body remains similar across tasks. No assumption is made that the sensory modalities of the two tasks remain identical, or that the exploration strategies or the learning algorithms are the same. If the latent dynamics of the two environment share some degree of similarity, we establish that the *reuse* algorithm provides improvements in exploration. We illustrate this on a real robot setup interacting with different objects in augmented

reality.

We then show that the *reuse* algorithm can exhibit scaffolding behaviour. This allows to guide skill acquisition through the exclusive manipulation of environments where no reward or fitness function needs to be defined. Additionally, we conduct experiments that show that exploration on real-world robots can benefit from reusing exploration trajectories produced on surrogate, simplified—even purely kinematic—simulations.

Throughout this thesis, our core contributions are algorithms description and empirical results. In order to allow unrestricted examination and reproduction of all our results, the entire code is made available.

Sensorimotor exploration is a fundamental developmental mechanism of biological systems. By decoupling it from learning and studying it in its own right in this thesis, we engage in an approach that casts light on important problems facing robots developing on their own.

Pour le lecteur francophone, un résumé détaillé est disponible: "Directions d'exploration".

Abstract en français

La robotique développementale a entrepris, au courant des quinze dernières années, d'étudier les processus dévelopmentaux, similaires à ceux des systèmes biologiques, chez les robots. Le but est de créer des robots qui ont une enfance—qui rampent avant d'essayer de courir, qui jouent avant de travailler—et qui basent leurs décisions sur l'expérience de toute une vie, incarnés dans le monde réel.

Dans ce contexte, cette thèse étudie l'exploration sensorimotrice—la découverte pour un robot de son propre corps et de son environnement proche—pendant les premiers stage du développement, lorsque qu'aucune expérience préalable du monde n'est disponible. Plus spécifiquement, cette thèse se penche sur comment générer une diversité d'effets dans un environnement inconnu. Cette approche se distingue par son absence de fonction de récompense ou de fitness définie par un expert, la rendant particulièrement apte à être intégrée sur des robots auto-suffisants.

Dans une première partie, l'approche est motivée et le problème de l'exploration est formalisé, avec la définition de mesures quantitatives pour évaluer le comportement des algorithmes et d'un cadre architectural pour la création de ces derniers. Via l'examen détaillé de l'exemple d'un bras robot à multiple degrés de liberté, la thèse explore quelques unes des problématiques fondamentales que l'exploration sensorimotrice pose, comme la haute dimensionalité et la redondance sensorimotrice. Cela est fait en particulier via la comparaison entre deux stratégies d'exploration: le babillage moteur et le babillage dirigé par les objectifs. Plusieurs algorithmes sont proposés tour à tour et leur comportement est évalué empiriquement, étudiant les interactions qui naissent avec les contraintes développementales, les démonstrations externes et les synergies motrices. De plus, parce que même des algorithmes efficaces peuvent se révéler terriblement inefficaces lorsque leurs capacités d'apprentissage ne sont pas adaptés aux caractéristiques de leur environnement, une architecture est proposée qui peut dynamiquement choisir la stratégie d'exploration la plus adaptée parmi un ensemble de stratégies.

Mais même avec de bons algorithmes, l'exploration sensorimotrice reste une entreprise coûteuse—un problème important, étant donné que les robots font face à des contraintes fortes sur la quantité de données qu'ils peuvent extraire de leur environnement; chaque observation prenant un temps non-négligeable à récupérer.

Dans une deuxième partie, l'algorithme *reuse* est proposé. Il permet d'exploiter dans un nouvel environnement inconnu les trajectoires d'explorations établies dans

un précédent environnement. L'objectif est d'améliorer l'exploration du nouvel environnement, avec l'unique contrainte que les deux environnements doivent partager le même espace moteur—ce qui est souvent le cas, étant donné que le corps d'un robot a tendance à rester similaire lors du passage d'une environnement à un autre. Aucune supposition contraignante n'est faite sur les espaces sensoriels des deux environnements, qui peuvent différer arbitrairement ; il en va de même pour les stratégies d'exploration et les algorithmes d'apprentissage. Si les dynamiques latentes des deux environnements sont similaires, l'algorithme *reuse* peut apporter une amélioration de l'exploration. Ceci est illustré sur un robot réel, qui interagit avec différents objets en réalité augmentée.

Une expérience permet ensuite de montrer que l'algorithme *reuse* peut démontrer une capacité à permettre l'acquisition de savoir-faire complexes, se reposant sur des savoir-faire plus simples. Cela permet de guider l'acquisition de savoir-faire en manipulant exclusivement l'environnement dans lequel le robot est plongé, sans avoir besoin de créer une fonction de récompense ou de fitness. De plus, des expériences sont conduites qui montrent que l'exploration dans le monde réel peut bénéficier de la réutilisation de trajectoires d'exploration obtenues en simulation, même si celles-ci sont simplifiées de manière importante.

À travers cette thèse, les contributions les plus importantes sont les descriptions algorithmiques et les résultats expérimentaux. De manière à permettre la reproduction et la réexamination sans contrainte de tous les résultats, l'ensemble du code est mis à disposition.

L'exploration sensorimotrice est un mécanisme fondamental du développement des systèmes biologiques. La séparer délibérément des mécanismes d'apprentissage et l'étudier pour elle-même dans cette thèse permet d'éclairer des problèmes importants que les robots se développant seuls seront amenés à affronter.

Contents

List of Figures

To anyone who will ever dare read this.

And to my little sister, whose accomplishments have already dwarfted mine; they ever-so-slightly make me aim a little higher than I would every time. I am more proud of her that she knows.

And to my brother, who is courageously finding its own path. As I am so very lucky to have found mine, I have only one thing to say: Marshall on! The view is pretty exciting up there.

And to my parents, for anything and really, everything.

Acknowledgments

To my family, my advisor, my colleagues.

To my advisor, Pierre-Yves Oudeyer, who has been more patient with me than I could hope for during the *five* years it took me to finish this few hundreds of pages. Pierre-Yves offered me, in the Flowers team, the time and freedom to grow as a scientist; I am profoundly thankful and equally indebted. The mutually agreed repayement plan, if I am not mistaken, is ongoing creation of new—and hopefully exciting—research.

To Paul Fudal, whose help, advices, support and ingenuity was instrumental in the experiments.

To my flowery team, past and present. Each and every one of you has enriched my experience during these last five years. I will always remember Thomas Degris, ever-willing to get into a sparring match about reinforcement learning, AI or really anything—I miss our talks. Adrien Baranes, who bootstrapped much of the research strands I am now wandering, and whose singing is dearly missed in the lab. Thomas Cedeborg, larger than life, whose earlier political activism transpired in much of his way of being, and led to many fascinating exchanges. Clément Moulin-Frier, deceptively smart, bon vivant, a real peer. I know that our paths will cross again at conferences, and the prospect makes me thrilled. Olivier Mangin, who has held more than one time my life in his hands. Camping without a tent in the middle of November in wine and boar-infested country, or in a prarie between two mountains with wild horses gallopping all night will never fade away. Jonathan Grizou. Finding similarly minded scientist is rare enough: I am happily bracing for future emulation, mate. Jérôme Béchu, first engineer of the team, made me feel at home in the lab right from the beginning. Working alongside you was always fun and interesting. Sao Mai, who, during my first month, offered me a roof until I found mine. Pierre Rouanet, true backbone of the team from its inception; he provided the fundations for some of the most impressive work we do at Flowers, and his technical expertise constantly sets a standard. Matthieu Lapeyre, who comes from the same town as me, and almost singlehandedly birthed a humanoid. His PhD monstruous impact is inspiring: not only did he stake new ground, but he got people to follow him there. William Schueller, who patiently gave me access to a physicist perspective on my work, and to an exceptional multi-cultural one on life. Sebastien Forestier, a curious mind from which I predict many great things. Alexandra Darrieutort, who expands our research horizon, and always brings delicious cakes or crepes that bring the team together. I'll race ahead and watch my back; I know you're coming. Benjamin Clément, who rarely ever refused a coffee-break—even when I was high on sleep-deprivation—, and allowed my curious self to observe his work with children. Didier Roy, which always has a spark in his eyes, and a joyfully mocking remark ready. His work with children and robots is testament to an forever open and curious mind. Steve Nguyen, that readily came help me teach robotics when our course became so popular the number of students exceeded the number of chairs. Baptiste Busch and Yohan Mollard, who have replaced me as dwellers of our experimental room. Baptiste is always jovial and up for a good laugh. Yohan provides a precious no-nonsense and focused approach to engineering. Thibaud Munzer, officemate, always up for discussing the definition of learning or the origins of abstractions. His calm demeanour hides a critical mind. Haylee Fogg, who crossed half the planet, with her engineering expertise and ever positive attitude, to come work with us.

To Nicolas Jahier, Nathalie Robin and Marie Sanchez each of who have consistently shielded us from much administrative complexity, and helped us get occasionally propulsed around the planet.

And Manuel Lopes, who I am prevented to say anything about, for he is on the jury—and I am grateful for that. I'll say this: the look he gave me the first time I mispelled his name, I only did it once.

To the many fantastic members of our neighbor teams, Phoenix and Potioc.

To Inria Bordeaux Sud-Ouest, who gave me a scientific home. Being able to do science is always a privilege, but the wonderful and beautiful working place at Inria make it a pleasure to. The importance of the terrific work of administrative staff on my ability to focus on doing science cannot be overstated, and I would like to particularly thank them.

To the ENS Lyon and the French government, whose historical idiosyncrasies created a extraordinary system that afforded me unique opportunities while granting me at the same time with a healthy dose of humility.

> *To all the researchers that came before me, and graciously lent me their shoulders to stand on, and without who this work would be very different.*

Open Science

New ideas may happen a long way away from data, but justifications, in science, should never be far from it. You can expect a lot of figures and plots in the following pages. Presenting a plot without making the code that generated it available amounts to asking the reader to trust and believe the author to do what he say we does, and to be able to know that he does, in fact, what we say he does. In other words, a naked plot amounts to have to assume that malice or incompetence are absent.

We would rather ask our reader to doubt everything, to examine every detail, to reproduce, to challenge. This is why we provided the code for *every plot that appears in the following pages*. And the code is not far away. No need to send mail or to roam websites. A link overlays the plot picture, and is additionally present in the caption. The exact piece of code for this figure is exactly one click away. You might sometimes need a cluster or a hardware setup to reproduce the results, but many plots only need a few minutes to be generated. Installing the programs for those plots should not take much more time. We encourage the reader to regenerate some of those plots, and to poke experiments, to tweak the parameters, and try new values. To find, in fact, flaws that our diligent work missed. Don't hesitate to contact us in that case.

Many of our experiments rely on the whims of random number generators which need to be initialized. We might have looked over the problem, and settled for a random initialization at the time of execution. We decided to run our experiments under fixed seeds. That provides the advantage to be able to reproduce some simulation plots exactly[1]. If an experiment is only run once, the seed is 0. For repeated experiments, random seeds have been generated using code that is itself available.

All the code is available under the Open Science License, which include all provisions of the LGPL with the addition of the following statement:

> *If you publicly release any scientific claims or data that were supported or generated by the Program or a modification thereof, in whole or in part, you will release any modifications you made to the Program. This License will be in effect for the modified program.*

The Open Science License ensures that uses and modification of this code throughout the scientific community remain available, reproducible, and verifiable to all.

[1] Because our programs are in Python, which guarantees that the random sequences for the same seed don't vary across versions, anyone should obtain the same plots.

The english version of this text is found here: "Exploring Ahead".

Directions d'exploration

Les roboticiens sont des démiurges.

Ils inventent les corps, les esprits qui les habitent, et le plus souvent, créent de toute pièces le monde qui entourent ces derniers.

De fait, les roboticiens sont eux-mêmes leur plus formidable obstacle.

Le risque, en effet, est que les roboticiens, créateurs à la fois des problèmes et des solutions, adaptent les problèmes aux solutions, et non le contraire. Cela peut mener à inventer et à étudier des problèmes artificiels, qui contribuent peu à l'avancée de la science, tout en évitant systématiquement les problèmes difficiles, en les modifiant en des versions plus simples d'eux-mêmes chaque fois qu'un obstacle un tant soit peu insurmontable est rencontré.

Mais il existe un autre risque, plus pernicieux, et plus fondamental. C'est d'élaborer des robots depuis une perspective humaine, en choisissant des caractéristiques qui font sens pour l'observateur extérieur, mais qui n'en ont aucun pour le robot lui-même et pour son expérience égocentrique du monde. En d'autres termes, le risque est que les caractéristiques qui font que les robots sont faciles à élaborer, à contrôler et à comprendre pour les humains rendent difficile pour le robot lui-même l'interaction avec le monde, et limitent fondamentalement ses capacités.

Une illustration de ce phénomène est trouvée dans la manière de créer des robots : typiquement, le corps, la partie matérielle, est créée et finie avant que la partie logicielle ne commence à être conceptualisée. Cela permet de découpler les deux activités, et de fait, les deux savoir-faire, a priori différents. Et cela permet de se débarrasser de la myriade d'interactions qui seraient à prendre en compte si le corps et l'esprit du robot étaient conceptualisés ensemble. Le logiciel joue ici le rôle du fantôme dans la machine, l'investissant et l'animant après qu'elle ai été créé. Les organismes biologiques ne fonctionnent pas de cette manière. Une telle organisation est certainement adaptée à la programmation des robots des lignes d'assemblage. Mais cette manière de procéder, celle utilisée pour programmer les applications des téléphones portables, est répétée pour des plateforme qui représentent l'état de l'art de la recherche en robotique, telle que l'iCub, le robot PR2 ou Baxter. Les chercheurs utilisant ces plateformes doivent trouver comment programmer des produits matériels achevés et difficilement reconfigurable[2].

Ce paradigme fonctionne bien pour certaines lignes de recherche, mais est problématique pour d'autres, tel que la locomotion à pattes. L'élaboration de jambes divorcées des algorithmes de marche qui seront utilisés pour les actionner a produit des robots qui requièrent des algorithmes précis, à faible latence et gourmands en puissance de calcul, tout en était peu robustes à la moindre perturbation inattendue de leur environnement.

La robotique évolutionnaire a attaqué ce problème directement, en proposant des algorithmes inspirés de la sélection naturelle pour automatiser la conceptualisation des robots à partir de l'évaluation directe de leur comportement, permettant à la morphologie et aux programmes de contrôle de s'adapter l'un à l'autre.

Cela étant, même en robotique évolutionnaire, un élément humain clé reste présent dans le processus d'élaboration: la fonction de fitness. Elle encode le but du processus évolutionnaire, et est entièrement décidée par l'expérimentateur, avant le début du processus, souvent d'une manière extrêmement spécifique. Elle peut par exemple représenter la distance parcourue par le robot au cours d'un intervalle de temps. La conséquence la plus immédiate est de créer des robots qui ne savent faire qu'une seule chose. Cela est aggravé par la tendance des algorithmes évolutionnaires à souvent se révèler plus malins que l'expérimentateur, en produisant des robots dont le comportement maximise la fonction de fitness tout en étant complètement inacceptable pour l'usage prévu dans l'esprit de l'expérimentateur. Les exemples en ce sens incluent l'exploitation de bugs dans le simulateur physique, et la production de robots qui couvrent la plus grande distance parce qu'ils acceptent de s'autodétruire pour aller plus vite. Ces considérations sont de sérieux problèmes qui s'ajoutent au fait que définir un goal n'est pas nécessairement la meilleure manière d'y parvenir, comme le montrent les travaux de Stanley and Lehman (2015).

Cependant, le problème le plus fondamental est autrepart : il tient dans le fait même que l'expérimentateur choisit les buts que le robot va poursuivre avec un zèle infaillible. Le problème est qu'il n'est pas clair à quel point l'expérimentateur humain est qualifié, ou même à la bonne place, pour décider des buts d'un robot, une entité possédant une incarnation et des processus cognitifs complètement différents des humains.

C'est sur ce point que certaines lignes de recherche de la robotique développementale cherchent à se démarquer du reste de la robotique. Elles étudient des robots qui doivent créer eux-mêmes les buts qu'ils poursuivent, en utilisant leur propres *système motivationnels*.

La robotique développementale est née de la réalisation que créer des robots "adultes", avec des capacités et des savoirs préformés, fonctionnels dès la sortie de la ligne d'assemblage était trop difficile. La programmation du sens commun, par exemple, a prouvé être

[2] Un effet notable de cette approche est que de nombreux travaux en robotique listent comme aspect positif de leur travail la capacité de s'adapter à n'importe quel robot, quelque soit son incarnation matérielle. Bien que semblant désirable, les conséquences plus larges d'un tel enjeu de recherche en font un but potentiellement dangereux.

remarquablement laborieuse. Observant que les humains acquièrent naturellement leur sens commun pendant l'enfance, il a été proposé de créer des robots "enfants", qui seraient équipés de capacités d'apprentissage leur permettant d'acquérir des savoirs et savoir-faire qui feraient sens pour eux, pour leur propre corps et leur propre environnement.

Les systèmes motivationnels, à leur tour, sont aux objectifs décidés par un expérimentateur ce que les capacités d'apprentissage sont au savoir-faire préformés. Ce sont des fabriques à objectifs, de la même manière que les capacités d'apprentissage sont des fabriques à savoir-faire. Ils permettent aux robots de se créer des buts qui sont adaptés à leur propre corps, leur environnement, et leur niveau d'expérience actuel. Les systèmes motivationnels se couplent aussi naturellement avec les capacités d'apprentissage, parce qu'il y a trop de choses à apprendre dans des environnements même modérément complexes; ils permettent de sélectionner quelles activités poursuivre, et de fait, quoi apprendre et quoi ne pas apprendre.

Tout cela nous amène au sujet de cette thèse : l'exploration. Les robots qui choisissent leur propre buts, qui acquièrent des savoir-faire par eux-même ont besoin d'explorer leur environnement, et ceci pour deux raisons. La première, de manière à acquérir de l'expérience, qui peut à son tour être utilisée pour modifier leur comportement (c'est le processus d'apprentissage). La seconde raison c'est que l'exploration sert à découvrir de nouveaux buts à poursuivre.

Dans cette thèse, nous nous sommes concentrés sur l'exploration des espaces sensorimoteurs, c'est-à-dire les espaces qui permettent d'exprimer la relation entre une action motrice et le retour sensoriel qui lui correspond. De plus, nous avons uniquement considéré l'exploration qui est conduite par le robot lui-même, sans guidage social ou savoir externe. D'où le titre: "L'auto-exploration des espaces sensorimoteurs chez les robots".

La thèse a trois objectifs. Le premier est d'établir l'exploration comme un problème scientifique. Le second est d'étudier certaines stratégies d'explorations simples, et l'impact que différentes variations ont sur elles, de manière à permettre de construire une intuition sur l'exploration et de former une base sur laquelle des stratégies plus élaborées peuvent être construites. Le troisième objectif est de commencer à explorer comment les capacités d'exploration d'un robot peuvent s'améliorer au cours du temps, à l'aide de l'expérience accumulée. Cette thèse remplit ces trois objectifs, même si seulement de manière spécifique.

Pour établir l'exploration en robotique comme un problème scientifique, on commence, dans le chapitre 1, au tout début : la définition d'un robot, l'impact que le fait d'avoir un corps a sur l'expérience que le robot a du monde, et pourquoi tous les problèmes en robotique ne peuvent être résolus par des simulations suffisamment ambitieuses du monde réel dans la tête du robot. La conclusion est que pour être efficace dans la partie non-structurée du monde réel, les robots ont besoin de passer par une longue phase de développement, de sorte à créer pièce par pièce les savoir-faire, les connaissances et le sens commun nécessaire pour faire face aux imprévus des situations dans lesquelles ils se trouveront dans le futur. Pendant cette phase de développement, leur capacité d'exploration sont cruciales.

Ensuite, nous formalisons le problème de l'exploration : explorer est créer accès à différents aspects de l'environnement. L'exploration n'est pas seulement spatiale : un robot peut explorer la réaction d'un objet avec lequel il interagit, comme par exemple les différents sons que l'objet est capable de produire. Cette définition de l'exploration nous permet de mettre en évidence une distinction importante entre l'exploration et l'apprentissage. Apprendre est modifier son comportement grâce à son expérience. Cela fait de l'apprentissage un concept distinct de l'exploration ; on peut apprendre sans explorer : c'est ce qu'un système de prévision météo fait. Et on peut explorer sans apprendre : c'est ce que font les robots aspirateurs, qui arrivent à couvrir une pièce sans jamais en apprendre la forme. Bien sûr, la plupart du temps, on désire combiner l'apprentissage et l'exploration.

Maintenant, pour faire de l'exploration un problème scientifique, il est nécessaire d'avoir un moyen de l'évaluer de manière quantitative. Puisque l'exploration crée accès à différents aspects de l'environnement, une manière de l'évaluer est de mesurer la *diversité* des réponses sensorielles que le robot est capable de générer. La diversité est une bonne mesure pour nombre de raisons : c'est un concept qui s'adapte à beaucoup de domaines, c'est une mesure intrinsèque; le robot lui-même est capable de la mesurer, sans perturber son comportement—ce qui n'est pas possible de faire si on veut évaluer, par exemple, sa capacité de prédiction. Ceci permet, de manière additionnelle, d'envisager partager des dispositifs expérimentaux avec d'autres domaines dans lesquels inspecter le processus de réflexion de l'explorateur est difficile, tel que les sciences cognitives.

Cela nous amène à un point important et inévitable : l'état de l'art des travaux similaires aux nôtres. La notion d'exploration et de diversité a bénéficié de peu d'attention explicite en robotique, en dehors de l'exploration spatiale[3]. Mais beaucoup de domaines proches mènent des travaux qui se rapportent aux nôtres. En robotique développementale, l'étude des motivations intrinsèques est pertinente; la diversité peut être utilisée et est utilisée dans certains algorithmes de cette thèse en tant que motivation intrinsèque. De plus, un intérêt croissant au cours des quinze dernières années a été observé pour la diversité comme mesure et outil algorithmique en informatique, dans des disciplines aussi variées que les ensemble de classification, l'optimisation par essaims particulaires et les systèmes de recommandation. En science cognitives, la diversité comportementale a fait l'objet de nombreux travaux, même si la quasi-totalité des données quantitatives ont été collectées sur des expériences d'exploration spatiale.

Dans le cadre de nos expériences, nous avons introduit une mesure de diversité appelée couverture-τ. Elle mesure le volume de l'union des balles de rayon τ centrées autour des points de retour sensoriel observés lors de l'exploration. Si le retour sensor-

[3] L'exploration spatiale est un cas spécifique d'exploration, pour lequel le déplacement dans l'espace sensorimoteur est déjà maitrisé.

iel est diversifié, les points sont loin les uns des autres, et la superposition des balles est faible : le volume de leur union est élevé. Si le retour sensoriel est peu diversifié, la superposition est importante et le volume moins important pour le même nombre de points.

Le deuxième but de cette thèse a été d'étudier les algorithmes d'exploration. L'idée ici a été de choisir l'un des algorithmes le plus simple possible, et de l'étudier sous différentes conditions. La simplicité de l'algorithme a été justifiée par deux facteurs. Le premier est que cela permettait de comprendre les résultats dans leur moindre détails sans devoir suspendre l'intuition du lecteur. Le comportement de la régression linéaire locale, ou d'algorithmes plus complexes en espaces à haute dimensions peut se révèler complexe, et c'est pourquoi nous avons opté pour une méthode plus simple, basée sur la perturbation de plus proche voisins. Et deuxièmement, en restant simple, l'espoir est que l'intuition gagnée puisse être réutilisée dans un champs plus large de situations qu'un algorithme plus complexe et plus spécifique.

L'une des premières contributions de l'étude a été de clarifier qu'explorer l'espace moteur était inefficace à cause des contributions *combinées* de la haute dimensionalité *et* de la distribution hétérogène de la redondance de l'espace sensorimoteur (c'est-à-dire, le nombre d'actions motrices différentes qui produisent le même retour sensoriel). La haute dimensionalité seule n'est pas suffisante pour rendre l'exploration de l'espace moteur inefficace.

Ensuite, nous avons analysé de manière systématique les contributions de chaque aspect de l'algorithme. L'impact de la distribution des buts a été étudiée, soulignant le potentiel que les méthodes qui dirigent leur buts représentent (Dans la majorité de cette thèse, les buts sont choisis de manière aléatoire). Les effets d'un mauvais modèle inverse ont été démontrés, et un algorithme pour l'exploration d'espaces sensoriels non-bornés a été introduit.

Les expériences suivantes se sont concentrées à démontrer comment des implémentations même rudimentaires de synergies motrices, de contraintes développementales et de démonstrations externes pouvaient avoir un impact positif sur l'exploration. Une leçon à retenir est qu'améliorer l'incarnation des robots offre potentiellement des gains à la fois plus larges et moins coûteux que d'améliorer les capacités d'apprentissage.

Jusqu'ici, toutes les variations algorithmiques étudiées n'ont pas fait usage explicite de mesures de motivation intrinsèque. La diversité a été utilisée seulement comme un outil d'évaluation. Au chapitre 4, nous introduisons un algorithme qui utilise la diversité pour choisir laquelle des stratégies d'exploration utiliser parmi plusieurs, et nous démontrons que cette méthode permet de s'adapter à différentes situations aussi bien que n'importe quelque mixture fixe de stratégies.

Le troisième but était d'étudier des moyens d'améliorer les capacités d'exploration du robot au cours du temps, à mesure que l'expérience s'accumule.

Pour comprendre l'enjeu sous-jacent, il faut considérer qu'une exploration réussie d'un environnement donné doit donner accès à différents aspects de cet environnement, c'est-à-dire, du point de vue du robot, produire une diversité de retours sensoriels. Pour produire une diversité de retours sensoriels de manière efficace, une connaissance de la dynamique de l'environnement est nécessaire, de manière à éviter sa redondance inhérente, c'est-à-dire, pour éviter d'exécuter des actions qui produisent les mêmes effets. Pousser et tirer sur une porte fermée est un bon exemple : ce sont deux actions différentes qui produisent le même effet—et donc aucune diversité sensorielle—et qui apportent une connaissance nouvelle de la dynamique de l'environnement. Si l'état de la porte avait été connu dès le départ, le robot aurait pu se concentrer sur des actions différentes, plus susceptibles de créer de la diversité. Cela explique la problématique de l'oeuf et de la poule qui touche la production de diversité: les connaissances nécessaires pour conduire une exploration efficaces sont les connaissances que l'exploration est censée produire en premier lieu. Cela signifie que le processus d'exploration peut s'auto-entretenir, mais peu aussi rester bloqué dans l'incapacité de produire des interactions suffisamment informatives, menant à de longues périodes d'exploration pauvre au début du processus dans les environnements difficiles.

C'est ce qui nous a poussé à trouver une solution pour améliorer l'exploration, notamment lors des premières phases de contact avec un nouvel environnement. Pour ce faire, nous avons introduit la méthode *reuse*, qui réutilise l'expérience acquise dans un environnement précédent pour en explorer un nouveau. Le coeur de l'idée est de sélectionner des commandes motrices qui ont produit une diversité de retours sensoriels dans l'environnement précédent, et de les réexecuter dans le nouveau. Cette méthode a le bénéfice d'être conceptuellement simple, et d'être agnostique aux modalités sensorielles de l'un ou de l'autre environnement, qui peuvent être arbitrairement différentes. La stratégie d'exploration et l'algorithme d'apprentissage utilisés dans l'environnement précédent n'ont pas besoin non plus d'être les mêmes que ceux de l'environnement actuel : la méthode peut réutiliser des données collectées de manière arbitraire. La seule contrainte est que les commandes motrices exécutées dans l'environnement précédent puissent être réexécutés dans le nouveau.

La logique derrière la méthode *reuse* peut être comprise en considérant comment la redondance fait en sorte que deux commandes motrices différentes produisent le même effet sur l'environnement: soit par redondance du corps, soit par redondance de l'environnement. La redondance du corps fait en sorte que deux commandes motrices différentes produisent les mêmes mouvements: le robot applique donc les mêmes forces sur l'environnement. La redondance environnementale fait en sorte que des forces différentes produisent le même effet, comme illustre l'exemple de la porte fermée. Typiquement, des effets différents parviennent à éviter à la fois la redondance de l'environnement et celle du corps. En changeant d'environnement, la redondance environnementale n'est pas conservée, mais celle du corps l'est, la plupart du temps. De plus, si les environnements sont similaires, une partie de la redondance environnementale est partagée. Ainsi, en réutilisant un ensemble de commandes motrices qui ont généré une diversité d'effets, la méthode *reuse* capitalise le savoir gagné sur la redondance corporelle, et de manière opportuniste sur la redondance environnementale.

Pour valider expérimentalement cette idée, nous avons conduit des expériences qui ont démontré la viabilité de cette approche sur un robot réel manipulant différents objets en réalité augmentée. Les résultats montrent que la méthode *reuse* est efficace lorsqu'elle réutilise l'expérience gagnée par la manipulation d'un objet (une balle) pour explorer un autre objet possédant un comportement significativement différent (un cube). La méthode est aussi robuste à des environnements non-similaires, lorsque la diversité créée dans un environnement ne se transfère pas bien à un autre. De plus, nous avons établi que choisir les commandes motrices à réutiliser via une mesure de diversité était plus efficace que de le faire de manière aléatoire.

Dans les expériences précédentes, *reuse* améliore les performances au début de l'exploration. Mais après suffisamment de temps, que *reuse* soit utilisée ou pas, le processus d'exploration arrive à des résultats similaires. Pour montrer que *reuse* peut faire plus qu'améliorer les performances pendant une durée limitée, nous avons élaboré une expérience qui montre que *reuse* peut rendre explorable un environnement qui ne l'est pas à premier abord. Un aspect intéressant de cette expérience est que l'exploration est façonnée non pas par une fonction de récompense externe, mais pas une manipulation de l'environnement et de la saliance des objets qui sont contenus dedans, de la même manière qu'une personne s'occupant d'un enfant pourrait faire.

Enfin, nous nous sommes intéressé à l'application de la méthode *reuse* à des situations où l'exploration de l'environnement précédent s'est déroulée entièrement en simulation, tandis que l'exploration dans le nouvel environnement se déroule dans le monde réel, sur un vrai robot. Transférer les résultats obtenus en simulation à la réalité a prouvé être une tâche difficile en robotique, un problème connu sous le nom de *reality gap*. Les résultats obtenus, quoique demandant d'être approfondis, sont excellents. Ils laissent entrevoir la possibilité d'utiliser des simulations grossières de la réalité comme des artifices cognitifs efficaces pour une exploration améliorée du monde réel.

Ainsi se termine cette thèse. Où aller, à partir de là ? Il y a trois directions de recherche qui se dégagent: la diversité en robotique, la recherche interdisciplinaire avec les sciences cognitives, et la robotique évolutionnaire et développementale.

Premièrement, la diversité en robotique. En 1255, dans son Commentaire sur les Sentences, Thomas d'Aquin avança le point suivant: un ange a plus de valeur qu'une pierre. Mais de là, on ne peut pas conclure que deux anges ont plus de valeur qu'un ange et une pierre[4]. Une version modernisée de l'argument de Thomas d'Aquin est proposé par Nehring et al. (2002) : "Un humain a plus de valeur qu'un chimpanzé. Mais de là, on ne peut pas déduire que 6000130000 humains et aucun chimpanzés ont plus de valeur que 6000000000 humains et 130000 chimpanzés." En d'autre termes, la diversité a de la valeur. Une telle observation peut être faite dans des domaines

[4] Pour ceux qui lisent le latin dans le texte: "*quod quamvis Angelus absolute sit melior quam lapis, tamen utraque natura est melior quam altera tantum*" (Lib. 1 d. 44 q. 1 a. 2 ad 6)

aussi différents que la biodiversité, l'art, la composition d'équipes, les portefeuilles d'investissement, les résultats des moteurs de recherche, les ensembles de classification, et même, le progrès scientifique. Dans Lehman, Clune et al. (2014), Pierre-Yves Oudeyer a remarqué que "parce qu'on ne comprend pas encore suffisamment ce qu'est l'intelligence, ou comment produire une intelligence artificielle générale, plutôt que de couper des directions de recherches, pour vraiment faire des progrès, nous devrions embrasser l'"anarchie de méthodes" de l'intelligence artificielle." En d'autres termes, lorsqu'on tâtonne dans le noir, la diversité est un outil précieux.

Il est tentant ici d'appliquer cette leçon à la robotique développementale, et c'est ce que tente de faire cette thèse: les robots développementaux, plongés dans la complexité du monde réel, et avec aucun autre choix que d'en faire sens à l'aide de leur capacités d'apprentissage et d'exploration, doivent tâtonner dans le noir pendant un moment. La quasi-absence de travaux sur la diversité en robotique développementale n'est pas à la hauteur du potentiel qu'elle promet d'apporter.

Il y a, cependant, de nombreuses façons d'abuser cette leçon. Premièrement, la diversité pour elle-même est difficilement justifiable, quoi que soit sa valeur intrinsèque. In particulier, un système motivationnel seulement dirigé par la recherche de diversité semble être une mauvaise idée. Certains ont avancés l'argument qu'étant donné que le nombre de choses simples est en quantité limitée, une exploration dirigée par la diversité conduira naturellement à découvrir des phénomènes de plus en plus complexes. La rareté de la simplicité, cependant, n'a jamais été justifiée en dehors d'exemples jouets, et les choses simples à découvrir et à apprendre semblent être en quantité suffisante dans le monde réel pour remplir plusieurs vies. Tout cela conspire à suggérer que les systèmes de motivation robotiques devraient favoriser une *diversité* de motivations, dont la diversité ferait partie. Des motivations en compétition et complémentaires devraient mener à des comportements alternant entre des phases d'exploration, lors desquelles de nouveaux aspects du monde sont découverts, et des phases d'étude concentrée sur un sujet, où des savoir-faire spécifiques seraient acquis.

Deuxièmement, le problème se pose de la manière utiliser l'expérience collectée lors d'une exploration dirigée par la diversité. Dans cette thèse, nous avons montré, via la méthode *reuse*, que cette expérience est précieuse pour explorer de nouveaux environnements. Mais explorer n'est pas le seul comportement qu'un robot développemental possède. La question de capitaliser et réappliquer l'expérience obtenue grâce à la diversité pour résoudre des problèmes précis reste ouverte, avec la question de savoir si une telle expérience est compétitive avec des approches plus directionnelles.

Enfin, de nombreux problèmes spécifiques à propos de la diversité n'ont pas encore de réponse satisfaisante. L'exploration dirigée par l'exploration diffère de l'exploration dirigée par la nouveauté en cela que les approches dirigées par la nouveauté ne peuvent pas contrôler explicitement la quantité de diversité qu'elles produisent. Maintenir une certaine quantité de diversité comportementale, en particulier lorsque l'environnement change et réduit les options disponibles pour le robot, peut uniquement être obtenu par la perspective globale qu'offre la diversité, et non avec la seule perspective locale de la nouveauté. Toutefois, la diversité requière plus de ressources computationnelles : quand est-elle nécessaire par rapport aux approches plus simples liées à la nouveauté ? Quelles sont de bonnes mesure de diversité ? La diversité a-t-elle un sens en haute dimension, où doit-elle est en permanence supportée par des représentations abstraites de faible dimension ?

Répondre à ces questions n'est pas facile ; une source d'intuition possible est offerte par les sciences cognitives. Comment les enfants utilisent-ils la diversité pendant leur développement ? C'est la deuxième direction de recherche qui semble prometteuse.

Il est notable que parmi toute la littérature disponible sur le jeu, l'exploration et la résolution de problèmes chez les enfants et les animaux, les mesures quantitatives de diversité des interactions qu'ils engagent avec le monde et des solutions qu'ils tentent sont pratiquement absentes. Les études s'arrêtent souvent à de vagues descriptions qualitatives. Des études quantitatives sur l'utilisation de la diversité comportementale dans l'exploration pourraient jeter une lumière précieuse sur les meilleures manières d'élaborer des systèmes motivationnels pour les robots. De plus, cette ligne de recherche, en vertu de sa méthodologie compatible, promet de permettre de conduire des expériences similaires sur des humains et des robots, menant potentiellement à des échanges et une émulation fructueuse entre les deux domaines.

La troisième direction de recherche est la robotique évolutionnaire et développementale, aussi appelée "évo-dévo-robo". La robotique évolutionnaire mimique le processus de sélection naturelle, tandis que la robotique développementale mimique le développement morphologique et cognitif des systèmes biologiques. La quasi-totalité de travaux, aussi similaires qu'ils puissent être, sont restés séparés jusqu'à maintenant. Étant donné l'intérêt porté à la création d'une intelligence artificielle similaire aux capacités humaines, cette séparation est déconcertante; après tout, les seuls exemples connus d'entités possédant une intelligence similaire à l'intelligence humaine ont été créés par une combinaison de ces deux processus.

Combiner la robotique évolutionnaire et la robotique développementale pose un énorme problème: le temps. Les échelles de temps du développement et de l'évolution—les durée de vie et les aeons respectivement—mettent déjà en difficulté leurs disciplines respectives. Combiner les deux semble donc totalement insurmontable. La bonne manière de voir ce problème est de remarquer que la taille de ce problème est telle que les progrès technologiques des, disons, 50 prochaines années ne la feront pas diminuer. En d'autres termes, attendre n'aide pas.

Une autre objection est d'avancer l'argument que les robotiques évolutionnaire et développementale sont deux jeunes disciplines, et ne sont pas encore suffisamment mûres pour être combinées. Bien que majoritaire spéculatif, cet argument pourrait se révéler être vérifié. Mais les difficultés rencontrées dans les tentatives de combinaison pourraient jeter une lumière précieuse sur les limitations de l'un ou l'autre domaine qui pourraient se révéler difficiles à découvrir autrement.

Ajouter une longue phase de développement à la robotique évolutionnaire pourrait voir l'émergence de nouvelles dynamiques, plus complexes, dans le processus évolutionnaire, et une amélioration du processus de sélection. Le travail de Bongard (2011) a jeté les premières bases, en montrant que le développement morphologique pouvait agir comme un véritable tamis, filtrant les comportements fragiles dans une expérience de locomototion à pattes. Dans l'autre sens, la robotique développementale pourrait bénéficier du processus évolutionnaire, qui pourrait réduire les décisions arbitraires que les chercheurs doivent faire pour le moment, notamment en terme des capacités de représentation et d'apprentissage qui sont données a priori aux robots.

La robotique évolutionnaire et développementale représente de toute évidence un défi formidable, et il est difficile de contester que les résultats potentiels le sont également. C'est un domaine d'investigation que l'on ne peut simplement pas se permettre de ne pas investir.

Au moins parce qu'il promet de réduire les tendances démiurgiques des roboticiens. Les roboticiens sont des démiurges ; l'évo-dévo-robo fait partie de la solution.

Part One

EXPLORATION

Writing is an exploration. You start from nothing and learn as you go.

E. L. Doctorow

I do not know what I may appear to the world, but to myself I seem to have been only like a boy playing on the sea-shore, and diverting myself in now and then finding a smoother pebble or a prettier shell than ordinary, whilst the great ocean of truth lay all undiscovered before me.

Isaac Newton

0

Exploration: A Simple Example

Let's consider the following problem: we are given a black-box that takes inputs and produces outputs. We know the values the inputs can take, but we don't know which outputs correspond to which inputs. We don't even know which outputs *can* be produced. We are given the opportunity to sample the black-box for a limited time. How much *diversity of effects* then, can be produced with the limited access we have?

This question defines an *exploration problem*. Here, the objective is to discover what effects—what outputs—the black-box is capable to deliver, and produce a representative subset of them. To answer such a problem is to provide an *exploration strategy*, i.e. a method that selects which inputs to try on the black box.

Let's take the example of a one-meter-long idealized robotic arm on a two-dimensional plane, made up of an open chain of joints linked by segments of equal length. The angles of the joints, which can take values between -150 and 150 degrees, uniquely define the posture of the arm, and therefore, the position of the end-effector.

Assuming the robotic arm is a black-box that accepts joint angles as input, and produces the resulting end-effector position as output, which *exploration strategy* can we consider to produce as much diverse end-effector positions as possible in a limited time?

A Tale of Two Exploration Strategies

The simplest exploration strategy is to try random angle values. For each motor command, the angle of each joint is chosen randomly between ± 150 degrees. This strategy

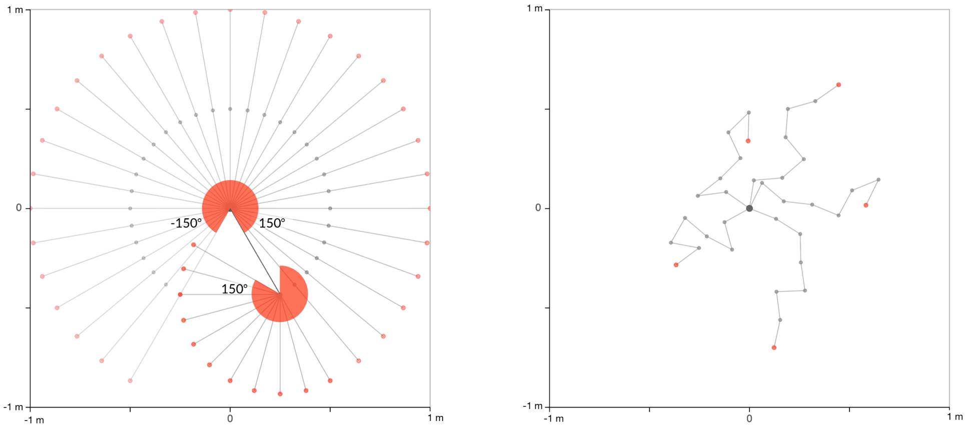


Figure 1: On the left, an illustration of the angle range for the two-joint arm. On the right, random postures for an arm with 7 segments, each of length 1/7th of a meter. [source code]

is known as *random motor babbling*[1].

Another strategy is to explore not the motor space but the sensory space. One approach is a *goal babbling*[2] strategy: goals—that is, points in the output space—are chosen, and motor commands must be found that produce effects that approach the goals as much as possible.

Of course, transforming goals into motor commands is not without problems: a goal can be impossible to reach—as would be for instance, a goal placed two meters away from our previously described robotic arm—, and, conversely, many different motor commands can satisfy the same goal. In other words, the *reachable space*—the set of effects that can be produced by the environment—might be a comparatively small subset of the considered *goal space*, and is possibly redundant.

In order to transform goals into motor commands, we use an *inverse model* of the environment. An inverse model provides a mapping from the sensory space—and therefore, the goal space—to the motor space. An inverse model is good when the motor commands it provides generates, when executed, effects as close as possible from their corresponding goals. In our black-box context, the inverse model cannot be known beforehand; it has to be learned incrementally, as the exploration progresses.

Let's note right away that the creation of inverse models is *not* the purpose of

[1] *Random* and *babbling* are not redundant terms here. Babbling was originally used to designate an infant seemingly meaningless production of vocalizations after the sixth week, and, later, to describe repetitive, and seemingly random, infant movements such as kicking. Those movements have since then proven to be far from random. We use *motor babbling* to describe the production of motor activations which are produced *for their own sake*, that is, with the purpose to find out what effect they yield. As such, *babbling* both implies that the action is not part of a planning strategy, and that its effect is not previously known to a satisfactory degree by the actor (i.e. babbling implies information seeking).

[2] In a fashion similar to motor babbling, *goal babbling* characterizes goals that are produced for their own sake, that is, with the purpose to find out if they can be reached.

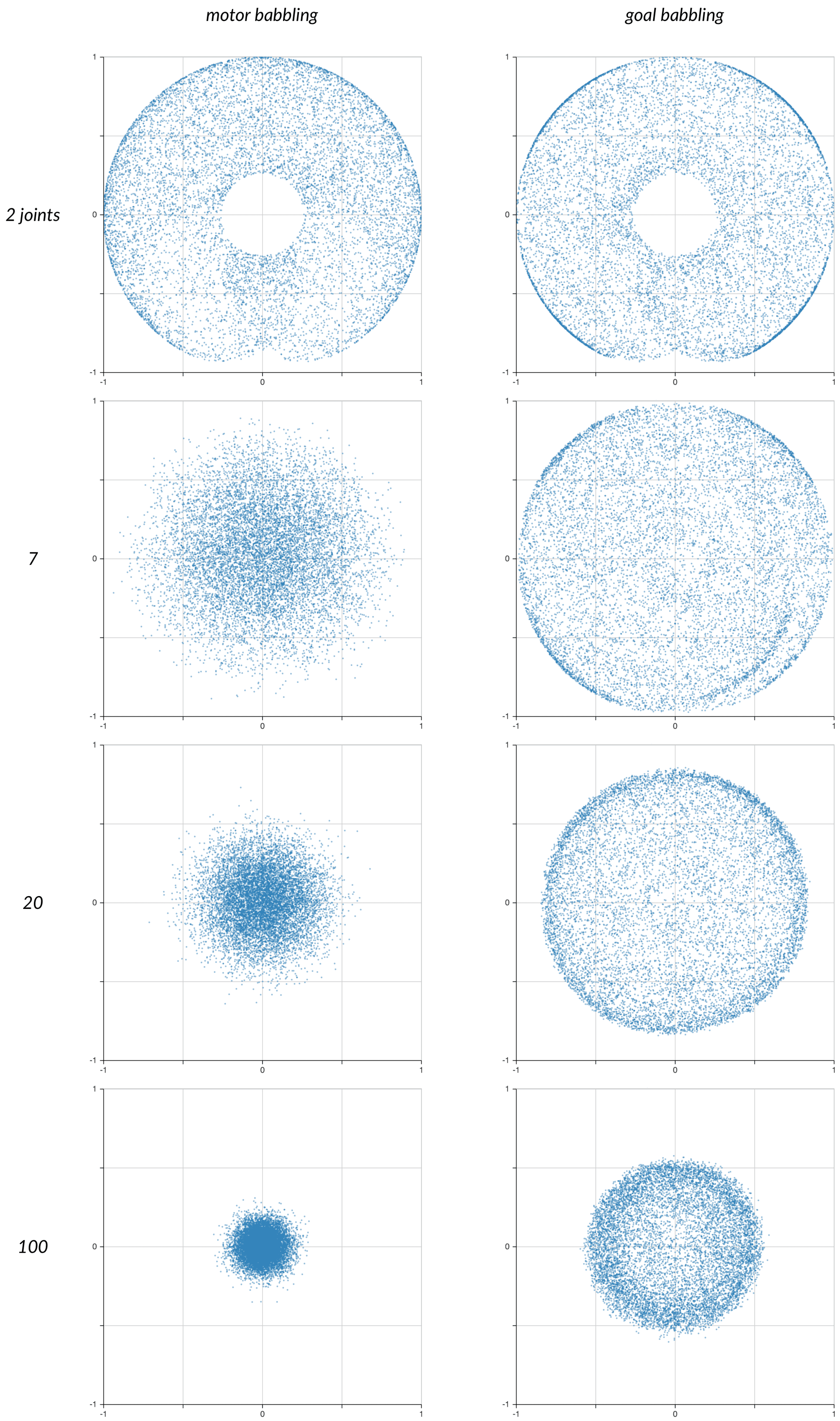


Figure 2: The goal babbling strategy is more effective than the motor babbling one at exploring the reachable space, especially when the robot arm possess a high number of joints. Still, the goal babbling strategy fails to explore all the reachable space when many joints are involved. The blue points represent the positions of the end-effector reached during exploration [source code]

the work we present in this thesis. The purpose is the study of behavioural diversity and exploratory behaviour. Whilst we use inverse model in many of the exploration strategies we study, they are considered here as tools, not ends.

For the purposes of our arm example, we employ a simple learning algorithm for the inverse model, since we want to stress that it is not the sophistication of the model but how exploration is conducted that makes the difference between motor and goal babbling.

Given a goal, the inverse model finds amongst previous observations the one with the nearest effect, then retrieves its motor command and adds a small random perturbation drawn within the legal range of the inputs to it, and returns the perturbated motor commands to be executed. Since such a learning algorithm relies on previous observations, the early phase of the exploration features a small number of random motor babbling steps in order to bootstrap the observations.

An Experiment

To compare the *random motor babbling* and *goal babbling* strategies, we consider four different arms, with 2, 7, 20 and 100 segments—the lengths of which are set so that the total length of the arm remains one meter across the configurations.

For the goal babbling strategy, random goals are created by drawing points randomly in the $[-1, 1]\times[-1, 1]$ square[3]. This strategy will be called *random goal babbling*, often abbreviated as *goal babbling* when no confusion is possible. The perturbation that the inverse model applies on motor commands is drawn uniformly from $\pm 5\%$ of the range on each joint value ($\pm 15°$).

Finally, the random goal babbling strategy is bootstrapped with 10 interactions of random motor babbling at the beginning of the exploration. For each arm, we run the two exploration strategies over 10000 interactions with the environment (henceforth, *timesteps* or *steps*) each.

Analysis

The results are available Figure 2. We observe a severe degradation of the area covered by the effects produced by the motor babbling strategy[4] as the number of joints increases. The 7-joint arm does not produce effects near the edge of the reachable space even after trying 10000 different postures. The 100-joint arm does not even cover a fourth of the reachable space.

[3] If you think that the goal space fitting the reachable space so well is highly spurious and actually straightforward cheating, you are deceptively perceptive. More details on this issue in section 3.2.

[4] The 'area covered by the effects produced by the motor babbling' is not yet a precise notion here—it will have to wait chapter 1. If one nevertheless needs one now, one shall for instance consider the area of the smallest disk containing all the effects.

This is easily explained. What random motor babbling is doing is providing an empirical estimation of the density of the redundancy[5] of the arm across the reachable space. For the 100-joint arm , the centre of the reachable space is orders of magnitude more dense than the outer edge, which leads to the distribution pattern of Figure 2. In sensorimotor spaces where the density of the redundancy is uniform, random motor babbling would produce an uniform distribution of effects over the reachable space, regardless of the level of redundancy[6].

This phenomenon is well illustrated by the 2-joint example. In Figure 3, the number of timesteps has been raised to 50000 compared to Figure 1: sensory areas where two solutions exists are twice as dense as the areas where only one does—as it should. In Figure 4, the two set of solutions have been separated, and overlaid with sample arm postures. One set of solution corresponds to arm postures where the second joint angle is positive, and the other where it is negative.

Under the goal babbling strategy, areas with different redundancy levels are explored uniformly. Along the edges of the reachable space in Figure 3 however, an increase of effect density can be observed. The reason for this is explained Figure 6:

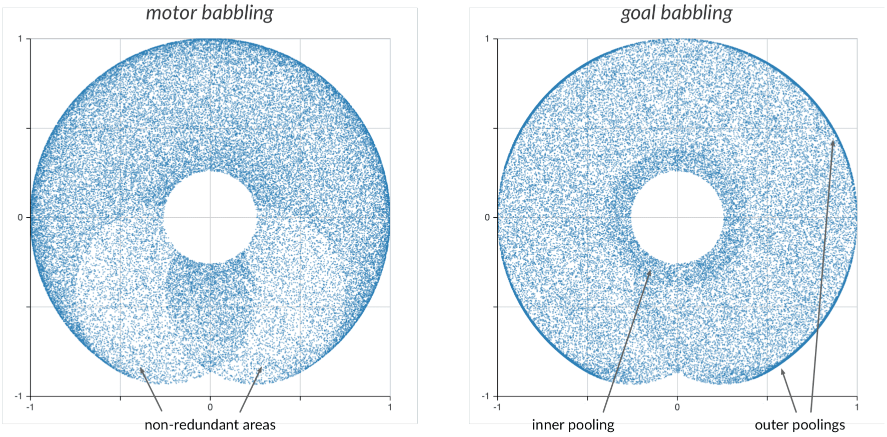


Figure 3: Goal babbling smoothes out the sensorimotor redundancy. The majority of the reachable space of the 2-joint arm is redundant. But due to the ±150° range of the joints, two areas where only one solution exists exist. While this difference in redundancy is clearly reflected in the motor babbling exploration, it is not present in the goal babbling one. Both figures show 50000 timesteps. [source code]

[5] For a discussion about the definition of redundancy see Conkur et al. (1997). In this thesis, we are interested in the difference between the redundancy of different areas of the sensory space. We define the redundancy of a subset B of the sensory space as the probability that an effect belong to B, given a random motor command, drawn uniformly from the motor space. Lenarcic (1999) provides an algorithm to quantify the redundancy of rigid, multijoint robotic arms, but the computation is only tractable for a small number of joints.

[6] A minor point of detail: here the robotic arm exhibits *kinematic* redundancy, i.e. there are more joints than necessary to obtain a given position of the end-effector. Musculoskeletal systems usually exhibit both *kinematic* and *kinetic* redundancy, where there are more muscles than required to apply the relevant forces on the joints. Typically, robots based on electrical motors do not exhibit kinetic redundancy, but those based on artificial muscles do. This thesis overwhelmingly uses examples of the former kind.

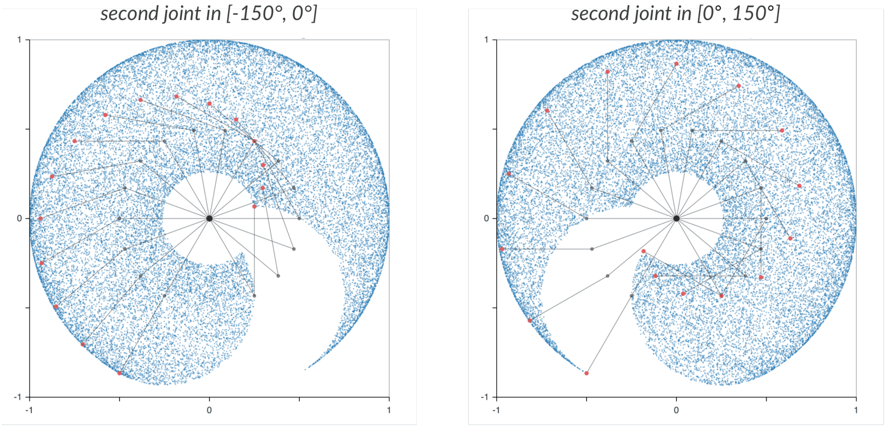


Figure 4: The two set of solutions are discriminated by the sign of the angle of the second joint. The two reachable areas are not superposable. Each figure shows 25000 samples. [source code]

because goals are chosen uniformly in the hyperrectangle $[0, 1]\times[0, 1]$, some goals are outside the reachable space. The effects resulting from these goals pool along the edges of it, in a pattern specific to the inverse model. Here, the perturbations induced by the inverse model on the motor commands produces a large impact on the position of the end-effector in the inner edge, and a correspondingly small one on the outer edge, as illustrated in Figure 6.

The goal babbling strategy consistently covers more of the reachable space than the motor babbling strategy, and in a more uniform manner. Nevertheless, the area covered by the strategy diminishes with the increase of the number of joints. This is due to the arm looping on itself more and more as the number of joints increases, as shown in Figure 5. These loops form an attractor for our perturbation-based inverse model. In simple cases, these local minima can sometimes be escaped. An example is given by the goal babbling exploration of the 7-joint arm in Figure 2. In the lower right quadrant, a first set of solutions left a visible pooling of effects, before being replaced

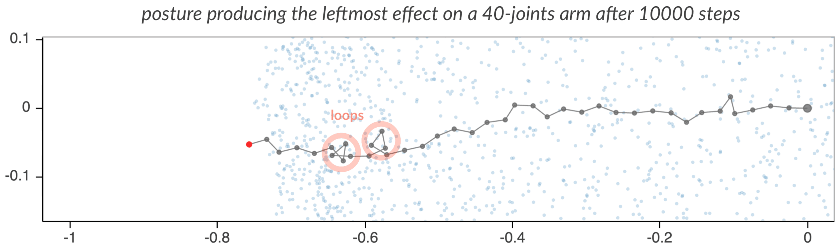


Figure 5: Loops appear on the arm as the number of joints increases, trapping the inverse model in local attractors. This is highly dependent on the initial motor commands produced during the random motor babbling bootstrapping phase, and reduces the areas covered, as seen in Figure 2. [source code]

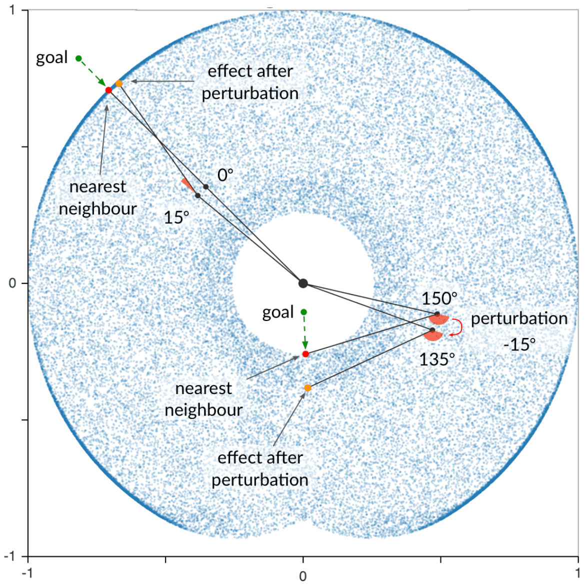


Figure 6: Goal babbling produces poolings of effects at the edges of the reachable space. The shape these poolings takes is a direct consequence of the inverse model. Here, if a goal is set near the centre, the nearest neighbour has his second joint angle near the limit, 150°; a perturbation of this angle produces a value in the range [150°, 135°], which produce a large effect on the position of the end effector. This leads to a distinctive inner ring of increased effect density. *A contrario*, a goal on the outer edge is associated with a nearest neighbour with a second angle joint near 0°, the perturbation of which (±15°) only results in a minor displacement of the end effector. As a result, we observe thin and dense strips of effects on the outer edges of the reachable space. [source code]

by better solutions coming from an adjacent area. The process is visible Figure 7.

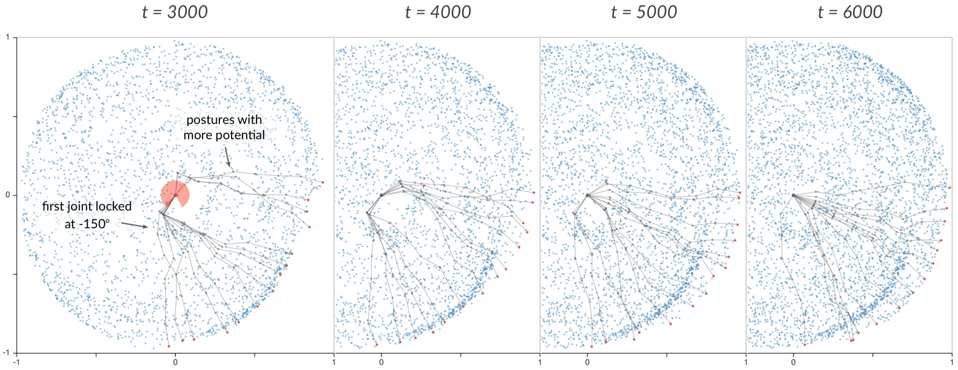


Figure 7: The progression of the goal babbling exploration of the 7-joint arm sees better solutions progressively replace solutions trapped in a local extrema because the first joint in locked at -150°. [source code]

This problem is linked to the minimal nature of the environment. In a more complex setting, it could be handled in any number of ways, the most reasonable of which would be to prevent the arm self-collisions, or having sensorimotor feedback of the convolutions of the arm, or having a natural rest pose. An unreasonable way would be to try to improve the learning algorithm to avoid this specific problem. As we'll discuss in the next chapters, each time the learning algorithms try to fight the complexity of the world with specific sophistication, there is probably a more ecological, more complication-frugal way to proceed. For instance, the physical world naturally prevents a robotic arm to pass through itself. Handling collisions would prevent loops, removing those local minimum from the sensorimotor space. Because of this problem is largely anecdotal in our context, we won't focus on addressing it in most of our experiments, but we will discuss parsimonious ways to deal with it in sections 3.6 and 3.7.

Despite the loops, the goal babbling strategy is better than the motor babbling strategy. It also benefits from the two-dimensionality of the sensory space, which remains so regardless of the morphology of the arm. Goal babbling separates the decisions about *what* to do from the ones about *how* to do it by making them in two different spaces, the sensory and motor space respectively. This affords goal babbling a direct, explicit way to encourage effect diversity by setting a diversity goals, thus fostering the objectives of an exploration problem.

So far we have shown that if a robot must explore its sensorimotor space without having any prior knowledge of it, several strategies can be conceived, and they produce significantly different results. In the next chapter, we will motivate why sensorimotor exploration is an important problem in robotics.

The Plan

This thesis is divided in two parts. The first part pursues two goals: first, it motivates the importance of studying sensorimotor exploration, defines exploration as a scientific problem, proposes an algorithmic framework, discuss the relevant previous work. Second, it systematically investigates variation of a simple exploration algorithm to provide a basis on which to consider more complex approaches. One of those, where a diversity metric is explicitely used as an intrinsic motivation for the exploration, is proposed at the end of the first part. The second part investigates how to improve exploratory behaviour with experience, and to do so introduces an method, *reuse*, to exploit exploration data from one environment to another.

Part One - Exploration

In chapter 1, we first expose the classic machine learning paradigm and contrast it with the interacting learning scenario of embodied robots. This allows us to make explicit the challenges robots face when learning. We take advantage of the discussion to expose and motivate the current trends in robotic research in which our work is set, in particular in relation with the notion of embodiment, development and evolution. We then propose to study the *exploration problems* robots face when discovering their body and their proximal environment, and we contrast it to studying learning problems. We introduce diversity measures to quantify and evaluate exploration, and present the *explorers* framework, that will express all the exploration architectures we develop throughout the thesis.

In chapter 2, we look at the existing literature on exploratory processes, exploratory behaviour and diversity. We begin with active learning, that proposed the first explicit algorithms for directed exploration. We then brush against the concept of self-organization, that underlies all biological organisms and many natural phenomena, and explicit how these processes create diversity, are challenging to predict, and are a promising venue for parsimonious robotic design. This also gives us the opportunity to discuss the concept of homeokinesis, a recently proposed method for sensorimotor exploration, and compare it to our approach. After that, we turn our attention to biology, and investigate sensorimotor exploration in fetuses, neonates and infants, and its relation to development. This leads us to the studies of exploratory behaviour in psychology, and in particular, to the theories on intrinsic motivation. We offer there a rapid historical perspective that leads us naturally to today, where computational approaches have joined the scientific dialogue. Amongst intrinsic motivations, novelty-based methods interest us particularly, for their direct relation the production of diversity. We finish our bibliographical review by a brief survey of the recent advances of evolutionary robotics that have advocated diversity as a robust fitness function, and highlight the steady rise in the use of diversity measure in computer science.

In chapter 3, we come back to the two-dimensional arm of chapter 0, and revisit many details previously ignored. The impact of goal distribution is investigated, and as a result, we provide exploration algorithms that can build and adapt the goal space during the exploration, removing the previously needed prior on bounds. Next, we modify the quality of the inverse model and study how it affects exploration, and argue than in some case, motor babbling is preferable to goal babbling, even when the heterogeneity of the redundancy is high. We then briefly discuss how motor synergies can improve exploration, and we provide a simple illustrative example. We repeat the same schema for developmental constraints and socially-provided demonstrations. As we discuss the merit of the presented work, we argue that effective exploration needs a multifaceted approach that combines many different phenomena.

In chapter 4, building on the challenge of the adequacy between the learning capability and the environment complexity discussed in the previous chapter, we propose an architecture that can dynamically choose among different exploration strategies by leveraging diversity as an explicit intrisic motivation. We illustrate the effectiveness of the methods on variations in learner quality and in exploration aggressiveness.

Part Two - Reuse

In chapter 5, we present the reuse method, that transfers motor commands from one task to another by enforcing diversity. A simple example of two kinematic planar arms is discussed.

In chapter 6, we discuss the existing literature on transfer learning, and formalize the reuse method. The two kinematic arms example of the previous chapter is analysed quantitatively, and we provide results that show that diversity produces consistently better performances than random reuse.

In chapter 7, an experimental setup with a real robot interacting with a virtual object is described, motivated, analysed, and critiqued. The feasibility of random motor babbling is discussed, and a series experiments shows that in the specific situation we study, reuse is both sensitive and resilient to task similarity, works even when the modalities are different, and can exploit random motor babbling data.

In chapter 8, we show that through environmental control alone, a diversity-driven agent can be guided towards sophisticated behaviours.

Finally in chapter 9, the reuse method is discussed in relation with the reality gap problem, and we show that even degraded models can be used with reuse to inform and improve exploratory behaviour.

Contribution

Our main contributions are:

- Defining and motivating the study of sensorimotor exploration in robots, as a critical part of developmental robotics.
- A detailed study of a simple example of sensorimotor exploration, that introduces different exploration algorithms under a single framework.
- A diversity-driven method for selecting exploration strategies in Multi-Armed Bandits contexts.
- A non-exhaustive review of the usage diversity in computer science robotics that points out a growing interest in the concept as an active tool rather than a passive measure.
- The reuse method, a diversity-driven transfer exploration algorithm.
- An experimental setup with a real robot and an augmented reality environment.
- An example of environment-driven (reward-free, fitness-free) development of behaviour in chapter 8.
- An new way to bridge the reality gap, that is robust to many innacurracies in the simulation.

Prudens interrogatio quasi dimidium scientiae.
[Judicious questioning is virtually half of science]
Francis Bacon

One asks the questions that one knows how to ask.
Richard Lewontin

1 The Sensorimotor Exploration Problem

In this chapter, we motivate our interest for sensorimotor exploration in robotics. We briefly expose the classical machine learning paradigm and show that robots present specific learning challenges that prevent us to consider them as just another instance of machine learning. We formulate the exploration problem and articulate its difference with learning, and argue that exploration should be studied in its own right. To that end, diversity measures and an architectural framework are introduced.

1.1 Classical Machine Learning

Abstract · Classic machine learning is inherently passive, and is geared toward studying and predicting phenomena that happen outside of the control of the learning algorithm.

Machine learning concerns itself with constructing and studying systems that can learn from experience[1]. The circularity of this definition is adequately resolved by Mitchell (2006), who defines machine learning as:

> *A machine learns with respect to a particular task T, performance metric P, and type of experience E, if the system reliably improves its performance P at task T, following experience E.*

[1] Many authors use the term 'data' rather than 'experience'. While data, i.e., numerically encoded information, seems to act as good catch-all for what systems are experiencing, we would refrain from considering this point settled.

This definition is operational. 'Reliably improves' imposes a useful constraint to distinguish learning from stochasticity, while 'following experience' does not require to formally establish a causal relation between the gain in experience and the gain in performance, only to measure it empirically ('*reliably* improves').

Requiring an improvement in performance does lead to some problems though. The performance during learning might not be monotonically improving. As a result, the machine may find itself not having strictly speaking learned at some point of the learning process[2]. In this manuscript, I will use a slightly different definition that, while less practical, encompasses more instances of learning:

> *A machine learns with respect to a particular task T, performance metric P, and type of experience E, if the system reliably changes its performance P at task T as a result of experience E.*

Or, in a shorter fashion[3]:

> *A system learns if it changes its behaviour as a result of experience.*

The change in behaviour—i.e. the change in response of the system to a given context—does not need to be noticeable. There must only exists a potential situation in the future where the behaviour of the machine would be influenced by the experience it just acquired.

The experience, typically, comes from a *phenomenon* producing outputs from a set of inputs. If we are studying the weather the inputs are the past weathers conditions, and the outputs the current ones. In the case of a voice recognition task, the recorded sounds are the inputs and the text transcriptions the outputs.

input
PHENOMENON
output

Figure 1.1: A phenomenon receives inputs and produce outputs.

An input and its corresponding output is an *observation*. The idea behind machine learning it to gather enough observations—the *training data*—, and feed them to a learning system, which uses them to create and update an internal model. The learning system can then be used to predict outputs corresponding to new inputs that are not present the training data[4]. As such, the goal of machine learning is to construct systems that can *generalize* from data.

This translates in our previous examples, to historical weather data being used to validate models that can then predict tomorrow's weather. And with enough samples

[2] One could of course choose P *a posteriori*, so that P effectively improves. This is unsatisfactory though, as in practice, P is most of the time given and cannot be modified.

[3] The role of the task T, the performance metric P and the experience E in the previous definition is to narrow the specific aspect of learning studied. Each is optional. Without them, any organism possessing a nervous system with non-null synaptic plasticity is constantly learning (Hebb 1949).

[4] We are restricting the discussion to *supervised* learning here, that is, a situation where the desired outputs are given by the environment.

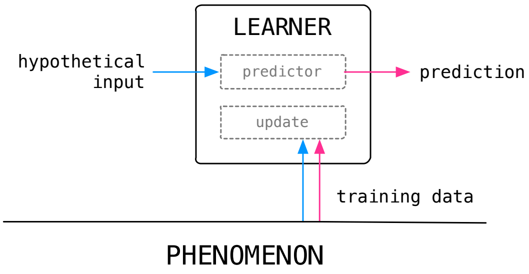


Figure 1.2: The canonical machine learning architecture.

of recorded vocalizations with their transcriptions, a computer program can build a voice recognition system that can recognize sentences it never encountered before.

The classic machine learning architecture can be summarized as depicted Figure 1.2. This architecture is general and has wide applications, ranging from detecting which edits are vandalism on Wikipedia (Adler et al. 2011), to helping medical diagnosis (Kononenko 2001), to classifying DNA sequences (Larranaga 2006), or detecting influences between famous artists of fine art painting (Saleh et al. 2014).

This approach thrives when data is abundant and matches the assumptions build into the models. In robotics, things are different.

1.2 The Trouble with Interaction

Abstract · We motivate why learning is important for robots, then describe the intrinsically interactive nature of robots, and discuss the unique challenges it entails.

Robots are synthetic systems that process information from sensors in order to self-produce motion. In other words, robots are synthetic systems that are *situated*—they perceive the world, and *embodied*—they can act in the world, and they are able to decide how to act contextually.

1.2.1 Learning Robots

Abstract · Learning in robots has several roles: increasing robustness to change, finding empirical solutions where theoretical models fall short, escaping task-specificity, granting self-sufficiency, allowing social exchanges, and providing insights into the learning mechanisms of biological systems.

Not all robots learn, i.e. not all robots change their behaviour with experience. Many robots used in industrial contexts repeatedly produce motions that have been programmed beforehand, and they do not improve them as they repeat them. For instance, picker robot on an assembly grab items on a treadmill and place them somewhere else[5]. Such a picker robot is able to *adapt* its behaviour to the position of the items that arrive on the treadmill, but it is not able to *learn*, as its performance will not change as it picks more and more items. This difference between adaptation and learning is a crucial one. Another example is a robot walking in a 2D maze. Assuming that the maze is simply-connected, that is, that all the walls are connected to the outer boundary, the robot, by following the left wall all the time—a method known as the *left-hand rule*—, will successfully be able to find the exit. Yet, even if the robot is presented repeatedly with the same maze, its performance at finding the exit will not improve. The robot successfully adapts without learning.

Adapting without learning has its limits; in many contexts, learning is a mandatory part of successful behaviour. We discuss some of the most important roles of learning in robots.

Robustness to Change

Robots that do not learn are inherently brittle to changes that are orthogonal to their adaptation capabilities[6]. For instance, the picker robot is perfectly able to handle changes in the position of items, but might not be able to handle those that involve changes in height, size, or weight, even if they are within the range the hardware would be able to handle.

If a robot does not learn, then the way it must behave for all the situations it can encounters must be pre-programmed into the robot. This makes programming robots an expensive task in complex environments. As a results, the tendency is to make robot environments simpler, and robot task-specific, so that programming the robots remains reasonably simple—this is typically the case in industrial settings.

Box 1.1: The *What* of a Robot

Robots are mechanical systems, but not all mechanical systems are robots: A bicycle can't self-produce motion, hence it is not a robot. A system that can self-produce motion is an *automatic* system. A clock self-produces motion, but it is not a robot as it has no sensors and does not process information.

But a washing machine that can measure and control how fast it is tumbling is a robot—robots don't have to be mobile or exhibit their motion in plain sight. A speaker connected to a microphone through a computer is technically a robot, as the membrane vibrates, but it is a degenerated case to which few robotics techniques apply. In contrast, a system producing and perceiving its vocalization through a motorized trachea is a robot, and such systems are active venues of research (Sasamoto et al. 2013).

Let's remark here that, were we to remove 'synthetic' from the definition, animals—humans—would be a subset of robots.

[5] See for instance this video for an illustration.

[6] By definition, all robots have some adaptation capabilities—even if they are just used to correct motion errors—because their sensory inputs informs their actions.

Robots that do not learn have the success of their behaviour tightly constrained by the set of assumptions that were made during their conception. If the environment evolves and violates any of those assumptions, the robot's behaviour may not be successful anymore.

Because in many real-world situations changes cannot be satisfactorily anticipated during the conception phase, robots that do not learn are restricted to controlled environments, where the range of situations they can encounter has been enumerated. Learning abilities allow robots to evolve in environments that were not in their designer's minds. To tackle more complex environments, robots need to learn.

Empirical Solutions

When sophisticated behaviour in a complex environment is required of the robot, preprogramming a successful behavioural strategy may be difficult. It requires to accurately modelize the phenomenon, and the robot, and then use sophisticated deductive reasoning to derive an appropriate plan for the robot actions.

A model of the robot or the environment may be difficult for any number of reasons. The poor build quality of the robot may induce unpredictable hardware variations that make any modelization inaccurate. Wear and tear may make any model quickly imprecise[7]. The robot structural architecture may be too complex: robots equipped with pneumatic muscle or soft limbs are notoriously difficult to simulate to any useful precision (Daerden et al. 2002; Trivedi et al. 2008). The minute physical details of the surface of the robotic hand and object to be manipulated (friction coefficient, surface deformation, compliance) may be difficult to measure, and their impact difficult to anticipate even with state-of-the-art physics theory. Preprogramming walking into human-size humanoids has proven challenging, requiring sophisticated techniques using detailed models (Hirai et al. 1998; Kaneko et al. 2009). And even then, the behaviour is limited to stringent assumptions:

> *The pattern generator, which is constructed using 3D linear inverted pendulum model and the preview controller, provides a stable motion including walking in real-time, on the assumption that the floor condition of the real environment is ideal and is the same as that of simulation model.*
>
> Kaneko et al. (2009, p. 12)

In those situations, pre-computing a good and robust controller for the robot is expensive or impossible.

Preprogramming a complex behaviour into a robot is akin to ask a child to look at a bike, and to think really hard to figure out how the bike works, and how he should position himself on it, and how he should coordinate its legs precisely to push the pedal while balancing himself to go forward without falling. Preprogramming the

[7] Which is not to say that wear cannot be handled: many methods have been developed to detect damage and wear in industrial settings, see Chandola et al. (2009, p. 17).

strategy into the robot is similar to hoping that after this period of intense reflection, the child will be able to walk up to the bike, and ride it successfully on the first try.

Clearly, the child does not need to understand how the bike works, nor does he needs a working knowledge of Newtonian physics to ride a bike. And even if it did, that is, if you took a mechanical engineer that never learned how to ride a bike, gave him as much observational time he desired, he would still not be able to proceed this way.

The child does not even need to keep good representation of the behaviour once a successful solution is found. An interesting study in this regards looked at the explicit knowledge of skilled typists of the computer keyboard (Snyder et al. 2013). They gave the typists 80 seconds to fill the 26 letters of a blank printout of the keyboard (all the other keys being represented). The subjects identified 57% of the keyboard correctly, got 22% wrong, and were unable to remember the rest. Successful behaviour does not need a good representation, nor a good understanding of reality.

Instead, for robots, a simple empirical trial-and-error strategy until the success is detected may suffice: using the real-world is a computationally frugal—and remarkably precise—way to simulate the real-world. And it bypasses the necessity for expensive and practically inaccurate models. The robot does not necessarily need to understand exactly why a motor command produces a specific behaviour to acquire successful behaviour: learning abilities allow to figure out complex behaviour without using sophisticated deductive reasoning.

Escaping Task-Specificity

Another role of learning that has emerged recently as an acknowledged research ambition, is for robots to escape *task-specificity*. In an industrial context, it makes sense to specialize each robot to its task. It increases productivity and efficiency, and is a cornerstone of assembly-line design.

Out of controlled environments, robots may be able to accomplish more than one task. Honda's ASIMO robot is able to detect movement, faces, sound, to recognize when a handshake is offered, to walk, to run, to detect obstacles, to shoot into a football. And many other abilities. But ASIMO is not able to acquire new behaviour. It cannot learn how to ride a bike. It is *specific to the tasks hpe was given*.

To be useful in social contexts or evolving environments where new tasks are created contextually, robots must be able to acquire new behaviour and master new skills on their own so that they can remain useful.

Self-sufficiency

Those three objectives of robotic learning all participate to a fourth and broader one: granting robots self-sufficiency. Self-sufficiency is the ability to carry out one's objective without the intervention of an another entity not intrinsically necessary for the task[8].

Many learning systems are not self-sufficient. For instance, predicting the weather is a demanding endeavour. Historical weather data must be aggregated and encoded; the current conditions must be continuously updated. Simulation of the weather across the globe must be run on supercomputers. Because faster simulations afford better resolutions, which in turn affords better accuracy, the programs are continuously optimized, and the hardware regularly updated. The simulations themselves exploit state-of-the-art science, and teams of meteorologists and engineers are constantly improving the prediction models. If the prediction is unusually bad in a particular instance, the roots of the problem are investigated and possibly addressed. Unless you include the humans running the weather predicting system into the weather predicting system itself, it cannot be considered as self-sufficient.

In the case of robots, we can distinguish two types of self-sufficiency: from experts, and from others. *Self-sufficiency from experts* is hindered each time an open-skull intervention is performed, that is, any intervention where an expert is needed to modify, repair or upgrade the robot software or hardware.

This is related to the concept of *operational closure* that separates a living entity from the rest of the world, in the context of *autopoiesis* (Maturana and Varela 1973). A robot is self-sufficient from experts if its hardware or software envelope—its operational closure—does not have to be breached. The weather prediction system's envelope is breached every time an upgrade is made to the system. Essentially, self-sufficiency from experts precludes any *redesign* of the robot once it has started functioning.

Self-sufficiency from experts in an important precondition in order to deploy robots outside industrial settings. In that context, learning capabilities grants self-sufficient robots some measure of self-redesign, since redesign cannot come from an outside entity anymore.

Self-sufficiency from peers relates to the robot requesting or needing help from a human or another entity who has no direct access to the robot's mind. A robot that requests demonstrations is not self-sufficient. A vacuum robot is self-sufficient from experts, but its self-sufficiency from peers is as great as its capacity to not get trapped in a corner, and as how rarely it needs to be emptied.

Self-sufficiency is a continuum. A robot is self-sufficient from experts insofar as it never needs repairing. As a consequence, avoiding dramatic damage is part of self-sufficiency. But damage cannot always be avoided: robots are never perfectly self-sufficient. Some robots are more adept at avoiding damage than others: they are more self-sufficient. Likewise, better learning capabilities may reduce the number of human demonstrations needed before a skill is acquired, thus improving the robot self-sufficiency from others, without achieving it completely.

A helpful parallel can be made with humans: humans are self-sufficient from ex-

[8] The task determines if another entity is needed, not the abilities of the agent. Greeting people in a museum intrinsically needs those people to achieve the task. Moving a reasonably weighted box does not need an outside entity.

Box 1.2: Self-sufficiency versus Autonomy

Self-sufficiency means that one can carry its mission while only relying on its own. In contrast, the meaning of autonomy, as Smithers (1997) (see also Paolo et al. (2008)) argues, has been diluted in the robotics community and is rarely employed to make 'clear and useful distinctions'; the term has been employed in senses as diverse as mobile, self-sufficient, intelligent, or able to make decisions.

Autonomy is rarely employed in robotics in the sense in which it is overwhelmingly used in philosophy, medicine, politics, biology, or law: an autonomous entity is an entity capable of self-determination and able to make its own laws or principles. In this sense, autonomy is a property that is internal to the system, and bears little relation with competence or capacity. Self-sufficiency, in contrast, is defined entirely in the relation the system can entertain with its environment. By way of illustration, a territory can be declared autonomous—able to make its own laws—without possessing the economical capability to not rely on foreign aid (be self-sufficient), or the political competence to govern itself efficiently.

In this thesis, we refrain from using the words *autonomy* and *autonomous* altogether.

perts insofar as they do not need medical care. And humans' self-sufficiency from peers is low in infancy (that does not mean that they are not *autonomous*—see Box 1.2), and increases throughout development.

Social Interactions

Most social interactions amongst humans assume that learning takes place. Even if only to remember someone face or name, learning must happen. Previously discussed information and expressed preferences are expected to be remembered, if only approximatively. And when didactic exchanges happens, individuals are expected to be able to acquire simple skills. The ability to learn is therefore crucial if robots are to have normal social interactions with humans.

Insights into Biological Learning

Finally, some of the research studying learning in robots is motivated by gaining insights into how biological systems learn. Robotic instantiations of mechanisms and structures found in biology provide important scientific tools to study them in repeatable, reproducible, and bias-controlled settings, that are difficult or impossible to achieve with biological systems (Webb 2000, 2001, 2002; Ijspeert, Crespi et al. 2005, pp. 190–193; Ijspeert 2008, pp. 647–648). One crucial advantage of robots for learning research is that they can be tested with learning deactivated, and they can be made to forget at will. We'll go back to this in section 1.3.

In other instances, the learning situations require specific conditions. For example, Blumberg et al. (2013) proposes a robot model to study the functional value of muscle atonia (lack of tone) for sensorimotor learning during sleep; such a study would be cumbersome in animals.

In many ways, the communication between biology, psychology and neuroscience on one side, and robotics on the other is still in its infancy. But because robotics offers unique and operational experimental opportunities, the dialogue is poised to strengthen as research advances.

This a pivotal ambition of the developmental robotics (Weng 2001; Lungarella et al. 2003; Asada et al. 2009), the evolutionary robotics (Nolfi 2000; Lipson 2005; Doncieux, Mouret et al. 2011; Bongard and Lipson 2014) and the biomimetic (Beer et al. 1998; Vincent et al. 2006; Ijspeert, Crespi et al. 2005; Ijspeert 2008; Pfeifer, Lungarella et al. 2007) research fields, which we'll talk more about in sections 1.2.4 and 1.2.5.

So far we have discussed the role that a learning ability has for robots, essentially answering the question 'why should we want robot to learn?'. We have distinguished six different roles: robustness to change, finding empirical solution to complex problems, tackling new tasks, self-sufficiency, engaging in normal social interactions, and providing insights into biology and neurosciences. These roles are often overlapping, and this list should not be considered as exhaustive. In the next section, we tackle the implications that being a robot has on learning.

1.2.2 Acting, Learning Robots

Abstract · Robots are agents, not pure learners from the classical machine learning paradigm. They are immersed in the real-world, and can exert control on their data sources. As a consequence, they face specific challenges and opportunities when learning.

Robots learn from experience collected through interaction with their environment. The environment fits nicely as a phenomenon: inputs are *motors commands*—the actions the robot executes—, and outputs are *sensory stimulations*, i.e. what the robot perceives through its sensors. Yet, the necessity of interacting with the real world to collect data has a profound impact on many aspects of learning that precludes us from considering a robot as just another instance of a classical machine learning problem.

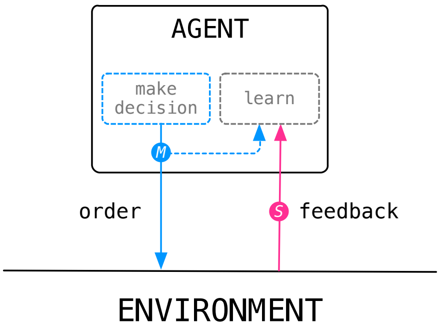


Figure 1.3: An agent makes decisions that affect the environment.

The most immediate consequence of the inherently interacting nature of robotic learning is that robots are never only passive learners, they are actors—*agents*—, and they must *act to learn*. Because of this, predicting the environment is not the only point of learning anymore: *controlling the environment* is a desirable goal too.

There are two main reasons why the robot has to act to learn. First, because while the real-world may be full of diverse events unfolding all the time, the current environment of the robot may not exhibit information pertinent for the robot's motivations. Acting introduces directed variability towards specific elements of the environment. And second, because acting allows the robot to learn the consequence of its own actions on the environment.

This active role in the environment introduces a new issue that is not present in the classical machine paradigm: if a robot must act to learn, then what should it do? The production of a motion with a motor is not a neutral act: it supposes that a decision was made at some point. A decision reduces all potential actions the robot could produce into just one, acted out. In comparison, the weather system never makes a decision, never acts; it receives the data and only *predicts* to the best of its ability.

Interacting with the real-world creates many challenges: observations are expensive to acquire while action possibilities far exceed the interaction opportunities. No reasonable assumption can be made about the homogeneity of the space with regards to stochasticity, redundancy or noise, and observable phenomena are not always learnable or controllable, which leads to high variability in the value of an observation. Moreover, many phenomena are not observable, and the interaction possibilities offered by the environment are unknown. We detail each of these points below.

Interacting Is Expensive

Since a robot is interacting with the physical world, each interaction requires time and energy. Time, in particular, intrinsically limits the number of interactions the robot

Box 1.3: Forward and Inverse Models

Typically, predicting the environment is done using an *forward model*: given some hypothetical action and an environmental context, the forward model makes a prediction about the consequence of the action on the sensory receptors of the robot. In order to produce a specific desired effect in the environment, an *inverse model* is used. Given a specific desired effect, i.e., a goal, the inverse model infers actions to produce it. In a learning robot, models are dynamically learned as the robot interacts with the environment.

In industrial settings, forwards and inverse models rely on a precise representation of the robot, from which kinematic and dynamic models are derived. But this is a specific case.

Forward and inverse model do not imply a representation of the world: they encode the relationship between actions and effects. In our bike example, a forward model might establish a relation between the pedalling rate and the speed of the bike, without the need to represent the mechanical processes involved in that relation. And the corresponding inverse model will indicate that to go faster, the pedalling rate must increase.

Forwards and inverse models are the product of an engineering approach, and finds their origins in control theory. Their presence in biological systems, and the form this presence takes is still debated (Ito 1972; Partridge 1982; Miall et al. 1996; Wolpert et al. 1998; Kawato 1999; Johnson 2000; Loeb 2012; Oztop et al. 2013).

Box 1.4: The Separation Between an Agent and its Environment

It should be stressed that the separation between the agent and its environment is not the same as the robot/rest-of-the-world one, or the hardware/software one (Bertschinger et al. 2008). The agent usually consists of a subset of the software of the robot; everything else constitutes the environment, as Figure 1.4 illustrates, including any software component that helps process and carry the motor signals to the effectors, or processes and transmits the sensory signals from the sensors. In particular, the environment includes the body of the robot, which, for our purpose, is only qualitatively different from the rest of the environment insofar as it provides the interface to it.

Engaging with an analogue world, impactful choices must be made to decide at which level of abstraction to interface the agent. It can be high-level, giving abstract orders and receiving semantic feedback ('move toward the door'/'door is open'), or low-level, sending torque commands s and receiving raw sensors data such as pixel matrices from a camera every 10 milliseconds.

In this manuscript, we will generally choose approaches where motor commands are low-level while sensory feedback is high-level (this is not innocent).

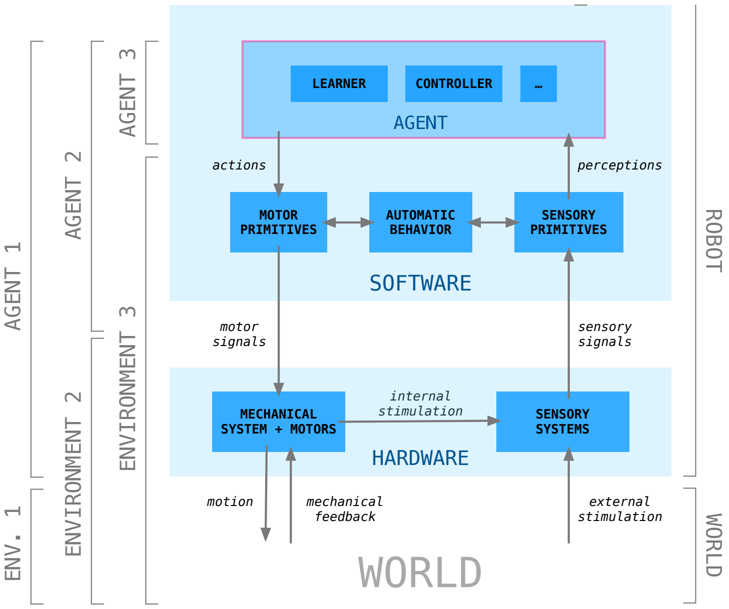


Figure 1.4: From a computational, open-skull perspective, the separation between an agent and its environment is not the same as the separation between the robot and the world. But for an external observer situated in the environment, they are the same, as the agent is identified by its operational closure (Maturana and Varela 1973). Note that the automatic behaviour component might include any number of subsystems, including other agents. The agent is discriminated from the rest of the software as an arbitrary perspective. In particular the agent is not necessarily omniscient or omnipotent over the software. See Pfeifer, Lungarella et al. (2007) for a similar diagram for biological systems.

is able to undertake. This holds regardless of the span of time considered: minutes, hours, days or lifetimes. In the same way the experience a single human has of the world is limited, so shall be the experience of any agent situated in the world. Time is a fundamental limitation because it won't go away with more computing power or cleverer algorithms, which are the two usual ways computer science tackles problems.

Besides time, interaction opportunities may be limited by other resources. For instance, in the case of a robot learning to engage socially with humans, individuals willing to interact with the robot may be few in a given context, and the robot might for instance be able to achieve only a few dozen of interactions per day[9].

Interaction Possibilities Are Many

The time cost of an interaction is made acute by the magnitude of the number of actions that are possible. A simple robotic arm with seven motors, each capable of 100 different positions (several orders of magnitude less than most servomotors can achieve) can exhibit 100^7 different arm postures. At a rate of one posture per second, that translate into 6 million years.

In any moderately complex situations, the number of distinctive interactions with the environment far exceeds the available time resources. It precludes any exhaustive exploration of the motor space, or any coverage of it to a useful resolution.

If the mapping between the robot's action and their consequences is simple enough, the robot may be able to generalize from a few interactions, i.e., a good forward model may be derived (Box 1.3). In most cases though, the mapping is too complex, unobservable or cannot be learned due to the limitations of the robot learning abilities. In that case, uncertainty over the consequence of the majority of the robot's possible motor commands is unavoidable.

Heterogeneity of Stochasticity, Noise and Redundancy

Many machine learning algorithms assume that noise and stochasticity is homogeneously distributed over the learning space (Loeb 2012; Oudeyer, Baranes et al. 2013). With robots, the noise and stochasticity is usually heterogeneously distributed. For instance, a robot bumping violently into a wall will experience high sensory stochasticity, while turning before the wall will elicit predictable environmental feedback.

Likewise, the amount of noise a camera is experiencing depends on the luminosity, and the same goes for the human vision system; in an unevenly lit room, the changes in sensory noise can be sharp. Motor noise might also be a concern. The noise around the position of a joint is usually not negligible, and impacts motion especially if the joint is proximal. Furthermore, over the range of possible values the joint can take, the noise might not be uniform. This typically happens in the neighbourhood of the extremities of the working range, or when part of the motor—for instance a specific

[9] A robot engaging in five-minute interactions with ever-willing humans, 24 hours a day, can only hope to collect 24×60/5 = 288 interaction experiences per day.

cog in the gearbox—is damaged. The noise can also be dependent on the forces being applied to the joint. Which means that the joint's level of expected noise dynamically varies with the posture of the rest of the arm and during physical interactions with the world.

Heterogeneity is also present in the redundancy of the mapping between actions and sensory feedback, as illustrated in chapter 0. In a typical robotic setup, many actions generate the same effect, while some effects are only produced by a handful of actions. This makes random action sampling ineffective.

Observable Is Not Predictable Is Not Controllable Is Not Learnable

A key assumption embedded in learning algorithms is that the entire domain considered can be learned (Oudeyer, Baranes et al. 2013).

An observable effect might not be predictable or controllable. For instance, a dice roll might be observed, but cannot be predicted or controlled (under reasonable assumptions). Likewise, I might correctly predict the trajectory of a cloud in the sky, but I have no control over it: predictable is not controllable.

And, a phenomenon might be potentially predictable and controllable by the robot, but not practically learnable: the phenomenon's complexity might outclass the learning abilities of the robot. Or, there might not be enough time to learn it[10]. Or, the phenomenon might require previous knowledge or a skill that the robot can acquire, but did not, yet. For instance, the ability to reach is required to be able to grasp. And running is easier to learn after being able to walk.

All is Not Observable

Let's get the uncertainty principle (Heisenberg 1927) out of the way: even in the best of circumstances, the knowledge one can have of a situation is inherently limited. But that makes no practical difference in most practical robotic settings.

Humans and robots alike are limited in their knowledge of the world in much more important ways. First, they are situated, i.e. they occupy a specific place in the environment, and they get their information about the environment from this perspective only. They cannot see behind an opaque object, they cannot hear sounds from behind a soundproof glass[11], they cannot feel an object they are not touching.

Second, there are many unobservable phenomena going on in the environment at any given time. Deductive abilities might be able to estimate unobservable information from observable cues, but in many instances, this is not reliable. For instance, the state of mind of someone else is not directly observable, and their knowledge and skills is not either.

Hence, any representation that a situated agent can form of the world is limited.

[10] One might for instance consider a phenomenon corresponding to a linear system of rank n+1 with only n samples allowed. Or, more simply, to enumerate all possible outcomes of a 6-dice roll with only 5 samples allowed.

[11] robots might able to *see* the sound though; Davis et al. (2014)

But it gets worse: the set of possible future states of the world is not known either.

Possibilities Are Unknown

Having a clearly defined set of all possible in a given situation outcomes is possible—if the robot is placed in a known, simple, controlled environment, and a description of the set of possibilities is spoon-fed by an engineer.

When the environment grows more elaborate, the task of the engineer becomes more difficult, and the description of possibilities more complicated. Physical phenomena at play are more difficult to grasp, initial conditions are only partially known, the number of interacting entities grows, and the number of interactions grows quadratically. Quickly, describing all possible outcomes is impossible, even with complete knowledge of the situation. In other words, the frontier between what is possible, even if unlikely, and what is definitely not possible gets blurry.

A fortiori, it is impossible for a self-sufficient robot only able to gain partial knowledge of the situation through situated sensory acquisition, and whose grasp of theoretical physics is arbitrarily bad, to derive all possible outcomes of a given situation. Therefore, the robot must act in a world where what is possible to observe is uncertain.

All Observations Are Not Equal

Interactions possibilities are many, but each is expensive, making an exhaustive approach unfeasible. And each observation does not yield the same information gains. Therefore, in a robotic context, the value of an interaction for learning a task varies dramatically. This leads to a potentially high opportunity cost for every interaction the robot chooses to undertake: each interaction yielding poor observations decreases the total amount of information the robot can hope to gather over its limited interaction budget.

Therefore, a good exploration strategy, that efficiently select actions to maximize the information they bring is necessary. It makes the exploration strategy employed to choose *what* to do as important as the performance of the learning abilities the robot is provided with. And the strategy must match the abilities of the agent: actions should yield learnable observations.

And the ineffectiveness of random sampling, the robot's limited knowledge of the situation, and the uncertainty on what the environment offers in terms of possible interactions make finding a good exploration strategy non-trivial.

1.2.3 Learning Before Acting?

So far, we have argued that robots must interact with the world in order to learn. Because it bypassed the need to use complex representations, that were difficult or impossible to acquire in the first place. Historically however, robotics has seen the development of many approaches were learning happens before acting is performed. In this section, we explore the advantages and pitfalls of such approaches.

A seductive proposition is to equip a robot with enough high definition sensors so that it can capture and build an accurate representations of its environment, in which it can then simulate its behaviour and the one of the environment by drawing on huge prepared databases of information about all conceivable objects and entities (Suh et al. 2007; Lemaignan et al. 2010; Tenorth et al. 2013), learn from this disembodied experience, and then act by exploiting the knowledge gained, without ever experimenting haphazardly in the real world. The robot would only be executing well-laid plans.

Granted, it seems that, for a child or a mechanical engineer, thinking about riding a bike is not the best way to learn how to actually do it. But neither of them has the ability to run complex physical simulations involving hundreds of moving parts in its head. And besides, interacting with the bike in the real world is slow. The robot, on the other hand, is able to run thousands of simulated tries per second. After all, we build planes and cars from simulations: simulating a bike, then, is easy. And even if the simulations are not perfect, their sheer amount should counterbalance the inaccuracies enough to be able to derive a successful behaviour.

This approach *is* seductive, if only because it promises that most problems can be solved by throwing enough computing resources at it: if the representation is accurate enough and has predictive capabilities, simulations can be run, and simulation's time is only dependent on the processing power available. Future technological progress will take care of any current lack of processing power. And even if simulators are still limited, future advances will make them more and more accurate: at some point, this two-pronged approach will be enough for most practical situations.

This approach extracts the problem from a context—the real-world—where it is time- and energy-consuming to solve into another, the simulated world, where there is no irreversible consequences for one's action, where the action costs are comparatively low, and where all the techniques of computer science can be brought to bear. In other words, as soon as the problem has been adequately transferred into the simulated world, all the difficulties of the real world we exposed in the previous section are discarded, and finding a solution becomes much easier.

Optimal Control

A reason for this approach is historical. Industrial robotics have developed mature, powerful theories that allow to compute *optimal* controllers. In an industrial setting it is highly useful. Optimizing repeated movements reduces costs. Given the availability of such a powerful toolkit, developed and field-tested over several decades, it seems natural to want to apply it onto new robots.

One of the most important motivation behind the research for optimal control algorithms is that they reduce the design of robots to their hardware: engineers are motivated to build the best possible robot bodies, with the guarantee that optimal control algorithms will exploit them to their fullest. The hardware and software problem are decoupled, with the later only needing a precise description of the hardware to automatically adapt to it. In theory, such an approach would allow unrestrained originality in designing the hardware. In practice, the opposite happens. Optimal control algorithms are only applicable to a narrow range of hardware (typically, rigid body with electrical servomotors), and as such reduce the choices that guarantee that the algorithms will work.

Indeed, those methods require accurate models of the robots. But as robots become more and more underactuated, compliant, flexible, even soft, and engage complex and uncertain environments, such methods impose a set of assumptions that is increasingly at odds with the ecological context of the robots. Optimal control generally requires a known, observable, computable cost function, as well as a precise and computable inverse model of the robot, and a low and homogeneous level of noise. The constraints on the objective function limits the range of tasks that can be handled, the inverse models are difficult-to-impossible to establish and are computationally expensive, and the assumption on the noise, as we have discussed, is unreasonable.

The optimality approach has been criticized by Loeb (2012): such assumptions are most often not found in biological organisms[12], who empirically derive *good-enough* behaviours instead by trying random motor activations on a high dimensional motor space where the density of useful solution is high (Raphael et al. 2010). By trying different random values, and creating perturbations of the most promising ones to approach the nearest local minima, animals are able to quickly evolve and acquire good behaviours. Proceeding this way has the advantage of providing the organism with a repertoire of useful, *diverse*, solutions that confer robustness to the organism.

Simon (1969, pp. 28, 119) similarly advocated for settling for such *satisficing* solutions:

> *We must trade off satisficing in a nearly-realistic model against optimizing in a greatly simplified model.* (p. 28)
> *In the real world we usually do not have a choice between satisfactory or more solutions, for we only rarely have a method of finding the optimum.* (p. 119)

There are current attempts to explain motor coordination as optimal control, in particular in the context of optimal feedback control (Todorov and Jordan 2002). And although the principle of evolution as an optimization process in often invoked to justify the biological plausibility of optimal control theory, Loeb observes that it is in fact at odds with evolution:

> *it is highly unlikely that a given mutation in musculoskeletal form of an individual will be accompanied by a simultaneous mutation in the control system*

[12] A small experiment illustrating how incomplete are our own conscious models of our limbs: close your hand into a fist without tightening your muscles, and then bend your wrist inwards as much as possible. What happens? If you are not flexible enough, your hand opens. You possibly did not anticipate this, illustrating blindspots in the explicit knowledge of the forward model of your hand, something you use all the time, every day, with remarkable efficiency and dexterity.

> *that might be required to make optimal use of the new body form*
>
> Loeb (2012, p. 763)

Settling for good-enough solutions not only creates robustness from mutation, but it also confers individual robustness to environmental changes because it brings a diversity of strategies that are not optimal to any specific environment, but rather efficient in many. It also enhances population robustness, because it creates behavioural diversity independently of genetic diversity.

In robotics, even when the theoretical assumptions are met for optimal control, practical considerations come into play: fast gait in legged robots requires high-bandwidth, low-latency sensory feedback for optimal control. Such a strategy is both taxing in energy and computational resources (for instance, a low-latency might require to simulate the immediate consequences of actions before the actual consequences can be perceived, using an internal forward model). Insects have evolved alternative control strategies that do away with centralized control—the legs communicate and synchronize through environmental mechasensory feedback—and allows them to react faster to obstacles on the ground than the speed of their neural pathways would allow (Cruse 1990; Espenschied et al. 1993). The performance of insects remains currently unmatched by robots.

So far, we have characterized the search of optimality as ecologically divorced from self-sufficient robots. But that does not disqualify the use of complex representations yet.

Benefits of a Full-Representation Approach

Using a full-representation approach, or a representation-based approach does not mean that one has to optimize in it (we will see how that might create a problem in chapter 9). Acknowledging the limitations in time, resources and realism, finding *operational solutions* is a possible approach.

The representation-based approaches have been extensively used, and, still are today. For instance, self-driving cars must be acutely aware of their surroundings. And they cannot experiment on the road for the sake of learning. A self-driving car should know how to drive safely from day one, and probably having an explicit, provable behaviour. Self-driving cars use Simultaneous Localization and Mapping algorithms (SLAM) (Thrun 2005, pp. 309-485) and array of sensors, to maintain constant, omnidirectional representation of their surroundings and their location in that representation. They must be aware of other cars, pedestrians, and correctly identify all the road signs and markings. A self-driving car is a robot. Why then, can't such an approach be used for any other robots?

First, it is important to remark that self-driving cars still face technical challenges, and have not be field-tested in large numbers, but non-engineers. While most of the technology may work, some of the remaining problems may prove very difficult to solve, and require radically different approaches. But let's assume that self-driving cars work.

The reasons why a full representation approach is feasible for self-driving cars are many: the dynamics of a car with a road can be subtle at times, but they are rather well-understood and can be computed to a useful degree of precision. A car has no limbs or articulated spine, nor does it grasp or manipulate anything; the set of different entities a car typically interacts with can be enumerated (vehicules, pedestrians, obstacle, etc.); most relevant information communicated through vision; it can sport heavy sensory equipment and carry large power resources. A car has no problem balancing itself at rest. Cars evolve in a road network that does not change rapidly. This allows to actually map the complete environment before the robot is allowed to roam in it, and that is what is done with self-driving cars: they rely on precise pre-captured maps of all the roads they evolve in, and merge those with their sensory data.

With humanoid robots, none of those assumptions can be made. For humanoids, having a perfect, up-to-date representation of themselves and their immediate surroundings for simulating behaviour is unrealistic.

This leads us to conclude that, while a full-representation approach is tempting because it can easily be *exploited*, it is impractical, expensive, most of the time unfeasible, always limited, and often unreliable.

Embodiment

One of the most impotant argument against representations was made by Brooks (Brooks 1991c,a), provocatively:

> *Representation has been the central issue in artificial intelligence work over the last 15 years only because it has provided an interface between otherwise isolated modules and conference papers.*
>
> Brooks (1991b, p. 1)

In rupture with his contemporaries in the artificial intelligence community, Brooks advocated an approach to constructing entities capable of intelligent behaviour without giving them symbolic manipulations abilities or representations, but by rather letting behaviour emerge from the interaction of the entity with the world (Brooks 1990, 1991b,a), using cognitive architectures where there was no centralized control centre to be found.

The main point advanced by Brooks is the one of *embodiment* (Brooks 1991c; Varela 1991; Hutchins 1995; Hendriks-Jansen 1996; Ballard et al. 1997; Clark 1997; Arkin 1998; Lakoff 1999; Pfeifer and Scheier 1999; Beer 2003), that postulates that intelligent behaviour can only emerge from a rich-enough interaction between the brain, the body, and the environment:

> *It is hard to draw the line at what is intelligence, and what is environmental interaction. In a sense it does not really matter which is which, as all intelligent*

> *systems must be situated in some world or other if they are to be useful entities.*
>
> Brooks (1999, p. 169)

As such, an intelligent agent must be firmly embodied in the real world[13], and should not rely on methods that allow him to escape towards simulated representation, for such an approach is inherently fated to fail. In the context of embodiment, the body in an inseparable component of cognition, not just merely a vehicle for it: it generates sensorimotor couplings that create specific regularities in the sensorimotor flow, i.e. the body structures our relation with the world.

More specifically, robots' bodies play the same role as our own: they offer an interface to the world. This interface is all but neutral: it is specifically situated, i.e. the location of our bodies gives us a specific point of view, it is heavily mediated (Taylor 1995), and it is embodied. This mediation influences how we perceive and think about the world (Pfeifer and Bongard 2006). In particular, we constantly think about the world in relation with the capacities of our bodies (Gibson 1977): I can sit on this chair, this table is too heavy for me to move it by myself, this door can be pushed. We do not perceive the world as it is, but how it relates to us.

> *The scale of human observation and experience lies within the narrow bounds of inches, feet or miles, all measured in terms drawn from our own selves or our own doings. Scales which include light-years, parsecs, Angström units, or atomic and sub-atomic magnitudes, belong to other orders of things and other principles of cognition.*
>
> Thompson (1917, p. 24)

Because our knowledge of the interactions the world offers changes with learning, our immediate perception of the world is dependent on the knowledge and skill we possess.

The existence of mirror neurons, that fire both when one performs an action or observe someone else perform it (Pellegrino et al. 1992; Gallese et al. 1996; Oztop et al. 2013), shows how our morphology impacts our perception: the body is computing the recognition of other's actions (Umiltà et al. 2001; Rizzolatti et al. 2001), removing the need to rely on abstract, deductive and representational cognitive processes.

Embodiment implies that it is equally impossible to comprehend the embodiment induced by a different morphology without experiencing it directly. This is because our body gives us access to our own embodiment experience directly and constantly, it is never just a cognitive process. By way of example, to understand (some limited form of) the embodiment of current humanoid robots, which often have no flexible ankle and no articulated spine, one has to strap ski boots, and a medieval armour, and experience the limitation for himself. Before doing so, anticipating accurately

[13] Embodiment postulates that intelligent behaviour can only emerge from interaction with a rich-enough environment. While environments different than the real world are not ruled out by the argument, proponents of embodiment generally argue that no such other rich-enough environment exists today, in particular not in simulation.

how they would affect everyday movements is surprisingly difficult. And, by the way, letting the world do the work of simulating the constraints with actual ski boots and armour is computationally and cognitively frugal, and much more accurate.

This call for an ecological approach to designing robots, where the morphology and cognitive functions are designed together, so that one can efficiently exploit the other. Smithers (1994) illustrates this point by observing that better sensors do not necessarily make the control problem easier: considering a wheeled robot equipped with IR sensors doing laps, he observes that, as the resolution of the sensors is increased, they are more sensitive to small variations—not noise—produced by the slight differences in the laps. This makes the control problem more difficult, because the control algorithms now have to discriminate between the informative and non-informative parts of the data, with respect to the control purposes. The solution here is not to make the movements of the robots more precise: the previous behaviour was already adequate, and that would make the control algorithm even more sensitive to occasional accidental variations. Rather, the sensory abilities of the robot should be designed to be compatible with its cognitive and motor abilities, with respect to the precision required in the behaviour.

A recent advance on the notion of embodiment is brought by Pfeifer and Iida (2005), who introduces the concept of *morphological computation* (Paul 2006). Under this principle, part of the computation necessary to accomplish a task can be done implicitly by the morphology of the agent, reducing the amount of explicit control, i.e. cognitive resources, that must be dispensed. The principle has been exemplified in biped locomotion (Paul 2004). *Passive walkers* are purely mechanical bipedal structures, that, when placed on an inclined plane, transform potential energy into kinetic energy, and stay balanced: they are able to walk. They demonstrate that a behaviour thought to require complex algorithms and fast communication pathways can be produced without computational control (McGeer 1990, 1992; Wisse et al. 2007). From there, reintroducing control on the passive walker can be done by acting on few parameters (Vaughan et al. 2004). This is a direct example of self-organization: the passive walker is creating an attractor of stable bipedal locomotion, and well-placed actuators, rather than modify explicitly the movements, nudge the walker toward a slightly different attractor, where walking is faster for instance.

Morphological computation is everywhere in the musculo-skeletal system. For instance, the soft envelope of our fingertips deforms to simplify grasping. In robotic manipulation, a spectacular example of this, pushed to the extreme, is vacuum grippers (Brown, Rodenberg et al. 2010). Instead of dealing with the complexity of an articulated hand, a gripper that passively adapts its shape to the manipulated object offloads much of the computational cost of grasping to a physical phenomenon. Even when articulated hands are required, equipping them with soft fingertips facilitates greatly manipulation. Morphological computation is also present in all our joints, whose compliance act as dampeners the small variations of the mechanical feedback of the world. Small perturbations are thus handled by the morphology, which acts as a filter that lets through consequential perturbations that warrant a change in motor activation.

This dampening also directly simplifies control by reducing the chaotic behaviour of the environmental response: small changes in motor commands will produce small changes in the produced effects. An example of this is the role of feathers in flapped flying:

> *in complete rigid wings, a small difference in stroke force between two wings will result in a drastic change of net aerodynamic force and moment on the body. Conversely, in a flexible wing, the change will be small due to the passive bending of each feather.*
>
> Shim et al. (2007, p. 757)

A step up from passive morphological computation, *mechasensory reflexes* are typically outside of the agent's control, but play a large part in organizing sensorimotor stimulation in a way that favours learning and control. In birds, flapping strokes during flying induce body oscillations on the order of ten times per seconds (Warrick 2002). Yet, the head remains largely isolated from these oscillations by the actions of optokinetic and vestibular reflexes (Maurice et al. 2004; Dickman et al. 2000). These reflexes are essential to ensure that the vision system and the maculae of the inner ear, which perceives accelerations, are able to function properly.

This example of (reflexive) action supporting perception illustrates well how indissociable one is from the other in embodied agents. This has been illustrated first by Ballard (1991), who introduced the idea that an active vision system, that is, a vision system which could move in the world to examine an object of interest under different perspectives, make vision much simpler from an algorithmic perspective.

In a take similar to Brooks, proponent of active (or *interactive*) vision reject the necessity of an explicit representation of the world:

> *An hypothesis of interactive vision claims that the brain probably does not create and maintain a visual world representation that corresponds detail-by-detail to the visual world itself. For one thing, it need not, since the world itself is highly stable and conveniently 'out there' to be sampled and resampled*
>
> Churchland et al. (1994, p. 36)

This idea, again, is that abandoning complex representation makes the cognitive and computational problems simpler. Humans keep a overall idea of the visible world, and resample the world as the need arise (O'Regan 1992). That ability of the vision/action systems to 'avoid' as much as possible higher cognition has been illustrated recently by a study done by Perfiliev et al. (2010), where objects flying at high velocity are launched at subjects from the side. In all instances, the subjects choose to (try to) grasp the object with the hand closest from the incoming direction (if the object came from the left, the left arm would be used). The truly remarkable observation is that this arm selection happened at a latency too low to allow for a voluntary decision to be made or planning to happen. Perfiliev et al. (2010) proposes that an innate neuronal mechanism can guide reaching of the arm towards a specific goal. The experiment was reproduced for humans, monkeys (Rhesus and Japanese macaques) and cats.

In conclusion, the same features that make these systems easier to control and to learn from make them harder to modelize and simulate (Anderson et al. 2005), especially from an egocentric perspective: not only embodiment, morphological computation and automatic behaviour remove the need for a full-representation approach to compute behaviour, they also defeat its possibility by hiding much of the complexity from the conscious experience of the agent. In such a context, learning cannot happen without acting.

In many ways, the theories of embodiment echoes and parallels the emergence of the dynamical system theory in developmental psychology (Thelen 1995; Thelen et al. 1996; Smith and Thelen 2003; Thelen et al. 2007), that emerged from concurrent advances in the comprehension of dynamic systems in physics and mathematics. According to this theory:

> *Development can only be understood as the multiple, mutual, and continuous interaction of all the levels of the developing system, from the molecular to the cultural.*
>
> Thelen et al. (2007, p. 258)

In the next section, we will see how robotics has embraced development.

1.2.4 Acting, Learning, and Developing

Robots must act, should learn, and, to learn, they have to act: that much has been established. It remains to be decided, however, what form should take the robot *before* learning begins. In other words, how much knowledge should the robot have about the task to be learned, and which learning abilities should be bestowed upon the robot in order to learn the task. In short: how should robots be born?

For *developmental robotics* (Weng 2001; Lungarella et al. 2003; Asada et al. 2009), the answer lies in developmental processes. Babies have little knowledge of the world, but in a span of two decades, they become fully functional adults. Developmental robotics proposes to reproduce similar growth and maturational processes in robots. That is, robots should learn like children do. The motivations are multiple (see also Pfeifer and Bongard (2006, pp. 141–145) for a discussion).

The first one is that creating a robot that can learn as a child provides a single reusable platform that can acquire many different behaviours, skills and knowledge. Or: *program once, learn anything*. In his seminal article, Turing (1950) alluded to this:

'Our hope is that there is so little mechanism in the child brain that something like it can be easily programmed.'. Decades of research have regrettably proven otherwise. The child's brain is incredibly complex, and reproducing its mechanism into a robot has proven anything but easy. A developmental approach to programming robot is certainly harder than to make a robot learn a specific behaviour.

But doing the latter requires engineers to program into the robot task-specific structures and knowledge that provide an appropriate context for learning. Having engineers designing an adult brain from scratch introduces considerable human bias into the cognitive abilities of the robots. Essentially, the robot is not only told what to do, but also *how to think*. Even if those robots still learn and adapt, they do so in a limited and *task-specific* fashion: acquiring another behaviour requires new cognitive structure to be spoon-fed by engineers. When complex, adaptive behaviour is needed, the work of engineers becomes the one of demiurges.

This gives us the second motivation for developmental robotics: remove as much *designer bias* from the cognitive abilities of robots as possible. Embodiment has a tremendous impact on the development of cognitive abilities of humans and is crucial for typical human behaviour. To replicate this phenomenon and its beneficial effects in robots, roboticists should try to program only general cognitive mechanisms into robots in the first place, and give them the time and the opportunities to discover and grow into their body by themselves, occasionally nudged by social guidance. An interesting implicit assumption here is that human are not competent to program another entity's mind explicitly. First, because they never did it for themselves - much of our individual cognitive development is implicit and self-organized (see for instance Byrge et al. (2014)). And second, because we are limited by our own embodiment, and cannot effectively think what it fully means and represents to have a different one, as discussed previously.

A significant goal of such an approach is to get robots to acquire common sense. For instance, a system asked to build a tower out of wooden cubes might decide that it is a good idea to start by placing the topmost cube, and then, having it stay suspended in the air, to arrange the other blocks beneath it. Acquiring common sense—that we, as humans, take for granted—is frustratingly difficult for robots. Efforts have been made to amass common sense in symbolic databases (for instance, Kochenderfer et al. (2003)), and one could argue that the cube tower problem could be fixed by adding the law of gravity to the robot knowledge. But then one would have to also consider reaction and friction forces, that are no less instrumental. To accurately take those into account, the mass and surface characteristics of each cube would have to be measured. We are fast falling into a full-representation trap, while forgetting that children do not need explicit, symbolic knowledge of Newtonian physics to build wooden castles.

Developmental robotics proposes that robots acquire common sense over a lifetime of experience, by engaging with the world in a similar way children do: through play. Robots should not be directed through specific useful tasks early in their development, but discover the world and its properties on their own. This should not only afford robots common sense, it should afford them common sense adapted to their body and cognitive capacities.

This leads us to a third motivation of developmental robotics: solving the symbol-grounding problem for robots. As articulated by Harnad (1990): 'How can the semantic interpretation of a formal symbol system be made *intrinsic* to the system, rather than just parasitic on the meanings in our heads?' (emphasis his), the symbol-grounding problem inquires about the mechanisms that create meaning in humans. For robots, this means discovering their own ontology; the alternative is to have humans put the meanings in the robots' head directly (Suh et al. 2007; Lemaignan et al. 2010; Tenorth et al. 2013). The literature and debate around the symbol-grounding problem is extensive, and we do not wish to get into it there. Let's just say that children satisfactorily solve the problem. The hope is that robots, given childlike abilities, will figure it out as well.

An interesting consequence of the symbol-grounding problem is of consequence: if symbols, i.e. language, emerge from our sensorimotor experience, then language is dependent on our embodiment. And language is central to the ideas we can form and express: embodiment has influence on what we can think.

A fourth and transverse motivation of developmental robotics is to use robots as scientific tools for understanding biological processes, structures and behaviour, as it has already been discussed (Lungarella et al. 2003). The example of the symbol-grounding problem is illustrative: if we manage to reproduce the phenomenon in robots, it would provide interesting hypotheses for the underlying psychological and neurological mechanisms in humans. And consequently, avenues of investigation for psychologists and neuroscientists. Developmental roboticists are heavily inspired by studies of biological systems. In return, and as the field progresses, the robots have the potential to inform us how to think about our own cognition and those of animals.

The relation between learning and development is subject to different perspectives. Kuhl (2000) proposes four different ones. That development and learning are distinct and do not interact with each other, that learning happens in the context of development, which is the traditional view (Piaget et al. 1953), or that the relation between learning and develompment is more complex, and involves reciprocal influences This is the point of view defended by Kuhl (2000) and by Oyama (2000). The last perspective is the one that does not recognize any reasonable conceptual differences between learning and development. Thelen et al. (1996) defend this position: development is a multi-timescale dynamic system, and learning is just one of its facet.

1.2.5 Acting, Learning, Developing and Evolving

Robots must act, should learn; when learning, they have to do so by acting, and they should preferably go through an extensive developmental process that embeds the learning process into a favourable context. Let's briefly take one more step: they should probably evolve too.

Evolutionary algorithms (Rechenberg 1973; Holland 1975) mimic natural selection, variation and hereditary processes. Candidate solutions are described by their genetic code, which is translated into a phenotype, which is evaluated according to a fitness function. The best performing members of the population are selected, and their genetic code undergoes random variations and mating combinations with other successful solutions. The limited assumption on the fitness landscape has made evolutionary algorithms powerful global optimization methods, useful in complex domains.

The application of evolutionary algorithms to robots—*evolutionary robotics*—has taken off in the last twenty years (Cliff et al. 1993; Meyer et al. 1998; Nolfi 2000; Lipson 2005; Floreano and Keller 2010; Doncieux, Bredeche et al. 2015). One important aspect of evolutionary robotics is that candidate solutions are evaluated by their behaviour rather than by their phenotype.

The motivation for evolutionary robotics is manifold. First, similarly to the developmental approach, it is the only existing process that produced intelligent entities so far. Second, it further allows to remove *designer bias* (Lipson and Pollack 2000). As such, the evolutionary process is not restricted to morphological or hardware considerations, but encompasses morphology, neural architecture, cognitive inborn abilities, and the processes directing and regulating development. Some approaches evolve controllers on a fixed morphology (Zykov et al. 2004), while others evolve morphology with fixed (or non-existent) controllers (Auerbach et al. 2010; Cheney et al. 2013), or co-evolve morphology and behaviour (Sims 1994; Lipson and Pollack 2000; Lehman and Stanley 2011b).

In the last few years, a new domain of inquiry has started to take shape: evolutionary developmental robotics—or *evo-devo-robo* (Jin et al. 2011; Xu et al. 2014). The work of Bongard (2011), for instance, evolves a population of gait controllers, with robots that start with small limbs, and grow them during the experiment. The morphological changes can be perceived as creating a developmental pathway through which acquiring robust behaviour is easier because controllers are filtered by their success on multiple morphologies. Delarboulas et al. (2010) proposes an approach where a robotic platform evolves controllers on-board. Those controllers are selected by using a self-driven fitness that aims at maximizing the sensorimotor entropy, and thus the behavioural diversity of the robot. The platforms further encourage development by comparing the behaviour of each controller against all of their ancestors, and encouraging diversity.

An evolutionary approach seems necessary because, as much as development reduced the problem to the creation of a robot child, it effectively leaves us with the task of designing a body, and mind, and the developmental process to make them grow together. Aside from designer bias, it seems that the complexity involved in such an endeavour, in a great part due to the high coupling of those three aspects, is beyond current human direct, explicit ingenuity (Harvey et al. 1992). Evolutionary processes require a lot of resources and time, but they are a proven way to obtain the result we seek.

1.3 The Exploration Problem

Abstract · Exploration problems are behavioural problems. They make less assumptions about the agent than learning problems, and are suited to analyse developmental processes.

In this thesis, we are concerned with *exploration processes*[14], and we study them as *exploration problems*.

Why Exploration?

> *the induction of novel behavioral forms may be the single most important unresolved problem for all the developmental sciences.*
>
> Wolff (1987, p. 240)

Exploration is a major mechanism in the production and control of behavioural diversity. Which, according to Pfeifer and Bongard (2006), is a crucial component of intelligent behaviour:

> *In spite of all the difficulties of coming up with a concise definition, and regardless of the enormous complexities involved in the concept of intelligence, it seems that whatever we intuitively view as intelligent is always vested with two particular characteristics: compliance and diversity. In short, intelligent agents always comply with the physical and social rules of their environment, and exploit those rules to produce diverse behavior.*
>
> Pfeifer and Bongard (2006, p. 16)

[14] We will use interchangeably *exploration processes* and *exploratory behaviour*, although they may conjure different images in the reader's mind, and may differ in contexts not considered in this manuscript (namely, a behaviour suppose an agent, while a process does not).

Behavioural diversity is a factor of individual robustness: the individual maintains a repertoire of varied interaction possibilities, some of which which will remain relevant the next time the environment changes. Moreover, behavioural diversity provides variability even in the absence of genetic or phenotypic diversity, and improves on them when they are present. This point was recently heeded by the evolutionary robotics community, as we will see in detail in section 2.6. It also impacts the dynamics of evolution. Individual exploratory behaviour translates in the spreading of the species, which in turn affects entire ecosystems, in particular when invasive species are introduced (see section 2.4).

Conversely, diversity has a profound impact on the development of behaviour and cognition. For instance perceptual narrowing, the sensitivity specialization observed in the first year in infants, is influenced by the diversity they are exposed to (Byers-Heinlein et al. 2013).

But actively fostering diversity in the interaction with the environment through exploratory behaviour is equally pivotal. Motor exploration begins *in utero*, and is the driving force behind the creation of the body map and the acquisition of gross and fine motor skills in infants. Neonates are able of sophisticated goal-directed exploratory behaviour (Hofsten 2004), and goal-directed babbling toward objects has been demonstrated in three-months old infants (Sommerville et al. 2005).

In active perception, exploration, as *information seeking behaviour* (Gottlieb et al. 2013), is necessarily present: 'We don't simply see, we *look*.' (Gibson 1988, p. 6).

In fact, humans and animals are intrinsically motivated to explore, and to seek, amongst other, novelty (section 2.4). When learning by trial and error, when playing, when displaying creativity, children are constantly adopting exploratory strategies to figure out what possibilities the world offers (Piaget et al. 1953), while, at the same time, coping with its formidable complexity (Keil 2003). Exploratory behaviour allows to *control the amount of diversity, of complexity* they subject themselves to (Kidd et al. 2012, 2014).

For learning agents interacting with an environment, exploration is the primary way to obtain learning data. In the reinforcement learning frameworks, the importance of exploration is underscored by the importance of the exploration/exploitation trade-off.

For self-sufficient robots, directed exploration through intrinsic motivations has been recognized as a crucial component of the development of rich behaviour in a cumulative learning perspective. Stated differently, directed exploration is a fundamental adaptation strategy for handling new, unknown environments. Intrinsic motivation mechanisms allow to establish a functional dependency between the robot's exploratory behaviour and its experience. This dependency ensures that the robot is directing its exploratory resources towards activities where significant information can be gained.

Moreover, exploratory behaviour enables the robot to create an estimation (even partial, even flawed) of what effects are possible to produce in a given environment, as a result of its own actions. This is important for any number of reasons, but one them is planning.

But perhaps the best motivation for studying exploration in self-sufficient agent is that *exploration is a precursor to learning*. Exploratory behaviour allows to discover learnable interactions—i.e. affordances (Gibson 1977)—in the environment, *before* they are considered as learning problems. For Eleanor Gibson (Gibson 1988), babies are not endowed with the abilities to perceive affordances, but must spend their first years discovering affordances in their environment. For instance, understanding mirrors, for a child, entails first to produce variability in the environment that allows to detect that the interaction proposed by the mirror is unlike other objects[15] (Loveland 1986). Then, a comprehensive exploratory behaviour must be carried out to amass enough observations to figure out what the mirror does. One could argue that the exploratory behaviour in front of the mirror is in fact highly structured, and fall in the child-as-a-scientist paradigm (Gopnik 1997; Schulz and Bonawitz 2007; Gweon et al. 2008; Gopnik 2012). But in many instances, random behaviour is just as informative (or not significantly less informative) that carefully crafted interventions (Cook et al. 2011). Similarly, an crawling infant will progressively discover the 'traversability of a surface of support'. But as experiments from Gibson et al. (1987) pointed out (and in contrast with a walking child), the surface will be engaged before it has been learned, or even accessed for its traversability. Infants have a lot to learn in their first years. Exploration cannot only be considered as a subroutine of learning behaviour. Exploration creates—provokes—contexts where new learning can happen. In other words, exploration happens at different levels, and is not just responsible for the trial-and-error behaviour that drives learning tasks. Exploration happens also *between* learning tasks, and greatly determine which learning tasks are engaged with by the infants.

Most roboticists have been preoccupied by solving problems, with few works seeking to discover them in vast unstructured environments. Without such an ability, a robotic agent can hardly pretend at exhibiting open-ended development.

In the next chapters, we will review some of these aspects more thoroughly.

Exploration and Learning

Exploration can exist without learning. The random motor babbling exploration strategy presented in the first example does not feature any learning behaviour. Likewise, the robot following the left wall in the maze displays a structured exploratory behaviour that guarantees success. It adapts but does not learn. The same can be said of the vehicles of Braitenberg (1986).

[15] Note here that we are not discussing the issue of self-recognition, for which the mirror has been a common experimental paradigm throughout psychological studies. The mirror here is only considered as an object creating singular sensory feedback (Loveland 1986).

Conversely, learning can happen without exploration. As outlined at the beginning of this chapter, a learning system needs not to interact with an environment: the weather system is fed data and predicts outcomes; it does not engage in any exploratory behaviour.

Learning and exploration do not apply to the same classes of entities either. Learning can apply to any system, while exploration, because it necessitates to act in an environment, applies only to agents.

Yet, exploration and learning often depend on one another. Indeed, they are highly complementary. Exploratory behaviour is directly related to information seeking: exploration's aim is to obtain information about the environment. This distinguishes exploration from actions that are purely motivated by exerting control over the environment (Gottlieb et al. 2013). Learning, on the other hand, is interested by exploiting the information gathered about the environment to inform and modify behaviour. Thus, for an agent interacting with an environment, and unless the environment is always providing all the necessary information to the agent without the need to elicit it, exploration is necessary for learning.

And because learning informs behaviour, it can, in particular, inform and improve exploratory behaviour: this is *directed exploration*. The interplay between the two, exploration feeding learning, and learning improving exploration, is at the heart of most interactive learning algorithms.

As an illustrative example, one can take an atypical case of learning: evolution. Evolution learns and explores, in directed and undirected fashion. Evolution's memory is the biosphere, and the organisms are candidate solutions. Natural selection is the learning mechanism of evolution; it keeps in memory only good solutions. Genetic mutation is undirected exploration, while mating is hybrid. While in its simplest form it is undirected, it holds the potential of directed exploration: sexual selection, i.e. when some members reproduce more when they are better at finding mates.

A Behavioural Approach

One may take exception of the examples of pure exploration—the random explorer and the robot in the maze. If exploration's purpose is to gain information, where is the information gain in those instances?

In response, we could consider another question: how can we distinguish between the robot 'mindlessly' solving the maze, and the one that explore the maze 'mindfully', conscious of the effectiveness of the left-hand-rule, and whose goal is to explicitly discover and remember the path to the exit? What about another exploration strategy, that uses a different decision mechanism to choose which direction to go at each turn, but which happens to always choose left on this specific maze? All exhibit the same behaviour. Making the distinction requires to look into the robot's head. All three robots, in fact, create access to the same information. Whether they capture, retain or exploit the information or not is a learning issue, not an exploration issue.

In this thesis, we take a behavioural approach at studying exploration. An exploratory process is considered with regards the information it creates access to—not how the information is used, or if it is remembered at all. Exploration, in other words, is evaluated from behaviour alone. There are several motivations for such an approach.

The first one is that it does not introduce assumptions about the agent's internal mechanisms of exploration into the evaluation. It does not assume that a specific learning mechanism is behind exploratory behaviour, and it does not try to evaluate learning as a proxy for evaluating exploration. The three maze robots have the same performance, the random motor babbling explorer can be compared to the goal babbling explorer, and a Braitenberg vehicle (Braitenberg 1986) can be compared to a SLAM robot (Smith and Cheeseman 1986).

Interestingly, when not having an open-skull access to the subject, discriminating learning from other mechanisms of behaviour is not necessarily trivial. For instance, motor babbling in babies, in particular repetitive kicking motions, have long been thought to be the result of hard-wired pattern generators (Hilgard et al. 1945). But evidences of learning have been found, by observing an improvement the uniformity of the repetition throughout the first year (Kahrs 2012). Similarly, discriminating intentional exploration from noise is not necessarily trivial, or possible. Loeb (2012) argues that the variability observed in human movements, even when subjects repeat the same movement, cannot be trivially attributed to the inherent noise of the musculo-skeletal structure, but can be interpreted as intentional exploration that can be interpreted, under a Bayesian paradigm, as generating enough relevant information to update the prior efficiently:

> *An observer has no way to know how much of the observed variability reflects this purposeful exploration versus computational noise [...]. The subjects themselves may not know.*
>
> Loeb (2012, p. 762)

The second motivation is methodological. Evaluating learning means taking a performance metric, evaluating the learning system, providing the system with experience or letting it acquire some on its own, and then evaluating the system again. The evaluation is quantified by the difference between the two performance values.

Computational learning architectures, and robots, are exceptionally suited for such an evaluation. The learning behaviour can be switched off during evaluation, using only exploitation mechanisms. This allows the learning behaviour to remain unaffected by the evaluation. Humans, on the other hand, do not have the capacity to stop learning. This makes any evaluation of learning a perturbation of the learning behaviour, which has to be accounted for.

Evaluating exploration through behaviour alone allows to use the same methodology on humans and robots. Gottlieb et al. (2013, p. 2) advocated more dialogue and integration between active learning in robotics, and the study of curiosity in psychology and robotics. Using measures than can seamlessly be used across fields is a step in that direction.

Being able to share the same methodology occasionally means being able to more easily create comparable experiments across fields.

Another assumption made when evaluating learning that the robotics and machine learning community often overlooks but that psychologists are acutely aware of, is that to evaluate learning on an agent, one has to have the means to compel the agent to submit to a controlled evaluation. This supposes several things. First, that the agent is able to submit to an evaluation, which means, in machine learning, that the agent can demonstrate either predictive or control abilities. Our random learner has neither of those. Second, that the system is able to understand and is willing to undertake an evaluation. Given the nature of robots, this is usually not a problem. When evaluating learning in infants or animals however, this is one of the main obstacles to the evaluation:

> *How much we have learned about infant behaviour and development in approximately the last half-century is testimony to the ingenuity, patience, and persistence of researchers in meeting and overcoming the formidable challenges posed by infants themselves.* (Bornstein 2014, p. 123)

Third, that the system is available at all. Learning evaluations monopolize the system, which has a significant cost in time and resources in machine learning, robotics and natural studies alike. If a robotic system is operating in real-time in a dynamic environment, the only way to evaluate its learning performance in the middle of the experiment without perturbations is to freeze the robot and environment, to perform the evaluation, and to resume the behaviour of the robot and the dynamics of the environment. For many environments, stopping time or resetting the environment to an earlier state is impossible. The experimenter must decide between more performance data or fewer perturbations.

And fourth, that a controlled evaluation of the specific learning abilities under investigation can be carried out and separated from other phenomena. Again, robots do not usually represent a problem in this regards, but this is a big problem in natural studies.

While all these suppositions seem easily handled by machine learning and robotic experiments, there is the danger that they influence the type of artificial cognitive architectures that will be created and studied by scientists. In other words, roboticists may tend to avoid cognitive architectures that do not have a clear and explicit switch somewhere that deactivate learning. Or robots that are difficult or impossible to reset to initial conditions—because, for instance, their bodies are irreversibly modified by the interaction with the environment.

Using behaviour to evaluate exploration avoids or reduce the importance of most of these problems. Exploratory behaviour still has to be elicited, usually in a controlled environment. But because the experimenter does not have to engage with the subject of study to conduct the evaluation, he can opportunistically rely on observations in the wild, or on the recorded behaviour of past experiments, or of since disappeared agents.

The behavioural approach avoids bias about the agents internal operation, and does not create nearly as much methodological difficulties than a learning evaluation entails. And it has the added benefits to be easily applied in different disciplines.

The third and final motivation for the behavioural approach is that is that learning performances are too limited to fully evaluate developmental processes.

This problem is particularly acute in the case of open-ended development. One ambition of developmental robotics is to create robots that do not stop learning, that explore their environments on their own and build solid foundations of knowledge and skills that make them capable. Evaluating open-ended development leads to challenges: On which problem should the robots be tested? Should it be the same for all robots or should it depends on their developmental trajectories? Even then, how can a meaningful set of learning tests be created when the set of skills that can be learned in the environment is difficult or impossible for the experimenter to establish?

The problem is analogous to a pair of twins, raised identically, which are one day given only one instruction: to learn what they want, and to do their best. One decides to learn the piano, the other opts for the rugby team. How then can their performance in these two tasks be compared? The problem does not go away with additional constraints: if one chooses tennis and the other rugby, the developmental trajectories are perhaps more similar, but not necessarily any more comparable from a learning standpoint.

Faced with developmental processes exhibiting self-organizing behaviours, current research, in particular on intrinsic motivation, regularly relies on behavioural measures to qualify and quantify the results alongside learning performances (Merrick and Maher 2009). For instance, in a multiobjective setting, Oudeyer, Kaplan and Hafner (2007), Merrick and Maher (2009) and Moulin-Frier, Nguyen et al. (2014) analyse the respective time the agent spends at each task. Stulp and Oudeyer (2012) also devoted a study to the self-organization of the behaviour elicited by a learning algorithm, PI^2_{CMA}. Delarboulas et al. (2010) evaluate the agent by the number of different locations it was able to visit over a continuous map, and Rolf (2013) uses a workspace coverage measure. In other words, the behaviour of a developmental process is as much worthy of scientific study as is the learning performance it produces.

To be clear, we are not advocating—at all—a strict closed-skull approach to the study of behaviour in computational agents. That would be ridiculous. But it is not because offline, open-skull measurements can be made that we should not avoid correlating them with the behaviour displayed *during* learning.

Formalization

We restrict the class of exploration problems we study in important ways: we consider one-step episodic environments, where one input corresponds to one output. And the environment has no memory or context, and does not change.

Environment and Tasks

An *environment* is formally defined as a function f from M to S. M is the motor space, and it represents a parametrization of the movements the robot can execute. It is a bounded hyperrectangle of $\mathbb{R}^{d_M}$, with d_M the dimension of the motor space. S is the sensory space; it is a subset of $\mathbb{R}^{d_S}$, with d_S the dimension of the sensory space. *Effects* and *goals* (desired effects) are elements of S[16].

A *task* is defined as a pair (f, n) with $f : M \mapsto S$ the environment and n the maximum number of samples of f allowed, i.e. the number of actions the robot can make in the environment.

Defining the environment as a function implies determinism. As no other assumption is made on f, non-determinism can be approximated by with a chaotic environmental feedback function, coupled with noisy motor communication pathways. We adopt a functional formulation here to avoid unnecessary burden[17].

Exploration

The objective of the exploration problem is to estimate what elements of S can be produced by f, i.e. to estimate the image of f, $f(M)$, designated as the *reachable space*.

An exploration strategy evaluates the function f, n times, providing a sequence of elements of M, $\mathbf{x}_0$, $\mathbf{x}_1$, ..., $\mathbf{x}_{k-1}$. Each $\mathbf{x}_i$ is evaluated as $\mathbf{y}_i = f(\mathbf{x}_i)$, and $\mathbf{y}_i$ is observed by the exploration strategy before $\mathbf{x}_{i+1}$ is chosen.

Each observation $\{\mathbf{x}_i, \mathbf{y}_i\}$ provides information on $f(M)$. Yet estimating $f(M)$, a possibly continuous, infinite subset of $\mathbb{R}^n$, from a finite set of points is not a well-defined problem. For this, we rely on a *diversity measure*, that is not necessarily known by the exploration strategy.

This last point is important. We do not assume that the agent has knowledge of the diversity measures that are used as evaluation. It certainly makes our work more difficult. Agents might explore with different goals in mind, and evaluate their own behaviour according to metrics we don't have access to. The choice of a diversity measure or the other can therefore be seen as arbitrary. This problem is not present, for instance, in reinforcement learning, where the cumulative reward defines an objective motivation for the agent, and an objective evaluation for the experimenter.

Yet, to allow comparing agents with different motivations, one cannot establish a

[16] We assume that S is known by the exploration strategy, but nothing prevents S to be set equal to $\mathbb{R}^{d_S}$

[17] For coincidental technical reasons, the simulation setup we will present in chapter 7 is not deterministic, as it turns out.

measure as better than all the others. In the context of the exploration problem, any exploratory measure, as a behavioural measure, is arbitrary. It is the responsibility of the experimenter to justify its interest.

Agents are, of course, free to use diversity measures to self-evaluate their exploration, and will we see such an agent chapter 4. There is just no guarantees that the experimenter will use the same one.

Discussion

One particularity of our approach is that we do not *a priori* evaluate the value of the diversity produced with respect to a specific objective. In rich environments, there are many ways to produce diversity easily which has little objective value without much effort.

We can, of course, define the evaluation measure so has to encode the achievement of a specific goal in it. But we do not do that here. In the example of the first chapter, we solved this issue by considering a sensory space that only encodes valuable diversity. This is certainly not a good way to proceed in a more general setting.

But this allows to focus on the production of diversity independently of other interests. The production of diversity is overwhelmingly studied in relation to its value for learning performance. This thesis does not focus on that.

Before moving on defining diversity measures, let's clarify the relationship between diversity and novelty. Novelty is a property of a newly acquired observation, in relation with observations already present in memory. Diversity targets the whole population of acquired observations.

An agent driven by diversity may either be motivated by maintaining a certain level of behavioural diversity toward a certain aspect of its behaviour, or be motivated to estimate the range of diversity a phenomenon offers. The first one, as we pointed out, can be motivated by a higher survival robustness and fitness: it keeps options open. In a specific situation, only one learned behaviour might be successful. Successful behavioural diversity can decrease with changes in the environment or the agent, and thus maintaining it may require ongoing exploratory behaviour. Moreover, as new skills, new affordances are regularly discovered by children, each of them may require to develop its own amount of diversity. This differs from simple novelty-seeking, as it allows to predict that agents will stop exploratory behaviour once cumulative novelty has reached a certain threshold on a task or phenomenon, regardless of the novel interactions available to them at the time.

This type of diversity is the one found in biological populations: there is a motivation to avoid inbreeding, and maintain genetic and phenotypic diversity. This is also precisely the purpose behind the diversity-driven evolutionary robotics methods we'll review in section 2.6.

The second motivation for diversity, sampling the whole range of a phenomenon, deals with understanding the possibilities the environment offers. This, in turns, allows for better exploitation. This is the purpose of the exploration of a Multi-Armed Bandits scenario, where the agent must sample different sources of rewards to find the one that is the highest (we discuss Multi-Armed Bandits scenario in chapter 4). Baranes, Oudeyer and Gottlieb (2014) proposed an experiment where adults were able to sample a set of different tasks, and found that they would sample the whole range of tasks, even as some were impossible. Yet, this experiment does not allow to discriminate between simple novelty-seeking and diversity-seeking.

A possible experimental framework would be to provide an unbounded set of tasks, that could not be sampled meaningfully during the allowed sampling period. The tasks would vary across several identified dimensions, some of which offering bounded variation. A simple novelty-seeking agent would sample tasks in no particular direction, amassing a set of diverse observations. A diversity-driven agent would preferentially explore along dimensions of bounded variations, aiming at understanding globally the possibilities offered by the task set on specific aspects. Note that the second strategy is also better in the long run: it tries to exploit the combinatorial nature of the task set by decomposing the diversity along dimensions. If the relationship between dimension is not too non-linear, this results in an estimation of the diversity exponentially faster than the novelty-seeking approach.

We are not aware of any work aimed at discriminating in children between a motivation over novelty and the two-type of motivation for diversity we have identified.

1.4 Diversity Measures

Diversity measures quantify exploration and allow to compare exploration algorithms. In the following, we consider different general-purpose diversity measures. Each of them expresses different assumptions about the explored space. An overarching assumption lies on the locality of the explored space: an area of the sensory space is considered explored if the nearest observed effects are not too far.

Two classes of diversity measures can be distinguished. *Global diversity measures*, that evaluate the exploration with respects to the possibilities offered by the environment. And *intrinsic diversity measures*, that quantify the diversity of a distribution of effects without regards for the reachable space.

Global Diversity Measures

In this section, we assume that the reachable space, $f(M)$, is known by the diversity measure. As we only target sensory diversity, no measure is sensitive to the motor commands $\{\mathbf{x}_0, \mathbf{x}_1, ..., \mathbf{x}_{n-1}\}$—they only take the effects $E = \{\mathbf{y}_0, \mathbf{y}_1, ..., \mathbf{y}_{n-1}\}$ into account.

Maximum Distance Measure

An overarching assumption we make in all the diversity measures we expose here is that a given area of the reachable space $f(M)$ is qualified as explored depending on how far it is from a produced effect—that is, from a point of E. As such, the *maximum distance measure* provides a global quantification of the exploration of the reachable space.

Definition 1. *Assuming that $f(M)$ is bounded, we define the* ***maximum distance measure*** *as the maximum distance between a point of $f(M)$ and its nearest neighbour in E, over all points of $f(M)$, i.e.:*

$$sup_{\mathbf{y}\in f(M)}\{min_{\mathbf{y}_i \in E}\|\mathbf{y} - \mathbf{y}_i\|\|\}$$

The maximum distance measure does not discriminate between situations that represent qualitatively significantly different explorations. In Figure 1.5, the two explorations have the same maximal distance value, but one manages to produce effects distributed over more than half the reachable space, while the other only produces one effect. To mitigate this, we now introduce averaged distance measures.

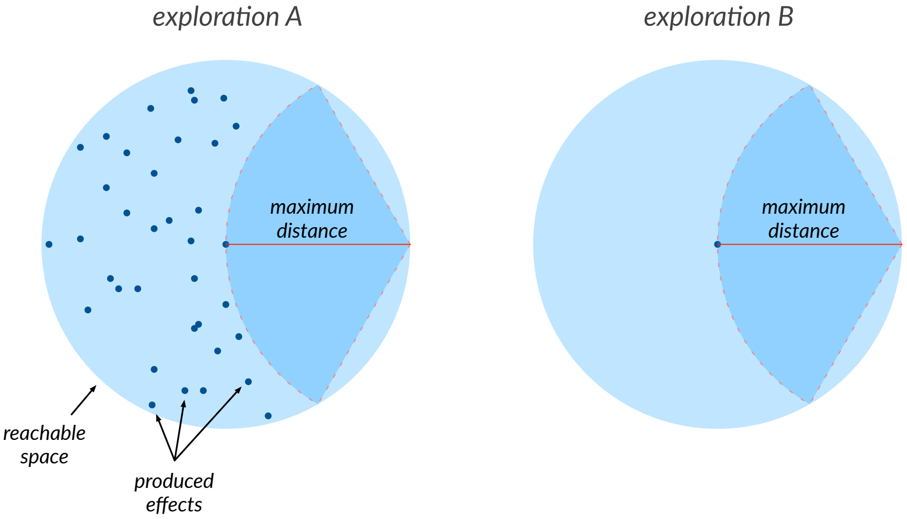


Figure 1.5: While exploration A produces more diverse effects than exploration B, they have the same maximal distance measure.

Average Distance Measure

To avoid the pitfalls of the maximum distance measure, we can compute the average, rather than the maximum, of the distance between the produced effects and the reachable space.

If the set of reachable effects is finite, of cardinal N_f, then, the average distance is defined as:

$$\frac{\sum_{\mathbf{y} \in f(M)} \{min_{\mathbf{y}_i \in E} \|\mathbf{y} - \mathbf{y}_i\|\|\}}{N_f}$$

Usually however, we will consider continuous reachable spaces. In this case, we need to integrate. For this, we are constrained to reachable spaces $f(M)$ whose volumes are defined and non-null. Using Lebesgue integration, the volume is defined as:

$$volume(f(M)) = \int_{f(M)} 1 \, \mathrm{d}\mathbf{y}$$

$volume(f(M))$ is the Lebesgue measure of $f(M)$—often noted $\lambda^*(f(M))$ in the literature.

Definition 2. *Assuming $f(M)$ compact, with $volume(f(M)) \neq 0$, we can compute the* average coverage error *as the average distance*[18] *of $f(M)$ to C.*

$$\frac{1}{volume(f(M))} \int_{f(M)} min_{\mathbf{x} \in C} \|f(\mathbf{x}), \mathbf{y}\| \, \mathrm{d}\mathbf{y}$$

Computing the average distance is usually intractable, and perfect knowledge of $f(M)$ is a limiting requirement. Moreover, the average distance also does not consider isolated points of $f(M)$, as they receive a weight of zero during the integration, which may not be desirable. As such, the average distance is not a practical diversity measure. To solve those problems, we discretize the reachable space.

Testset-based Average Distance

Definition 3. *Given a set of points T belonging to $\mathbb{R}^n$, the testset-based average distance of a finite set of points $E \subset f(M)$ is defined as:*

$$\frac{1}{card(T)} \sum_{\mathbf{y} \in T} min_{\mathbf{y}_i \in E} \|\mathbf{y} - \mathbf{y}_i\|$$

The testset-based average distance allows to evaluate how well the observed effects cover a manually defined set of goals of particular interest.

[18] This is similar to the Hausdorff distance (Hausdorff 1914), but averaged to avoid giving too much importance to outliers. See Schütze et al. (2010) for a formal definition in the discrete case.

We can also define a squared variant:

$$\frac{1}{card(T)} \sum_{\mathbf{y} \in T} min_{\mathbf{y}_i \in E} \|\mathbf{y} - \mathbf{y}_i\|^2$$

The squared variant can be understood as a mean square error (MSE) estimator for the nearest neighbour inverse model. The nearest neighbour inverse model returns the motor commands corresponding to the closest observed effect from the goal. As such, it can be understood as providing a baseline for learning performance against which any more complex inverse models can be measured. Consequently, a squared testset-based average distance is compatible with evaluating both exploration and learning.

We now present two methods to create a testset that approximates the average distance measure: one applicable when no isolated points are present in $f(M)$, and another that takes into account isolated points.

Lattice Restriction

The idea behind lattice-based testsets is to be able to select a finite set of points T in $f(M)$ with an arbitrary low maximum Hausdorff distance between T and $f(M)$. We restrict our discussion to the case where $f(M)$ is bounded.

Given a point lattice $\mathcal{L}$ over $\mathbb{R}^n$, if we consider the restriction of $\mathcal{L}$ to $f(M)$ we obtain a finite testset. If we further constraint T so that for every point of T, there is an open neighbourhood of that point included in $f(M)$ (thus ignoring isolated points), then we can approximate the average distance measure by reducing the coarseness of the lattice[19]. Furthermore, because T is a subset of $f(M)$, the measure lower bound is zero.

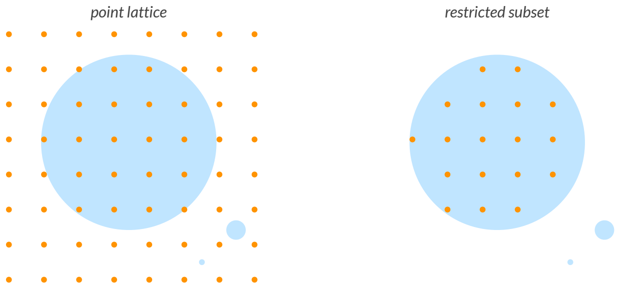


Figure 1.6: The restriction of the lattice to the reachable space provides and adequate testset for the coverage measure, but misses the small region because of the high coarseness of the point lattice. The isolated point is not considered.

Such a testset provides a tractable method for evaluating exploration. By modifying the coarseness of the lattice (i.e., the norm of the vector of the basis of $\mathbb{R}^n$ from

[19] We do not claim that the limit is equal to the average distance measure when the coarseness goes to zero. For practical purposes, and on the reachable space we consider, the approximation is sufficiently precise.

which $\mathcal{L}$ is generated), we can balance the precision and the computational cost of the measure.

Lattice Adaptation

In order to take isolated points into account in a robust way, we adapt the lattice using a two-pass nearest neighbour algorithm.

We define the *proximal subset* P of the lattice as the set of points in $\mathcal{L}$ that are nearest neighbours of a point in $f(M)$. Formally, for each point $\mathbf{y}$ of $f(M)$, we consider the minimal distance from $\mathbf{y}$ to $\mathcal{L}$. The points of $\mathcal{L}$ that are at minimal distance of $\mathbf{y}$ are its nearest neighbours. P is the union of the nearest neighbours for all points over $f(M)$. Since $f(M)$ is bounded, P is finite.

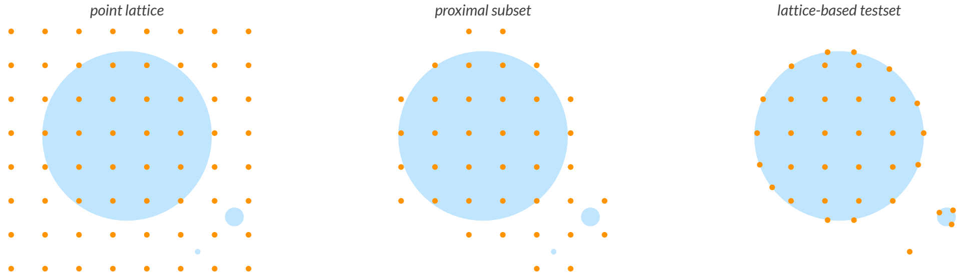


Figure 1.7: Considering a reachable space, here in blue, and a lattice, the proximal subset P of the $\mathcal{L}$ is first selected, and then projected onto $f(M)$. The testset correctly takes the isolated point and the small region into account, even when the lattice is relatively coarse.

For each point p of P, we now consider the set of its nearest neighbours in $f(M)$, and choose one at random if more than one exists. The *testset* T is the union of these nearest neighbours[20] over P. Figure 1.7 illustrates the process. The testset does correctly take into account isolated points in $f(M)$, even if the lattice is coarse.

Heterogeneous Evaluation

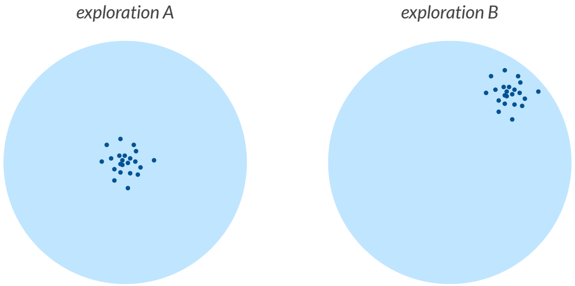


Figure 1.8: Same exploration pattern, but different testset-based distance measure.

So far, the measures presented make the assumption that each part of the reach-

[20] If no point of P has more that one nearest neighbour in $f(M)$, T can be understood as the projection of P onto $f(M)$.

able space has the same exploration value. An experimenter might want to give more weight to some areas of the reachable space. This can be easily done by considering different lattices with varying level of coarseness for different areas of the reachable space, or by manually defining the testset so that its local density matches the exploration weight of the area. Another option is to manually define the weight of each point of the testset. We will not consider such measures in this manuscript.

Moreover, the measures have been evaluating the exploration with respect to the entire reachable space. These measures were aimed at discovering the whole range of effects that were possible to create in the environment. As we discussed previously, diversity can also be aimed at producing enough cumulative novelty. The next measure we introduce estimate this type of diversity.

Estimating exploration over the entire reachable space requires both knowledge of $f(M)$, and $f(M)$ bounded. Although in simple cases, those requirements are reasonable, they are not in complex environments. This leads to undesirable side-effects. In Figure 1.8, the same exploration pattern is evaluated differently with regards to its overall location in $f(M)$. Moreover, in large reachable spaces, with a very limited number of interactions, it rewards explorations where the observations are far from one another. While this may be desirable, a less aggressive diversity measure might better suit other situations. For instance, a representative sample of diversity over a local area of the sensory space might be easier to exploit or to learn from.

Intrinsic Diversity Measures

We introduce now a diversity measure that does not require $f(M)$ to be known or bounded, avoids the pitfalls of Figure 1.8, and is only sensitive to the distance between effects up to a defined threshold.

Threshold Coverage

Threshold coverage considers the volume of the union of the set of hyperballs of radius τ—the threshold—that have for centres the observed effects. Figure 1.9 illustrates this for the two-dimensional arm.

Definition 4. *Considering a set of points C belonging to $\mathbb{R}^n$, and $\tau \in \mathbb{R}^+$, we define the τ-coverage of C as:*

$$\mathit{coverage}_{\tau}(C) = \mathit{volume}\left(\bigcup_{i=0}^{n} B(\mathbf{y}_i, \tau)\right)$$

with $B(\mathbf{y}_i, \tau)$ the hyperball of centre $\mathbf{y}_i$ and radius τ.

The τ-coverage measure is particularly useful when the reachable space is not known. For this reason, it is an *intrinsic* measure: the agent is able to compute it on its own

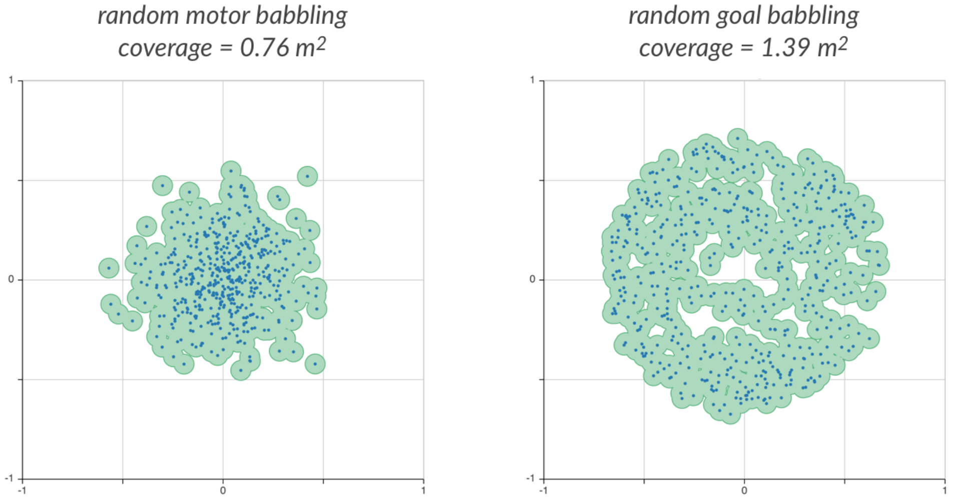


Figure 1.9: Threshold coverage quantifies the area of the space that has been reached at a given precision. The graphs show the coverage of random motor babbling and random goal babbling strategies from chapter 0 on a 20-joint arm over 500 timesteps. [source code]

without knowledge of the environment. We discuss an exploration strategy making use of it in chapter 4.

The τ-coverage is insensitive to how spread the effects produced are, if they are farther apart than the threshold. In Figure 1.10, the threshold coverage is the same, but the effects produced are very different.

exploration A
exploration B

Figure 1.10: Threshold coverage is insensitive to effect spread over the threshold. The two explorations have the same threshold coverage.

The τ-coverage is problematic in high dimension, because it turns out to be difficult to compute, even if we simplify the hyperballs to axis-aligned hyperrectangles (see appendix A on this issue). Therefore, we introduced an approximation of it, in the context of the chapter 4, that is available in appendix B.

Sparseness

Lehman and Stanley (2008) and Ollion et al. (2011) have proposed a diversity measure based on nearest neighbours. The *k-sparseness* $s_k(\mathbf{y}_i)$ of an effect $\mathbf{y}_i$ is the average distance from its k nearest neighbours:

$$s_k(\mathbf{y}_i) = \frac{1}{k}\sum_{j=0}^{k} \|\mathbf{y}_i - \sigma_j(\mathbf{y}_i)\|$$

with $\sigma_j(\mathbf{y}_i)$ the jth nearest neighbour of $\mathbf{y}_i$ in the observed effects.

The sparseness measure of the exploration is then defined as

$$\frac{1}{n}\sum_{i=0}^{n} s_k(\mathbf{y}_i)$$

The higher the k-sparseness measure is, the better the exploration. The measure is robust when only one cluster is present. However, if multiples distant clusters of effects are discovered during exploration, this creates high fluctuations in the sparseness value. The sparseness value increases abruptly when the new cluster is discovered, and decrease equally abruptly when the cluster contains k effects. This is illustrated Figure 1.11.

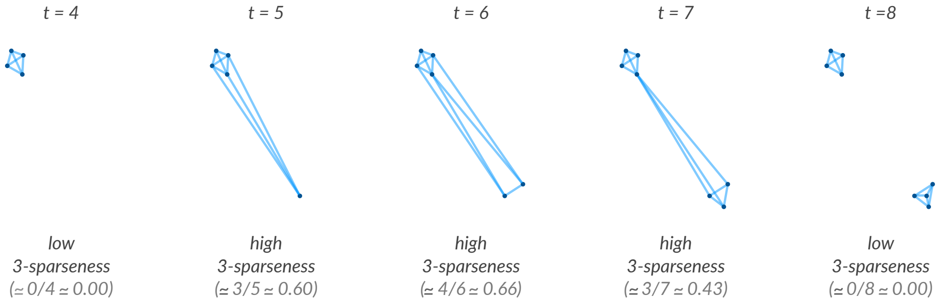


Figure 1.11: Sparseness can fluctuate abruptly because of clustering effects. Here, with k = 3, the initial 4-point cluster has low 3-sparseness. When a point is found in a new, distant cluster, the sparseness increases sharply. It stays high as a second and third point are added to the new cluster because each point of the new cluster must find part of their neighbours in the other cluster. When an additional point is added to the cluster, all neighbours are local, and the sparseness value decreases significantly. The sparseness is numerically estimated in this example by considering the intra-cluster distances negligible (note however that all the intra-cluster distances considered null here are involved twice in the sparseness value), and the inter-cluster distance equal to 1.0.

This can be avoided if $k = n$, but as Doncieux and Mouret (2010) remarks that implies making $n(n{-}1)/2$ distances computations (but also avoids having to compute the nearest neighbours). This is hardly feasible for high values of n.

A possible solution to the high number of distance computation is to take a page out of the book of particle physic engines. Faced with simulating a quadratic number of gravitational or electromagnetic interactions to compute between high numbers of particles, particle physic engines sometimes resort to a quadtree approximation. For instance, for mass interactions, the interaction between a particle and a distant cluster of particles is approximated to the interaction with the particle and a body whose mass is the same as the one of the distant cluster, and whose centre is the averaged centre of the cluster.

This approximation is constrained by the similarity measure used. Here, we only consider the euclidean distance as a similarity measure, which is particularly adapted to the quadtree approximation proposed. See Doncieux and Mouret (2010) for a discussion of the similarity measures considered in evolutionary robotics.

Alternatively, diversity measures such as sparseness that are brittle to cluster structure (see Olorunda et al. (2008) for other examples) can be supplemented by clustering algorithms, that decompose the diversity computation to a cluster basis. This has the added benefit to sharply reduce the number of similarity computation needed if many clusters are present. An additional diversity estimation—for instance, using sparseness—can be done between clusters.

Entropy

Delarboulas et al. (2010) have used the entropy of the sensorimotor stream to quantify behavioural diversity. From a set of discrete observations, computing entropy can be done by grouping observations into classes. Delarboulas et al. (2010) uses k-means and ϵ-means (Duda et al. 2001) to create classes. One could also use a non-parametric clustering method (see for instance Gershman et al. (2012)) to avoid imposing a specific number of classes.

Given p classes, with respectively $n_0, ..., n_{p-1}$ members in each class, the entropy (Shannon 1948) is defined as:

$$-\sum_{i=0}^{p-1} \frac{n_i}{\sum_{j=0}^{p-1} n_j} \log \frac{n_i}{\sum_{j=0}^{p-1} n_j}$$

Note that Delarboulas et al. (2010) clusters over the whole sensorimotor observations (sensors *and* motor) whereas we are just interested by a sensory clustering. The entropy is a robust measure that benefits from a solid theoretical background, but it is sensible to the number of classes. It also does not take into account the relation between the classes: two observations can be arbitrarily close, but belong to different classes, and considered completely differently by the entropy measure.

Measures of Biodiversity

Diversity measures have a long tradition of usage in ecology to quantify species diversity. As species form natural classes, entropy measures are often used in those domains (Pianka 1966; Hurlbert 1971; Whittaker 1972; Peet 1974; Cousins 1991; Lande 1996; Purvis et al. 2000; Davies et al. 2011). The three most used measures are species richness (the number of different species, regardless of their abundance), Shannon information (as presented previously, Shannon (1948)), and the *Gini index*

(Gini 1912)[21], that represents the probability that two individuals chosen randomly are from different species (also used as the *Simpson concentration* (Simpson 1949), that expresses the opposite probability: that two individuals chosen randomly are from the same species).

Page (2011) distinguishes between three types of diversity. Diversity of types which defines diversity amongst classes of entities, such as biological species. Diversity within a type, which quantifies the variations of entities of the same class. And compositional diversity, which describe the diversity that arise from the arrangement of different entities, such as genes. Of course, defining classes can be arbitrary, and a variation can become a diversity of type depending on the perspective.

This list of measures hardly exhaustive. The use of diversity in machine learning is further discussed in section 2.7.

1.5 The Explorers Framework

To express and compare the algorithms investigated in this manuscript, we introduce the *Explorers* framework. The framework is largely strategy-agnostic, and can naturally express motor babbling, goal babbling and intrinsically-motivated exploration algorithms. The framework is designed around small, simple and well-understood modules that do not incorporate too much sophistication. Modules can be arranged in many ways to obtain diverse algorithms. This also makes the exploration algorithms easily reusable in a larger cognitive architecture.

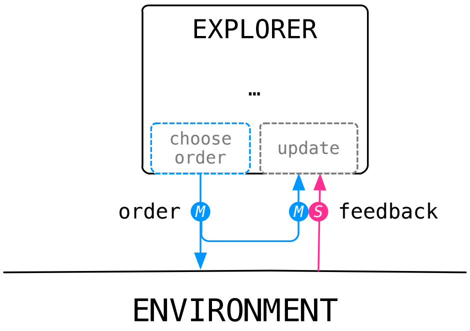


Figure 1.12: The only requirement for an explorer is to provide orders to be executed by the environment (which includes actuators).

[21] Interestingly, despite its influence, Gini's original work has never been translated from Italian (Ceriani et al. 2012, p. 421).

At the centre of the framework is the *explorer* concept. The explorer is the module communicating with the environment: it provides motor commands for the environment to execute and receives observations (Figure 1.12, the feedback signal and update component is subsequently assumed for all explorers and not pictured). But an explorer can also be integrated in a larger architecture where it does not have access to the environment directly. This makes hierarchical (or organized around a more general graph) architectures natural.

We can easily express a goal babbling architecture (Figure 1.13) in the *Explorers* framework. The explorer interacting with the environment allows to filter motor commands that are proposed by the inverse model, and eventually to select another goal if the motor command is not satisfactory or possible to execute.

In the interest model architecture, the interest model provides goals, and leaves the selection of the order to the learner (Figure 1.14). This is the architecture proposed by other architectural frameworks for goal-directed intrinsic-motivation (Hervouet et al. 2013; Moulin-Frier, Rouanet et al. 2014); the intrinsic-motivation component is considered as an add-on destined to guide a learning architecture. This leads to potential problems, where the learner is given the responsibility to act appropriately given an intention it did not originate, and may for instance produce orders that have already been tried without success. The explorer in our architecture is responsible for both originating the goal and ensuring that a suitable motor command is found for it. This allows to flexibly change goal or reject motor commands before execution for any reason. In other words, our architecture does not explicitly divorce decision from execution.

The loop between the explorer and the learner can be exploited to create architectures where one explorer has more than one learner. One goal can be dispatched to the inverse models of all the learners, and the explorer can filter the results to decide

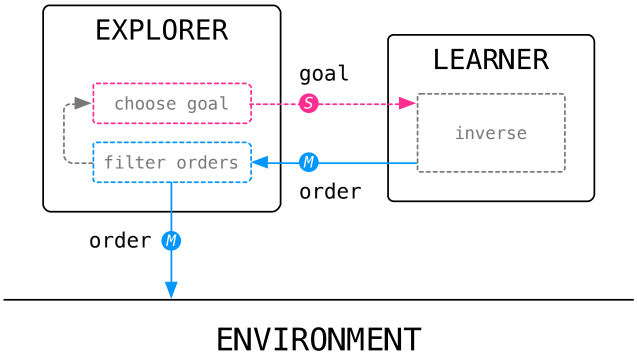


Figure 1.13: In this goal-directed algorithm, the learner does not interact with the environment. It is used by the explorer to create orders that correspond to goals expressed by the explorer. Orders can be rejected by the explorer for any reason, in which case another goal is chosen (grey arrow). The feedback is not shown here.

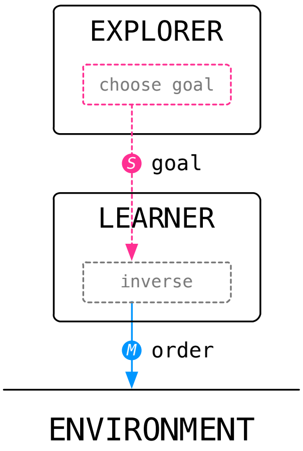


Figure 1.14: In the interest model architecture, the learner is the interface between the agent and the environment.

which order to execute. Heuristic about filtering orders can for instance be based on the confidence the learner expressed in their inference, if such a signal is available. If some learners are accurate but slow and expensive while others are fast but imprecise, the explorer can exploit the choice they offer by selecting which learners to poll given the situation and the resources (time, power) available. This type of architecture also allows the explorer to handle a set of heterogeneous goals which require different learners.

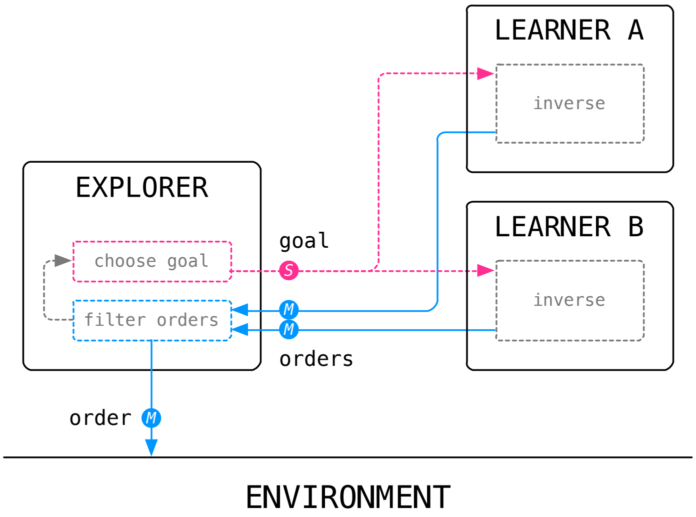


Figure 1.15: The explorer decides which of the two orders to execute once they have been generated by the learners as its disposal. Alternatively, it can preemptively choose to only ask one of the two learners.

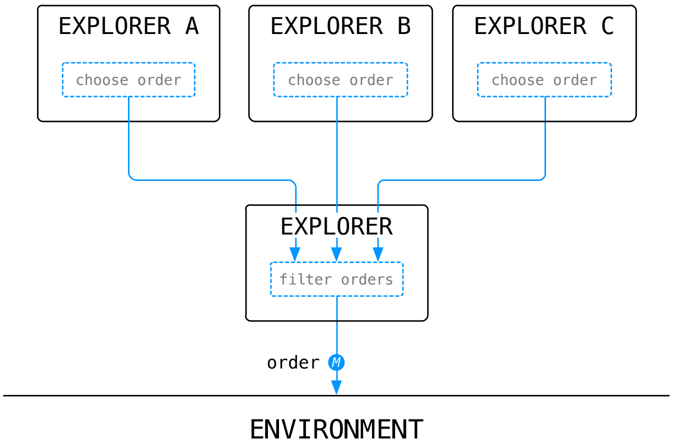


Figure 1.16: Hierarchical explorers are straightforward. In the experiments we are presenting, multiple explorers are active.

The framework handles straightforwardly hierarchical architectures where different exploration strategies are unified into a global explorer. Instances of such an architectures are investigated in chapters 3 and 4.

Science is exploration. The fundamental nature of exploration is that we don't know what's there. We can guess and hope and aim to find out certain things, but we have to expect surprises.

Charles H. Townes

2 Bibliographical Remarks

Exploration processes intertwine with a wide range of research strands. The aim of this chapter is to give a broad overview of some of the research related to exploratory behaviour and behavioural diversity.

We first discuss active learning, that introduced some of the first methods of directed exploration. Then, a brief exposition is done of self-organization, which permits to position some recent works in sensorimotor exploration. From there, we expose how exploration processes are involved in the early development of infants and review some of the history of the research on human and animal exploratory behaviour and motivation. Integrating both the ideas of active learning and the theories of motivation from psychology and neuroscience, we survey then the computational approach to intrinsic motivations, with an accent on novelty-based motivations. Finally, evolutionary approaches are discussed, in particular in how they use diversity as selective pressure.

2.1 Active Learning

In the context of classical machine learning, the idea behind *active learning* (Hasenjäger et al. 2002; Lopes and Oudeyer 2010; Lopes and Montesano 2014; Dasgupta 2011; Settles 2012) is that learning performance can be improved if the learning algorithm is able to choose the observations it want to make on the phenomenon (i.e., the inputs it wants to try). Of course, this is not always possible. Our weather predictor can't apply meteorological conditions across the planet just because it would dissipate some ambiguity in the models.

An easy, somewhat simplistic, analogy between classical 'passive' learning and active learning would be searching for the presence of a particular element in a sorted sequence. Passive learning would go through all the elements at random or in sequence, while active learning would employ a binary search, accessing the middle of the sequence and recursing on half of it, thus testing exponentially fewer elements than passive learning. Active learning not only takes advantage of the underlying structure of the data to select inputs that yield observations that are relevant to the problem at hand, but also takes into account previous observations to decide the next actions. Active learning is the computer playing 'twenty questions' on a given problem with the environment.

The majority of the active learning literature deals with classification tasks. Typically, large amounts of unlabelled data are available for the classifier, but obtaining a label has a cost. For instance, to train a classifier on sentiment analysis in social networks, there are huge numbers of posts available. But for each of them, a team of human must describe the sentiment expressed, in a consistent manner[1]. An active learning algorithm can dramatically decrease the cost of learning by only asking labels for data that significantly improve its classifying performance. For example, by only asking about users' posts that are near the decision boundaries, i.e, user posts the algorithms is not confident about. In fact, this approach is known as *uncertainty sampling* (Lewis et al. 1994). Other methods are guided by prediction errors (Thrun and Mitchell 1995), variance (Cohn, Ghahramani et al. 1996), disagreement between hypotheses (Cohn, Atlas et al. 1994), disagreement among a comity (Seung et al. 1992; Breiman 1996; Freund et al. 1997), or expected improvement (Jones et al. 1998).

Active learning is relevant to our discussion because it augments learning with directed exploration. The exploration is aimed at finding information in the environment that can best improve the knowledge of the agent. It is not surprising then that most methods drive exploration with concepts directly linked to the learning performance.

Settles (2012) distinguishes between three flavours of active learning: *pool-based approaches*, where a large, finite, set of unlabelled examples is available; *stream-based* approaches where unlabelled samples are observed sequentially, and the algorithm can decide whether to label or discard each one; and *query-based* approaches, where examples are parametrized by features with given ranges, and the algorithms can ask for any combination of feature values it wants.

Query-based approaches are closely related to the *optimal experimental design* problem in statistics (Fedorov 1972; Chaloner et al. 1995; Pukelsheim 2006), where, for

[1] Consistency is hard, and to detect it, more than one individual are sometimes employed to label the same data, which increases costs.

a given hypothesis, a minimal number of experiments must be designed to refute or prove the hypothesis. While traditional active learning focuses on classification, optimal experimental design generally deals with regression problems, and the two fields generally do not use the same techniques. An example of a regression problem is the modelization of the underground in geostatistics: core samples are collected over an area, and a model of the underground must be created[2]. Evidently, core samples are costly to collect, and the monetary incentives to get accurate models of mineral resources are sometimes high. Optimal experimental design fits the robotic agent problem nicely: an experiment is a motor command, costing time and energy, and its result is the sensory feedback.

Optimal experimental design has been applied to robotic learning to discover body schemas (Martinez-Cantin et al. 2010) or to learn forward kinematics (Cohn, Atlas et al. 1994; Cohn, Ghahramani et al. 1996). Bongard and Lipson (2005) use an evolutionary algorithm to synthesize a model of the robot from empirical data by co-evolving a population of candidate models and a population candidate *tests*. While the models attempt to explain the internal systems of the robot, the tests are aimed at extracting observations from the real robot to provide better information to the models: they are a set of experiments that are improved throughout the evolution process.

Aside from specific robotic applications such as these, active learning and optimal experimental design approaches typically make assumptions that make the methods they propose unfit to be directly applied to robotic setups. Two of them are particularly disastrous: the assumption that the model is completely learnable, and that the noise is homogeneous (Oudeyer, Baranes et al. (2013), see also section 1.2.2). For instance, Freund et al. (1997) proved that active learning could result in an exponential decrease in sampling to reach a given precision in some setting, but did so under the assumption of noiseless, deterministic environments.

Related to active learning, and more specifically, to optimal experimental design, an new paradigm has recently had a major influence in theories of child cognitive development. The *child-as-a-scientist* paradigm (Gopnik 1997; Schulz and Bonawitz 2007; Gweon et al. 2008; Cook et al. 2011; Gopnik 2012) considers the hypothesis, that, rather than acting randomly in the world, children act as rational thinkers, creating experiments and testing hypotheses through their interaction with the world in a manner structurally similar to scientific inquiry. Convincing experiments have indeed showed that preschoolers understand causality, can distinguish it from spurious associations, and construct interventions to do so (Gopnik et al. 2001; Schulz, Gopnik et al. 2007).

Yet, even if constructing and carrying informative interactions, i.e. interactions

[2] Geostatistics was actually also the historical motivation behind the development of Gaussian process regression methods, also known as *krigging* from the name of Daniel Krig, who used it to evaluate the gold resources in mines in South Africa (Krig 1951). His method was later formalized by Matheron (1962).

that afford maximal information gain, can decrease the number of interactions necessary to understand a phenomenon and disambiguate confounding evidence, in many situations, random interactions are only slightly suboptimal. In that context, rational experimentation, *a priori* requiring high cognitive resources, is not a particularly ecological behaviour. As Cook et al. (2011) points out (emphasis ours):

> *selective exploration of confounded evidence is advantageous even if children explore randomly (with no understanding of how to isolate variables): the more different actions children perform, the better their odds of generating informative data.*
>
> Cook et al. (2011, p. 352)

Here we find a major motivation for our work: behaviour that produces diversity is a key investigating tool in infants. Gweon et al. (2008) provided a study where infants presented with confounding evidence increased the variability of their exploration, even if that represented a physical effort. Schulz and Bonawitz (2007) and Bonawitz et al. (2012) reported similar results, where children preferentially engaged with a confounding toy, rather than to play with a new one.

Children seem to occupy an intermediary ground between random behaviour and rational experimentation, one or the other being favoured in function of the task, the difficulty, and the environmental and social conditions (Cook et al. 2011).

2.2 Self-organization

Self-organization is the property some systems have to self-organize, that is, to organize in such a way that the source of the organization is not found outside the system:

> *self-organization refers to exactly what is suggested: systems that appear to organize themselves without external direction, manipulation, or control*
> (Dempster 1998, p. 41)

> *Self-organisation is a dynamical and adaptive process where systems acquire and maintain structure themselves, without external control.*
> (Wolf et al. 2005, p. 7)

Let's take just one step away from the tautology with the definition of *organization* proposed by Ashby (1960, 1962). Given a system represented by a number of states S, and submitted to inputs I, the organization of the system is defined as the mapping f from $S \times I$ to S that describes the evolution of the state of the system in reaction to inputs. Under such a formalism, a self-organizing system does not change its organization, but specific inputs might move f into a different area of the state space where its behaviour is significantly different. This formalization is interesting, as it can be extended to express self-organization in learning systems, where the organization f changes to a new organization f' with each new input.

I will avoid the rabbit hole that a formal definition of self-organization entails—the question is not yet settled[3]—, and redirect the interested reader towards the contributions of Shalizi (2001), Wolf et al. (2005)[4] and Polani (2008) on this topic.

Self-organization can be found in many complex systems, at any scale: crystal formation (e.g. snowflakes, or ice in graphene nanocapillaries (Algara-Siller et al. 2015)), convection patterns, morphogenesis, ant foraging (Deneubourg et al. 1989), cerebral activity, school of fishes and flocks of birds, crowd (Helbing et al. 2005), sand dunes, financial markets, ecosystems (Arthur 1990), weather, planet rings, galaxy formations (Cen 2014).

Some characteristics are regularly found in self-organizing systems: they produce symmetry-breaking changes, self-amplification of—and resilience to—small disturbances, and dimensionally-reducing macroscopic effects (at least in the eye of the macroscopic observer) (Der et al. 2013; Der 2014). Many self-organizing systems are composed of many repeated elements interacting locally with one another, and subject to environmental pressure. In many of those systems, and in contrast with human engineering, no explicit design or intent can be found anywhere.

Der et al. (2012) propose the example of a uniform gas. Heated from the bottom, so that a sufficient gradient of temperature is created, it will display regular and stable convection patterns known as Bénard cells (Bénard 1901). Those cells break the symmetry that existed originally in the gas by amplifying small perturbations to create macroscopic patterns. At the same time, after an external, occasional, perturbation of an established Bénard cell pattern, the system will restabilize, possibly to different Bénard cells (i.e. the Bénard cell number can be different, or their location can change). Individual particles still obey the same laws of physics, but the organization of the system has moved to a different part of its state space, where its behaviour is significantly different. Still, no central organizational control is present, and the physical particles themselves are an integral part of the mechanism of organization.

A key—informal—insight to understand what self-organizing systems do is to consider that they tend to reduce the number of states the system can be found in. Whatever the initial state of the gas, if submitted to a gradient of temperature, it will eventually collapse into a convection pattern in the future: the system converges to a very specific region of the state-space, smaller than the set of state possible under current conditions. The same can be said about the snowflakes. Every snowflake is

[3] The controversy surrounding self-organization is such that Maturana, the father of autopoiesis, decided against using 'self-organization' entirely: '*I do not think that I should ever use the notion of self-organization [...]. Operationally it is impossible. That is, if the organization of a thing changes, the thing changes*' (Maturana 1987, p. 71). Incidentally, Ashby's definition provides a solution to his point.

[4] Of interest to the reader, Shalizi (2001) and Wolf et al. (2005) provide compelling arguments for a distinction between emergence and self-organization.

different[5]. But the set of possible states that ice can be found not being a snowflake is much larger; the self-organizing system of an icy cloud guarantees that out of random icy droplets flying around, snowflakes will be produced.

Another way to formulate this is to say that many self-organizing systems create *attractors*. Take the example of a ball dropped into a bowl. After a given time, as assuming that the friction of the ball rolling in the bowl in not null, the ball will end up at rest at the bottom of the bowl. However the ball is dropped, the future state of the ball is at the bottom of the bowl, at rest. This may not seem an example of self-organization. But the same gravitational force which is at play in the bowl example is responsible in much the same way for the creation of planets, planet rings and galaxies (Cen 2014), which are highly regular structures evolved out of the amorphous quark-gluon plasma of the early universe.

This gives self-organizing systems both an important sensitivity to environmental conditions and a resilience to them; the temperature and humidity conditions and their fluctuation during snowflake formation will greatly impact the geometry of the snowflake, but the final shape be hexagonal and symmetrical regardless. Likewise, the sand dunes will erase the traces of the caravan's passage, and the school of fishes will re-form after being traversed by a predator. As Wolf et al. (2005)'s definition stresses, a self-organizing system maintains its structure.

Biology is a heavy user of self-organization (Camazine 2003), in particular in the case of morphogenesis: the genetic code, when expressed, coordinates the execution of self-organization processes into biological organisms. In other words, evolution is weaving organisms with self-organization processes (see for instance Eggenberger Hotz (2003)). Evolution navigates the fitness landscape (or *adaptive landscape*, Wright (1932)) by trying combinations (sexual reproduction) of ever-so-slightly modified (random mutation) self-organization processes, encoded in the genetic code. This is hardly surprising: rather than having to explicitly specify in the genetic code where every brain cells should be placed, evolution only has to pick a self-organizing process whose characteristic is to converge towards producing a specific type of structural

[5] Some colleagues have questioned the use of the cliché. So. There are of the order of 10^{18} molecules of water in a snowflake. In those water molecules, 1 out of 3210 will have a *deuterium* atom—an isotope of hydrogen, ^{2}H, that occurs in a proportion of 1 for each 6420 hydrogen atoms—instead of a *protium* atom (the much more common ^{1}H isotope), thereby forming a semiheavy water molecule $^{2}H^{1}H^{16}O$. Following the same logic, 1 in 41216400 water molecules will have both, forming a heavy water molecule $^{2}H_2{}^{16}O$. Conversely, oxygen-18 and oxygen-17 are present in proportions of 1 in 500 and 1 in 2638 water molecules respectively. Which means that out of 10^{18} water molecules, there are 3.11×10^{14}, 2.43×10^{10}, 3.79×10^{14}, and 2.00×10^{15} molecules of $^{2}H^{1}H^{16}O$, $^{2}H_2{}^{16}O$, $^{1}H_2{}^{17}O$, and $^{1}H_2{}^{18}O$ respectively. By computing the binomial coefficients (using De Moivre's approximation of the factorial (De Moivre 1733; Pearson 1924, p. 403)), we can estimate the number of different possibilities of isotope distributions for the same snowflake geometry to $10^{2.06\times10^{16}}$. This is considerably higher than the estimated number of snowflakes that fall on Earth each year (6.6×10^{28}, see Pilipski et al. (2006)), or since its creation, 4.5 billion years ago (4.3×10^{39}). We'll leave a more precise calculation that takes into account tritiated water *and* the diverse combinations of hydrogen and oxygen isotopes *and* diverse isotopic fractionation phenomena (see for instance Jouzel et al. (1984)) *and* all the other planets where it snows in the universe as an exercise. Of course, the reader may question the relevance of distinguishing snowflakes by their isotope distribution. For geometrical differences, Pilipski et al. (2006) provide an analysis. And for the reader that questions the utility of the whole exercise *regardless*, I would point out that it provides the background for one of the only examples of *compositional diversity* (diversity by composition of common parts, here, isotopes) of this thesis. Earth isotope abundance data from the Commission on Isotopic Abundances and Atomic Weights, Berglund et al. (2011).

organization. That is, the specific location of each and every neuron in a group of neurons is not specified in the genetic code, but the structure that the group must respect is, in a constructivist way, through a self-organizing process. Using self-organization, evolution benefits from self-organization's resilience to perturbations—the brain structure will form reliably under a vast number of conditions—, while taking advantage of its sensitivity to external conditions, which preserves variability into the final structure, and thus maintains phenotypic diversity.

Self-organization is an underlying notion of much of the concepts of development (Thelen et al. 2007), and this would motivate by itself the exposition that it is given here. But self-organization pertains to our point in this thesis because, as just underscored, it creates constrained—or rather *functional*—diversity, without intent. Self-organization create contexts where a set of structural constraints are enforced, which allows the produced structure to play a predictable role in a larger system. But within the constraints, self-organization processes produce variability tied to environmental conditions.

Because of this variability, self-organization systems are also difficult to predict (Orrell 2007). Earth's weather system or the financial markets present important prediction challenges that still escape us in great part. And because of their complexity, they are difficult to simulate. As such, self-organization process present important challenges to any representation and predictive model of the environment.

Self-organization is also relevant to the discussion of a parsimonious approach to designing robots. Self-organizing processes related to morphogenesis offload much of the information necessary for the formation of an organism into the characteristics of the environment: the genetic code is not in itself a complete specification[6]. Biological organisms develop from a interaction between the genetic code and the environment; the genetic code is meaningless without the environment. The interested reader can advantageously consult the wonderful work of Oyama (2000) on this topic.

Self-organization is present in behaviour (Kelso 1995). It has been proposed as one of the fundamental mechanism involved in the acquisition of speech (Oudeyer 2006, 2013; Moulin-Frier, Nguyen et al. 2014), or, at a different scale, the cultural evolution of language (Steels 2012). Self-organization of behaviour emerged as a domain of investigation with the concept of homeostasis introduced by (Cannon 1932). In order to survive and eventually reproduce, a biological system needs to keep its physiological properties to acceptable levels. The first synthetic experiments on homeostasis were done by (Ashby 1940), who later introduced the term 'self-organization' (Ashby 1947).

Recently Der et al. (2012) introduced the concept of *homeokinesis*, that essentially turns homeostasis on its head. A system's behaviour driven by homeostatic principle

[6] Moreover, the genetic code would be nothing without the comprehensive cell machinery that caters to it, and that is transmitted to the offspring as much as the DNA strands: the cell machinery encodes also an important part of the information necessary for the morphogenesis.

will try to compensate for deviation from its equilibrium state. This leads to the system falling inert as soon as all needs are met. Homeokinesis, instead, directs behaviour by trying to investigate and reproduce deviations: such a behaviour is self-amplifying. A small perturbation can be magnified by a feedback loop that tries to reproduce it. This allows the system to escape local attractors and investigate different behaviours. Martius et al. (2013) showed that the behaviours created could be resistant to external perturbations.

The aims of homeokinesis are very close to ours concerning self-exploration. The goal is to propose mechanisms that can allow robots to discover their abilities and the possibilities offered by the environment themselves. Two important differences exist between homeokinesis and our work. First, they do not operate on the same systems: homeokinesis is primarily designed for low-level, high frequency sensorimotor loops while our algorithms—as presented here—target single-step, atomic, higher-level interactions with the environment. And second, one of the feature of homeokinesis is the high neural plasticity of its controllers; essentially, the discovery of a new behaviour erases the old one. The platform does not capitalize on its discoveries[7], and can be subject to behavioural loops. Our methods have a global view on explored behaviour, and strategies can compare newly discovered effects with all the previous ones, thus explicitly estimating and fostering *diversity*.

2.3 Sensorimotor Exploration in Fetal and Neonatal Development[8]

Neuronal activity influences neuronal development, in particular synapse formation and neuronal survival (Mennerick et al. 2000; Zito et al. 2002; Goda et al. 2003; Vanhoutte et al. 2003), even at a very early stage, before synapse formation. Decrease in activity translates into decreased neuronal proliferation, generally slower neuronal migration, and affect neuronal differentiation, driving in particular the proportion of excitatory versus inhibitory neurons (Spitzer 2006). As such, the importance of motor activations during the prenatal phase is fundamental for neural development.

In human infants, motor activity starts in the fetus at 9 weeks (Humphrey 1944), and at 10 weeks, translates into various arm and leg movements (Adolph et al. 2010). Some of those movements are coordinated between limbs, and some others are isolated, moving limbs or digits while the rest of the body remains still (Prechtl 1985; Prechtl and Hopkins 1986): this may allow a faster differentiation of the different

[7] The lack of long-term memory in homeokinesis is not inherent though, and could be added in any number of ways.

[8] This section is largely based on Adolph et al. (2010)'s and Hofsten (2004)'s account of prenatal and postnatal motor and perceptual development.

Box 2.1: The Price of Not Moving

While the reason why fetuses move is not yet fully understood, there is no question, however, that motion is a critical component of a normal morphological development. Moessinger (1983) performed an experiments on rats, where the fetuses were injected with curare during the last 3 days (out of 21) of gestation, artificially inducing *fetal akinesia*, a loss of voluntary movement. The curarized rats fetuses were born with underdeveloped lungs (pulmonary hypoplasia), shorter muscles leading to contracted joints (arthogryposis), undersized jaw (micrognathia), short umbilical cords, and an excess of amniotic fluid (polyhydramnios, possibly explained by reduced fetal swallowing), and generally reduced body, muscle and bone mass. A study by Rodríguez et al. (1992) confirmed those results, and further analysed the perturbation of the skeletal development. Moessinger (1983) remarks that, whereas curarized rats presented excess of amniotic fluid (polyhydramnios), lack of sufficient amniotic fluid (oligohydramnios) exhibits similar symptoms. This explanation is the same: lack of amniotic fluid leads to reduced opportunity of movements for the fetus.

The *Fetal Akinesia Deformation Sequence*, as it is referred to by the literature, is a striking example of the irreducibility of the organisms to the genetic code, as discussed in the previous section. Here, normal morphological growth is indissociable from behaviour: if the fetus does not move, it won't have a normal body. Interestingly, how much the mother moves is also a factor. Recent research in mice has suggested that increased exercise in mothers could reduce the risk of congenital heart defects in the fetus (Schulkey et al. 2015).

body parts in the brain. Those movements are not random. By the 14th week, two-third of hand motions are directed towards salient objects in the uterus: the fetus body, the umbilical cord, and the wall of the uterus (Sparling et al. 1999).

The specific reasons why fetuses move are still investigated. Some motions are provoked by external stimulation, others stem from self-simulation, while some appear to be just spontaneous. As Adolph et al. (2010) puts it, 'A primary reason why fetuses move is that they can' (see also Box 2.1).

At birth, the newborn undergoes a drastic environmental change. No longer supported by the amniotic fluid, the same movements require more strength. The dampening properties of the amniotic fluid are not present anymore, and stopping a movement becomes non-trivial. These two new environmental conditions taken together and combined with underpowered muscles ensure that neonates are restricted in their capacity to perform violent movements they could not control, and that could lead to injury.

The spontaneous movements of the fetus continue after birth, through short bursts of activity (Thelen 1979, 1981a,b), that extend throughout the first year. At one year old, it is estimated that infants have undergone more than 100 000 bouts of activity, commonly referred as *body babbling*, each of them involving repetitive movements. One of those activities is the kicking of the legs, that seems to prepare stepping motions.

As a clear evidence that those motions involve learning, the uniformity of the repetition of the kicking motion improves throughout the first year (Kahrs 2012). The original proposed explanation, involving hard-wired pattern generators (Hilgard et al. 1945; McGraw 1945), is further weakened when one observes that movements adapt to the fetus rapid morphological changes during development (Robison et al. 2005).

Furthermore, recent studies have shown that those movements play an important role in the organization of the body map of the spinal cord and the somatosensory cortex (Braun et al. 2001; Milh et al. 2006; Granmo et al. 2008).

Efforts have been made to construct models and simulations of those mechanisms. Simulation of the self-organization emerging from spontaneous movements have been conducted (Marques, Imtiaz et al. 2012; Marques, Völk et al. 2012). Yamada et al. (2013) and Sasaki et al. (2013) have used a simulation of a human fetus, along with its amniotic fluid and uterine wall to study the development of a spiking neural network and the establishment of the body map of the simulated fetus, driven by self-stimuli and stimuli from the environment. Lee (2011) has proposed a conceptual framework that attempts to explain how infants can switch from simple motor babbling to gradually more complex actions, and exhibit sophisticated play activity later in their development. Blumberg et al. (2013) argues for the value of twiching during sleep for sensorimotor development: the general muscle atonia of sleep would generate highly discriminable proprioceptive feedback, and propose a robotic model to study it.

The self-organization of the body map in infants is still the object of intense investigation. Our study of exploratory processes is in part motivated by this phenomenon, and how it could be effectively reproduced in robots: exploratory processes generate information that can fuel the self-organizing process, that could effectively bootstrap body and proximal environment discovery.

The work presented in this thesis, however, does not investigate this problem directly, and does not claim to bring any contribution to it either. The exploration process we investigate are considered in ecological contexts that have no resemblance to the developing body of a child.

Goal Babbling

In chapter 0, we compared two strategies, motor and goal babbling for a reaching task. Even as goal babbling has been shown to be superior, one can wonder about the ecology of a goal babbling strategy: it needs a coordination between motor action and sensory perception, needs for the consequences of actions to be observable, and observed, and needs the causal link between motor and sensation to be established. These abilities will no doubt be present in young infants, but they seem sophisticated for neonates? Since we advocated goal babbling to learn from scratch, without any previous information[9], if such a strategy cannot be carried by a neonate, the biological justification of our model, even as simplified and contrived as it is, cannot be made.

Neonates, it turns out, are able to sophisticated goal-directed actions (Hofsten 2004). They are able to direct their attention towards salient features and others' eyes (Haith 1980; Farroni et al. 2002), and are able to extend their arm towards their gaze (Hofsten 1982). Moreover, infants will tend to create sources of coordinated sensorimotor observations by putting their moving hand in their field of vision (Meer et al.

[9] A premise that is unreasonable anyway. The neonate does benefit from all its fetal sensorimotor experience.

1995), placing them specifically at the location of a spotlight when the rest of the field of view is dark (Meer 1997). At 3 months, infants are capable to form goal-based representation about the manipulation of object (Sommerville et al. 2005). The representation of actions, in infants and adult, seems to be goal directed: an action is perceived as the same if the goal remains identical, even if the method differs (Hofsten 2004).

In computational settings, the works of Baranes and Oudeyer (2010), Rolf et al. (2010), Jamone et al. (2011) and Hervouet et al. (2013) have demonstrated the benefits of goal-directed exploration over motor babbling in several contexts. And although Lee (2011) has termed his approach 'goal-free motor babbling', it in fact mixes goal babbling and motor babbling.

2.4 Psychological Studies of Exploratory Behaviour

The earliest studies of exploratory behaviour were done on the novelty-seeking behaviour of rats at the beginning of the 20th century (Small 1899; Slonaker 1912; Nissen 1930).

In 1937, Skinner, inspired by the works of Morgan (1894), Thorndike (1911) and Pavlov (1904, 1927) (and Watson (1913)), develops the theory of *operant conditioning*, which insists that all behaviour can be shaped by external rewards (Skinner 1938, 1957).

By 1943, the Hullian *theory of drive reduction* (Hull 1943), building on the homeostatic concept of (Cannon 1932), is established. According to it, behaviour can be explained by the need to reduce the tension of *physiological drives*. Drives can be distinguished between primary drives such as thirst, hunger, reproduction, sleep, fear, pain, and secondary drives that are not physiological, but learned from conditioning, such as money or citation counts. Drive can compete for control of behaviour: hunger can override sleep, or the opposite depending on the situation. This would account for the behavioural diversity displayed by animals.

A seductive aspect of the drive reduction theory is its mathematical formulation (Spence 1952). It offered for the first time a way for psychologists to compute behaviour, and it made the theory a prime target for *in silico* experiments.

However, the drive reduction theory (and the operant conditioning theory) fails to adequately explain exploratory behaviour. Exploration increases drives' tension without reducing the tension of any. In 1950, Harlow et al. (1950) observes that monkeys would play with mechanical puzzles for extended periods even when no reward was provided. Harlow argues that this playful behaviour could not be satisfactorily explained by any primary or secondary drive. He proposes a drive to manipulate, that belongs to another type of drive, *intrinsic* drives. His idea did not receive wide acceptance: the drive reduction theory was able to explain much of behaviour exhibited by animals and humans, with a mathematical formalism that Harlow's third drive was compromising. Harlow abandons his idea. At the same time, Montegomery starts a series of experiment on the exploratory behaviour of rats (Montgomery 1951a,b, 1952a,b), and propose an exploratory drive (Montgomery 1954; Montgomery and Segall 1955).

Since its publication in the 1930's, Piaget's research is slowly making its way over the Atlantic. Piaget insistence on exploratory behaviour driving cognitive development (Piaget et al. 1953) will have a profound impact on the study of exploratory behaviour.

Berlyne (1950, 1960, 1966) starts conducting experiments to explore the impact of novelty on behaviour in animals and humans, and postulates that:

> *When a novel stimulus affects an organism's receptors, there will occur a drive-stimulus-producing response (which we shall call 'curiosity')*
>
> Berlyne (1950, p. 73)

Berlyne proposes a curiosity drive, externally excited by stimulus conflicts, with a typical U-shape response to novelty: stimulus that are neither too novel nor too familiar arouse a maximal motivational response. Contemporary to Berlyne, Fowler attacks Berlyne's account of curiosity, noting that the stimulus producing behaviour is supposed to both evoke and satisfy the curiosity drive (Fowler 1965). Fowler proposes instead a boredom-based drive, which explains why investigative behaviour may be initiated before any relevant stimulus. Recent experiments have shown that boredom certainly plays a role in motivation: humans sometimes prefer negative outcomes (such as electric shocks) rather than doing nothing (Wilson et al. 2014)[10].

The same period sees other theories of motivation emerge. Festinger (1957) proposes a drive based on the reduction of cognitive dissonance: when presented with information that is not coherent with an individual's beliefs, he would experience discomfort and be motivated to reduce it. Kagan (1972) offers a similar view, building upon the work of Festinger, that formulated motivations as reduction of uncertainty, and recasts the work of Berlyne, Dember (1965), White, Hunt, (McClelland et al. 1953) and other in that perspective. The theory of Kagan and Festinger have been criticized for failing to explain why humans engage in activity that increase uncertainty; for Kagan, it is explained by a more cognitive form of uncertainty:

> *When behaviours seem to be aimed at increased levels of uncertainty, the more fundamental goal is often to resolve uncertainty surrounding an attribute of the self* (Kagan 1972, p. 60)

[10] Pascal called this one in 1662: 'j'ai dit souvent que tout le malheur des hommes vient d'une seule chose, qui est de ne savoir pas demeurer en repos dans une chambre.', which translates to: 'I have often said that all the sorrow of men came from one thing only, their inability to remain quietly at rest in a bedroom.' (Pascal 1662, Divertissement 186)

Dember and Earl (1957), Dember (1965), Walker (1964) and Hunt (1965) propose an alternative approach where humans do not strive to reduce uncertainty entirely, but rather to maintain an intermediate level of it, dubbed *optimal incongruity*. Amusingly, Kagan suggested that if one read the work of Walker (1964) and Dember (1965) while replacing every instance of 'optimal incongruity' by 'uncertainty', they would be great additions to his work[11].

More recently, Loewenstein (1994) proposed an iteration on these ideas: curiosity as a 'form of cognitively induced deprivation that arises from the perception of a gap in knowledge or understanding'. In other words curiosity is created by the difference between what the subject knowns and what it would like to know.

Contemporary of Berlyne, White (1959) attacks the Hullian theory of behaviour, and propose the concept of competence as a fundamental part of motivation. Similar ideas were formulated a decade before by Woodworth (1947, 1958). According to White, mastering a task would be motivating in itself, and not necessarily need external rewards—nor it would necessarily need for the task to be useful at reducing the drives' tension. deCharms (1968) proposes similar ideas, insisting on an *internal perceived locus of causality*, i.e., that the success at a task comes with the perception, from the subject, that the success is due to internal causes—that he, not an external event, is responsible for the success.

Twenty years after Harlow's experiments on monkeys, Edward Deci conducts similar ones on humans (Deci 1975; Deci and Ryan 1985). He uses a Soma puzzle, and ensures that participants are left innocently alone with it for several minutes. Not only participants play with the puzzle even where there was no objective reason too, but additional rewards manage to decrease the length of engagement in some cases. Not only the behaviour is not explained by the Hullian theory, but intrinsic motivation can decrease when the behaviour is also reinforced by a conventional reward. Intrinsic motivation do not necessarily entertain additive relation with primitive or secondary drives, hereby defeating attempts at claiming competence or novelty as just another drive[12,13].

Csikszentmihalyi introduced and documented the concept of *flow* (Csikszentmihalyi 1990; Csikszentmihalyi et al. 2005): human engagement is maximal when the task at hand is neither too complex or too easy, but matches the level of competence of the individual optimally. When in *flow*, subjects exhibit attention span far longer than the one observed in other situations. Csikszentmihalyi's work resonates strongly with White concept of competence White (1959). A similar concept in social learning is the *zone of proximal development* of Vygotsky (1978): the activities that are just hard enough that a learner cannot learn them on its own, but can if helped by a peer

[11] 'If one substitutes uncertainty for optimal complexity in the writings of Walker (1964) and Dember (1965), these positions become complementary to the one presented here.' (Kagan 1972, p. 57).

[12] Interestingly, an experiment by Festinger and Carlsmith (1959) on cognitive dissonance also showed that external rewards entertain complex, non-intuitive relations with motivation.

[13] One thing remains undisputed: (primary) drive tensions generally override exploratory behaviour (Cohen et al. 1968).

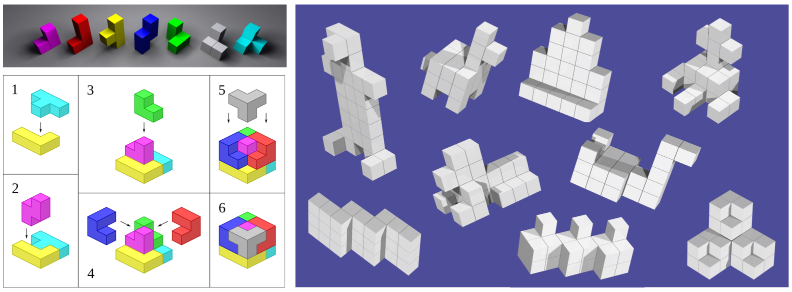


Figure 2.1: A way to assemble the pieces of the soma puzzle to form a cube. Many identifiable patterns can be produced. (figures by Dmitry Fomin, CC0 and fdecomite (modified), CC BY 2.0)

(usually a parent or teacher). Recently, some work has been proposed to use both concepts together (Basawapatna et al. 2013).

One of the most interesting aspect of intrinsic motivations is that they are highly dependent on the individual's experience and competence: they *stage* the acquisition of knowledge and skills by guiding the learning process towards learnable and/or new areas of the learning space. As such, they are an indispensable actor in the development of a self-sufficient agent.

Recent research corroborates intrinsic motivation theories. Kidd et al. (2012) has shown that children allocate their visual attention in order to maintain an intermediate level of complexity, avoiding sequence that are too simple (not enough information gain), or too complex (no possible or energy-expensive information gain). The results were reproduced in Kidd et al. (2014) for sequences of sounds: attention was contingent on intermediate complexity. Gerken, Balcomb et al. (2011) reported a similar result concerning learnability on 17-month old children. Two similar linguistic patterns were presented to the children, one learnable and the other not learnable (the result of a previous study, see Gerken, Wilson et al. (2005)). Children were shown to engage more with the learnable pattern; they avoided 'labouring in vain'. These studies confirm the theory that even from a young age, learning abilities influence exploration and attention.

In a recent study done on adults Baranes, Oudeyer and Gottlieb (2014) shows that when able to choose freely amongst an array of tasks (short video games), adults seek novelty and challenge. The concurrent presence of a motivation to increase competence and one to see new games shows that behaviour is the result of the interplay between *multiple intrinsic motivations*. Moreover, even as the task set featured unlearnable tasks, the range of tasks was sampled. This can be understood as a simple novelty seeking behaviour, or as a global diversity-seeking behaviour. Novelty-seeking is only interested in being presented with something not experienced before. Diversity-seeking implies a motivation to understand the range of variations that the task set offers. The experimental setup proposed by Baranes, Oudeyer and Gottlieb (2014) does not allow to disambiguate the two motivations.

Intrinsic Motivation in Animals

Recent studies of the motivations of animals have established interesting results. Wood-Gush, Vestergaard and Petersen (1990) and Wood-Gush and Vestergaard (1991) provide experiments that indicate that piglets seek novelty in their exploratory behaviour. The first experiment (Wood-Gush, Vestergaard and Petersen 1990) showed that after being confined in a bare pen, piglets spend more time examining a novel object introduced in their environment, than piglet confined in pens featuring straw, branches, logs, stone and creep feed, and therefore that offered richer interactions. In a second experiment (Wood-Gush and Vestergaard 1991), when offered the choice between entering two pens, one with a novel object and one with a familiar one, the piglet showed a strong preference for the pen with the novel object. Moreover, the novel object was linked with a significant increase in playful behaviour.

Interestingly, both experiments where criticized by Rushen (1993), arguing that methodological error did not allow to distinguish between an intrinsic exploratory motivation[14] and classical conditioned behaviour. Wood-Gush and Vestergaard (1993) offered a rebuttal, observing that even if environmental cue trigger the exploration, it does not explain the behaviour itself. Moreover, piglets, even if kept in pens where food is provided *ad libitum* show exploratory tendencies (similar observations were done on food-deprivation in rats not altering novelty-seeking behaviour, Hughes (1965)). And if the pens are bare and featureless—even when all physiological needs are satisfied—, abnormal behaviour is observed. Exploration seems to be not only intrinsic, but necessary, in a rich-enough environment. Although the Rushen response stays mainly technical, it also illustrates the confrontation of two schools of thoughts, where Rushen defends that all exploration is due to external stimuli.

Fear of novelty has often been invoked to argue that exploration could only be motivated by external stimuli: how to explain that animals would voluntarily engage in an activity that elevated stress levels while not providing any obvious reward? Experiments conducted by Misslin et al. (1986) measured the levels of corticosterone, a steroid hormone involved in stress response, in mice that were allowed to roam in familiar and novel environment. The results showed no physiological or behavioural indication of stress during the exploration of the novel environment. Stress was present, however, if the mice were prevented to return to the familiar environment, or were manually placed in the novel environment. Misslin et al. (1986) concludes that the ability to regulate one's own exploration is critical. This suggests that personal causation and control, which can be considered as motivations with the *internal perceived locus of causality* of deCharms (1968) and the competence drive of White (1959), also play important roles in emotional regulation during exploratory behaviour. This has led some psychologists (Duncan 1998; Poole 1998) to define exploratory behaviour, and more generally, the expression of intrinsic motivations, as a *behavioural* or

[14] Wood-Gush uses 'endogenous', instead of 'intrinsic', using a external/internal distinction rather than an extrinsic/intrinsic one. We will discuss the difference in the next section.

psychological need.

The study of novelty-seeking behaviour is not the prerogative of vertebrates: it has been reported in cockroaches (Darchen 1952) and bees (Lindauer 1952; Liang et al. 2012). This underscores the need for a systematic study of how cognitive abilities across species influence the diversity and complexity of intrinsic motivations. A recent study by Edwards et al. (2014) highlighted that capuchin monkeys did not engage with a learning task in the absence of an immediate reward: they were not intrinsically motivated by discovering causal knowledge. Moreover, capuchins learned better when the reward was present. Evidence in humans has shown opposite results (Kang et al. 2009): memory is increased when driven by intrinsic motivation. This suggests interesting venues of investigations for the developmental causes of the cognitive differences between humans and capuchins. As Edwards et al. (2014) points out (emphasis them):

> *Our results suggest that the gap between human and animal causal cognition is not merely a gap in competence, but is perhaps also a gap in motivation.*
>
> Edwards et al. (2014, p. 11)

Let's note that Watson et al. (1999) have observed instances of reward discarding in cynomulus monkeys, on a task with staged difficulty. Watson et al. (1999) conclude that in some instances, going to a higher difficulty seems more motivating than the food pellet resulting from finish the current level of difficulty. These two studies are insufficient to draw any definitive conclusions, but it points to two salient taxonomic units for further comparative studies: the catarrhines (Old World monkeys, to which cynomulus monkeys belong, and apes) and platyrrhines (New World monkeys, that include capuchins). Finally, Clark and Smith (2013) provide a study where chimpanzees would engage more with a cognitive task when food rewards were absent, highlighting a complex relation between reward and motivation.

A systematic study of animal capacity for intrinsic motivation is currently made difficult by the methodological difficulties of studying motivations in animals. Indeed, how to measure and discriminate between extrinsic and intrinsic motivations from animal behaviour remains an active challenge:

> *how best to measure exploratory behaviour in rodents remains a contentious issue.* (Brown and Nemes 2008, p. 442).

And the terms designating different motivations are often not precisely used:

> *Boredom, apathy and depression are often hypothesized to occur in animals housed in impoverished environments [...]. However, in very few cases has the use of these terms been validated empirically, and often no precise definitions are given.* (Meagher et al. 2012)

Others have stressed the current situation as well, and attempted to provide answers (Hughes 1997; Carter et al. 2012).

Another problem is that most studies will study behavioural traits for which there is a priori evidence of their presence. As Gosling et al. (1999) points out:

> *First, Antarctica will be discovered only if one sails south: The lack of evidence for a dimension does not necessarily prove the factor does not exist; studies may not have included the items relevant for the factor. To show that a dimension does not exist in a species requires that future researchers actively search for that dimension.*
>
> Gosling et al. (1999, p. 74)

Evolutionary Perspective

Another perspective that has seen a recent uptick is the evolutionary one. In the computational intrinsic motivation domain Singh, Lewis and Barto (2009) and Singh, Lewis, Barto and Sorg (2010) have brought forward very interesting ideas, suggesting mechanisms that could explain how intrinsic motivations originated from evolutionary processes—because, one way or another, they did (I go back in details about this in section 2.5). In biology, recent studies of animal behaviour have shown the heritability of exploratory behaviour (Dingemanse 2002), which has led studies to use rats selectively bred for novelty seeking behaviour (Stead et al. 2006; Ballaz 2009). Russell et al. (2010) performed a study of five males (normally bred) white rats individually introduced in the mammal-free, 9.5-hectare Motuhoropapa island in New Zealand, and tracked their position using GPS. Their conclusion was that the rats movements were apparently random, only mediated by a central place of foraging behaviour.

These studies stress the need to look at the relevance of intrinsic motivations in biological organisms beyond the individual level, at the species level. Intrinsic motivations generate exploratory behaviour. And exploratory behaviour directly contributes to the geographical dissemination of a species, which in turn, improves species survival. The argument can be made then that, even if intrinsic motivations can be detrimental for the survival of the individual, because it pushes it into unknown and uncertain—thus potential dangerous—situations[15], it can still be explained in a evolutionary perspective, because it increase the species geographical robustness (Russell 1983; Holway et al. 1999; Martin 2005; Taylor and Hastings 2005; Wright et al. 2010; Cote et al. 2010; Russell et al. 2010; Chapple et al. 2011, 2012; Liebl et al. 2012; Overveld et al. 2013) and reduces inbreeding.

Interestingly, exploratory behaviour has been studied experimental in large majority on rats. That can be explained by its convenient and heavy use across biological studies. But one also has to acknowledge that its capacity to disseminate and invade new ecosystems, matched by few, makes it a particularly favourable subject for those studies.

[15] In particular, exploration might push an individual in areas where few conspecifics are present. The reproduction rate of species in areas of low conspecific density is complex, and can be negative. This phenomenon has been dubbed the *Allee effect* (Allee 1931; Stephens et al. 1999; Taylor and Hastings 2005).

Of course, robots are not generally subject to these issues of species survival. Yet, as Merrick (2012, p. 231) remarked, current agents and robots have few types of motivation (see section 2.5), and few behaviours. But the trend is towards an increase in both, in the context of social exchanges with peers, and in particular in the context of populations of agents (Sequeira 2013) and swarm robotics. For those domains of investigation, the impact of intrinsic motivations on the population overall behaviour has potentially substantial consequences.

The consequences for individual robots are reversed, but are nonetheless important to consider: intrinsic motivations based on psychological accounts of behaviour in humans and animals do not necessarily lead to behaviour that is best at the individual level. Consequently, their transcription onto robots, where individual efficacy is sought most of the time, must be done with appropriate awareness of those factors.

2.5 Computational Intrinsic Motivation

The field of computational intrinsic motivation is situated at the confluence of psychology, active learning, and the design of embodied and disembodied agents.

Reinforcement learning (Sutton 1998) is a learning framework where the goal of an agent is to maximize a reward signal. Reinforcement learning has been successful in robotics, in particular because it offers a naturally incremental, online learning framework. However creating reward functions for complex tasks has proven to be difficult and frustrating, often relying on manual tuning from experts. Exhaustive analysis of small problems has shown that the best reward functions can be counterintuitive (Singh, Lewis and Barto 2009; Singh, Lewis, Barto and Sorg 2010). Moreover, static rewards signals and stable environments extinguish the acquisition of new knowledge and skills over time: once the task is learned, learning stops.

Spurred by the advances of the psychological models, intrinsic motivation systems were proposed as environment-agnostic motivational systems conductive of open-ended learning. Intrinsic motivation allows the agent to structure its learning trajectory by itself while complying with the environmental constraints, and prevents the extinction of learning. Intrinsic motivation drives are now recognized as a fundamental component of any self-sufficient robotic and biological system learning and exploring in an uncertain environment (Gottlieb et al. 2013).

Many different artificial curiosity drives have been proposed.

Schmidhuber proposed the compression driven progress measure (Schmidhuber 1991, 1990, 2009, 2010), basing the drive on how well a prediction module could compress the sensorimotor data the agent receives. In 2004, *intrinsically-motivated reinforcement learning* (Barto, Singh et al. 2004; Singh, Barto et al. 2005; Stout et al. 2005) was introduced, that computes the reward inside the agent, using objective feedback from the environment, with novelty as a possible drive.

At the same time, in the context of developmental robotics, Oudeyer (2004) and Oudeyer, Kaplan and Hafner (2007) propose the *intelligent adaptive curiosity* (IAC) algorithm, that partitions the sensorimotor space and seeks regions where the derivative the performance in prediction is maximal. Robustness improvements are later made by Baranes and Oudeyer (2009). Lee, Walker et al. (2009) propose a variant of IAC, called *category-based intrinsic motivation* that creates regions using *growing-neural gas* (Fritzke 1995) to create regions.

IAC is limited to low-dimensional problems, because the partition is done over the whole sensorimotor space. Baranes and Oudeyer (2010) proposed an new algorithm, SAGG-RIAC, that only partitions the sensory space and guides learning by choosing interesting areas of the sensory space: SAGG-RIAC is a goal babbling strategy, and was demonstrated on an 30-dimension motor space. Hervouet et al. (2012) later proposed improvements on SAGG-RIAC.

The idea of using learning progress to drive exploration was adapted to model-based reinforcement learning by Lopes, Lang et al. (2012), using cross-validation error to measure the evolution of model accuracy.

Other approaches include *empowerment* (Klyubin et al. 2005a,b, 2008; Salge et al. 2014b,a), where the agent is motivated to maximize its control over the environment. It is based on information theory; the agent maximized the channel capacity from the motors to the sensors. Another information theoretic approach is offered by (Ay et al. 2008; Martius et al. 2013), where the agent is motivated by maximizing *predictive information*. Maximizing predictive information encourages the robot to diversify his behaviour as much as possible, while keeping it predictable: this is the production of diversity constrained by the learning capabilities of the agent. Friston et al. (2010) argues that organisms are motivated by the minimization of the free energy, and that it elicits active sampling. Let's remark here that the free energy, also called *excess entropy*, is linked to the notion of self-organization (see Polani (2008, p. 28)).

This review is far from exhaustive, see (Oudeyer, Kaplan and Hafner 2007; Baldassarre and Mirolli 2013) for detailed surveys.

Extrinsic versus Intrinsic Motivations

The precise definition of intrinsic motivation, and its difference from *extrinsic* motivations has been the subject of debate (Baldassarre 2011). The internal/external dichotomy has been rejected. While there is consensus that intrinsic motivations originate in the agent, typical extrinsic motivations such as hunger generate stimulations that are expressed through proxies in the brain: all motivational signals are created by the brain (Baldassarre 2011, p. 2). Singh, Lewis, Barto and Sorg (2010) and Barto (2012, pp. 36–40) argue that extrinsic and intrinsic motivation form a continuum in biological systems, since they were created through a gradual evolutionary process. A computational account of this hypothesis is given by Singh, Lewis and Barto (2009) and Singh, Lewis, Barto and Sorg (2010), who proposed the notion of an *optimal reward function*. Given a fitness function and the distribution of environments, a reward function is evaluated in function of the expected fitness it generates across the distribution of environments. The optimal reward function is the one generating the highest expected fitness. It is a method that creates good rewards signals, robust to environmental variation, and that produces behaviours generating high fitness. Singh, Lewis, Barto and Sorg (2010, p. 12) argue that this framework provides a plausible explanation for the formation of both intrinsic and extrinsic motivations:

> *the difference between intrinsic and extrinsic motivation is one of degree—— there are no hard and fast features that distinguish them. A stimulus or activity comes to elicit reward to the extent that it helps the agent attain evolutionary success based on whatever the agent does to translate primary reward to learned secondary reward, and through that to behavior during its lifetime. What we call intrinsically rewarding stimuli or activities are those that bear only a distal relationship to evolutionary success. Extrinsically rewarding stimuli or events, on the other hand, are those that have a more immediate and direct relationship to evolutionary success.*

A contrario, Oudeyer and Kaplan (2008) propose an explicit definition:

> *An activity or an experienced situation, be it physical or imaginary, is intrinsically motivating for an autonomous entity if its interest depends primarily on the collation or comparison of information from different stimuli and independently of their semantics, whether they be physical or imaginary stimuli (i.e. measured by physical sensors or by internal "software" sensors) perceived in the present or in the past (in which case they will typically be internally represented and compressed by the brain through learning) or stimuli that are simultaneously present in different parts of one stimulus field.*
>
> Oudeyer and Kaplan (2008, p. 3)

While we will refrain from weighting in (too much) in the debate, it is important to understand the properties that are largely shared by intrinsic motivations: they tend to generates tensions that are *functionally dependent on the experience of the agent*, while this is not the usual case for extrinsic motivations.

Let's take the example of hunger, an extrinsic motivation. Hunger does not depend on the experience of the subject. It depends on his recent past—how long ago and how much did he eat. But it does not depend on the knowledge of the agent. Or on his skills. Without food, the subject will experience hunger and then starvation the same way.

On the other hand, curiosity is driven by the relationship between the environment and the agent's knowledge and skills (Ryan et al. (2000, p. 56): 'intrinsic motivation exists in the relation between individuals and activities'). And because those two are dependent on experience, it makes curiosity dependent on experience. As (Baldassarre 2011, p. 4) remarks, intrinsic motivation signals are characteristically transient in humans: they disappear and decrease as soon as the skill has been learned or the knowledge acquired.

It is important to note that we discriminate the motivation by the *tension* they create. One could argue that hunger generates experience-specific tensions. For instance, Given one's preferences, one may be hungry for a salad but not for a steak. While this is true, eating one or the other, however displeasing, will relieve hunger. More generally, whatever substance has relieved hunger in the past will do so in the present, even if preferences can change, in particular because of habituation mechanisms. The same cannot be said of curiosity: telling someone something he already knows will not satisfy his curiosity, even if it did in the past, when he had not had knowledge of it yet. Intrinsic motivations' *tensions* are usually dependent on experiences. Yet, not all are. Oudeyer, Kaplan and Hafner (2007) propose, for instance, several morphological intrinsic motivations that are not dependent on experience.

A point that we must address is the one of secondary drives, which are learned from conditioning from primary drives. As such, they seem to make non-intrinsic tensions dependent on experience. However, this is the drive that is dependent on experience, not the tension. The way to relieve the tension of those secondary drive is still a change of state of the agent, not a change of experience. A child can have internalized social pressure from his parents to have good grades, and this is the only thing motivating him to study. This is an extrinsic motivation. The way to relieve the social pressure is to consistently do his homework, listen in class, and obtain a good grade report he can present to its parents. Moral pressure and risk-taking notwithstanding, the child could find that not studying and cheating his way to exams or forging the report card would be an equally good solution to the pressure. The situation is solved by a change of state. A child that is intrinsically motivated to study might still feel social pressure, and cheat or forge his report card, but he will study nonetheless, because his motivation can only be satisfied by an acquisition of knowledge.

The distinction does not depend on the agent behavioural success. It does not depend either on the biological mechanisms underlying the motivations. It applies to humans, robots, as well as artificial agents.

In that context, it is easy to see why intrinsic motivations are suited for cumulative learning: they adapt to the accumulation of experiences, and lead the agent towards continuously acquiring novel information and avoiding stationary or repetitive behaviour. Intrinsic motivations produce structured exploratory behaviour.

Finally, let's remark that distinguishing motivations by how experience is involved is probably interesting and useful, regardless of how well it separates extrinsic and intrinsic motivations. Its applicability to computational intrinsic motivation in particular make it useful in the analysis of the contribution of different motivations to behaviour in heterogeneous motivational architectures.

Novelty versus Surprise

In our interest for exploration and the production of diversity, using a motivational drive driven by novelty seems the most straightforward choice. Novelty, here, is defined as 'different from anything known before'.

Novelty-based intrinsic motivation directs the agent behaviour to seek out stimulus that have not been observed before. Novelty is different from surprise: an effect can be novel without being surprising, because it has been correctly predicted (e.g. there is a new intern in the lab this morning, but I was told about it.). Conversely, an effect can be surprising but not novel, because it has not been correctly predicted but is familiar nonetheless (e.g. I am surprised to see my colleague at work this morning, I thought he was ill.). Surprise depends on the internal model of the agent, novelty only depends on its past history. Surprise is related to prediction error, and mediated by the confidence in the prediction: a wrong prediction with low confidence will generate less surprise than one made with high confidence. There are numerous subtleties to the distinction. The interested reader is encouraged to consult Barto, Mirolli et al. (2013).

In neurosciences, the difference is still investigated. Novelty requires to compare current stimuli against long-term memory, and therefore has to involve the hippocampus (Kumaran et al. 2007; Otmakhova et al. 2012). The neural response to surprise (i.e. the *prediction error*, sometimes called *contextual novelty* (Ranganath et al. 2003)) in visual stimulus, on the other hand, would originate in the superior colliculus (a midbrain region often studied for its involvement in eye movement, but which has a much larger multisensory role in directed attention) (Redgrave and Gurney 2006; Redgrave, Gurney et al. 2012). The superior colliculus exhibits strong habituation characteristics (Rankin et al. 2009): stimulus can receive a maximal response from the colliculus even if they are not new, if the previous presentation happened far enough ago in the past. Still Lisman et al. (2005) reports results on the involvement of the hippocampus in surprise (expected versus unexpected conditioned stimuli) detection.

Many computational approaches have proposed surprise-based (or habituation-based) drives (Bolado-Gomez et al. 2013; Lee and Meng 2005; Meng et al. 2005; Huang and Weng 2002, 2004; Marshall et al. 2004), and many use the term *novelty* to describe their methods. While novelty and surprise can be perfectly overlapping in simplified environments—thus justifying using surprise-detection methods to identify novelty—, in most studies, the difference is rarely acknowledged or discussed, in particular as a limitation of the method's applicability to more complex environments where novelty and surprise become distinct[16].

Yet, our approach investigates processes that produce diversity. As such, a novelty—not surprise—intrinsic drive is of higher interest to us (this does not preclude success-

[16] The methods will remain applicable and useful, but cannot be characterized as novelty-driven motivation methods.

ful diversity-producing exploration strategies to use a predictive-based drive, however).

Computational Novelty

The detection of computational novelty is related to two other problems. The first is anomaly detection (Chandola et al. 2009), where a system is monitored for any behaviour that deviate from an established norm. The application are numerous in industrial plants, medical care and automated security monitoring. The assumptions made in the anomaly detection context are not usually compatible with cumulative learning found in open-ended robot platform: a training set of exclusively normal behaviour is provided, and usually learned in a batch fashion. The second problems is outlier detection (Hodge et al. 2004; Chandola et al. 2007), which overlaps significantly anomaly detection, with a significantly different starting assumption: outliers are already present in the data, and no 'clean' normal dataset exists.

Let's note that given a diversity measure, the novelty of a new piece of data can be quantified by the difference in diversity before and after the data has been acquired. As such, any diversity measure defined in section 1.4 defines an implicit novelty measure. We'll use such a technique in section 4.2

Few implementations of intrinsic motivations rely on novelty. This is due to multiple reasons. First, many existing learning algorithms can be understood as being already inherently driven by novelty: R-max (Brafman et al. 2003) is a reinforcement algorithm where all states are given initial maximal rewards estimations that drive the exploration optimistically towards states that are not familiar. This illustrates the second reason: many learning algorithms are considered and tested in environments where the complete task is learnable, i.e., all states can be visited. In such a context, novel tasks are explicit: novelty is supervised.

Markou et al. (2003a,b) have proposed an extensive two-part review of novelty detection, underscoring the two different approaches: statistical methods, and neural networks.

In neural approaches, Marsland et al. (2002) has introduced Growing-When-Required (GWR) neural networks, that create a new node when the activation level of the nearest node of a new stimulus is below a given threshold. This can adequately compute novelty. Furthermore, the network keeps track of the amount of training that each node has received, hence allowing a less binary form of novelty for rare stimuli. This has been used by Neto et al. (2005a, 2007b,a) for visual detection of novelty.

In statistical approaches, most methods modelize the distribution density of the existing data, and characterize an observation as novel if it belongs to a low density area. To accurately modelize the distribution density however, large quantity of data are usually required or assumptions must be made (such as Gaussian distributions), which reduces their flexibility. Neto et al. (2005b,a, 2007b) have used Incremental PCA (Artac et al. 2002) for novelty detection in visual attention (and compared it to the GWR method). PCA creates a more compact representation of the data. If the new data cannot be represented precisely enough by the current representation, it is considered novel. Incremental PCA modifies the representation with each observation, integrating the new data to represent it accurately (and, hence, memorizing it).

Computational novelty is currently mostly used for visual tasks. For our own purposes, using the immediate improvement in diversity using a threshold coverage measure (section 1.4) was sufficient.

2.6 Diversity in Evolutionary Robotics

Abstract · Encouraging population diversity during the evolutionary process is a recent solution to two major challenges of evolutionary robotics: the bootstrapping problem and the early convergence problem.

Evolutionary robotics (Stanley 2011; Doncieux, Bredeche et al. 2015) aims to design robot morphologies, neural architecture and behaviours using algorithms inspired from the natural selection, variation, and hereditary mechanisms of natural evolution. As a subfield of evolutionary algorithms, evolutionary robotics distinguishes itself by evaluating the robots' behaviours rather than directly evaluating their phenotype, i.e. their morphology or their controller. As a subfield of robotics, evolutionary robotics distinguishes itself by having a global approach to the design of robots and their controller (Mautner et al. 2000), in contrast with the engineering approach that tries to decompose the design into independent problems to ensure modularity.

Evolutionary robotics regularly faces—amongst others—two specific challenges.

When the fitness of all members of the first generation is identical (typically because no rewarded behaviour was exhibited), the selection process cannot provide any progress toward a solution, and the algorithm is stalled. This is the *bootstrapping problem* (Mouret et al. 2009a). The canonical solution is to create a staged fitness function (Gomez and Miikkulainen 1997; Urzelai et al. 1998; Kodjabachian et al. 1998)—a proposition akin to a developmental constraint. The fitness function initially rewards solutions to simple problems and is progressively made more challenging to eventually match the real task. A related method, *reward shaping* (Dorigo et al. 1994; Mataric 1994), is used in reinforcement learning. In practice, such approaches require to design problem-specific fitness functions.

The second challenge is *early convergence*: the evolutionary process becomes trapped into a local extremum (Goldberg 1987; Bongard and Hornby 2010). This is due in particular to the fitness function having to play two roles: defining the problem to solve and guiding the search for a solution. If the fitness function is not carefully designed, it may only fill one of those roles properly. Such a fitness function is called *deceptive* (Mouret et al. 2009b), and is a case of over-exploitation.

To provide a problem-agnostic solution to these two problems, it has been proposed to consider selections processes that encourage behavioural diversity[17] in the population of candidate solutions. This has been proposed first in the classical evolutionary algorithm domain (Goldberg 1987; Sareni et al. 1998), and recently adapted to evolutionary robotics (Trujillo et al. 2008; Lehman and Stanley 2008, 2011a; Risi et al. 2009; Gomez 2009; Mouret et al. 2009a; Mouret 2011; Mouret et al. 2012; Doncieux and Mouret 2010, 2014; Krcah 2010; Delarboulas et al. 2010). Those approaches modify the fitness function to account for diversity.

The most common modification is *fitness sharing* (Goldberg 1987; Holland 1992): solutions close to one another share, i.e. divide amongst themselves the fitness score, in the same way individuals from the same ecological niche compete for resources. This method has proven itself empirically and has recently been theoretically proved as beneficial for simple cases (Friedrich et al. 2008).

Lehman and Stanley (2008, 2011a) proposes to abandon objective completely and focus on searching for behavioural novelty alone. The method proposed by Lehman and Stanley (2011a) is a novelty search: new solutions are compared for similarity against the current population and an archive of notable exemplars. Such an approach is shown to significantly outperform an objective-based one in a maze walk task. The authors also argued that it fosters open-ended exploration: because there are only so many ways to act simply, the candidate population is progressively guided towards more complex behaviours. Delarboulas et al. (2010) uses a similar approach where offsprings are compared against their ancestors, instead of their peers. Mouret (2011) shown that just maintaining the current population behavioural diversity, without considering past populations was enough to get good results, and avoided a growing computational cost for fitness.

This strand of evolutionary robotics research reinforces the idea that diversity producing processes are crucial for developing complex behaviours.

Developmental and the evolutionary approaches that encourage diversity remain significantly different. They happen at different timescales. One is concerned with the diversity of the individual in a online, incremental way, while the other happens at the species level and usually operates in batch evaluations of a complete generation. The work of Delarboulas et al. (2010) recently bridged that gap, by proposing an architecture where the evolution of controller happens online, and is driven by an intrinsically-motivated fitness function encouraging diversity at the individual and historical level (controllers are compared to their ancestors). In the context of that research, Delarboulas et al. (2010, p. 342) perhaps best expressed one of the greatest asset of using an intrinsically motivated approach with robots:

[17] Note that ensuring *genetic diversity* is different, and quite straightforward, as it can be controlled explicitly (Nguyen and Wong 2003).

the robot is rewarded here for what it gets (a rich sensori-motor experience) and not for what it does (going fast and circling infrequently)

2.7 Diversity in Machine Learning

Interestingly, a pioneer work on behavioural diversity in robotics has largely gone unnoticed. Balch (1997) and Balch and Parker (2002) introduced the notion of behavioural diversity in robot teams.

Another domain where diversity measures have been used is swarm optimization (Kennedy et al. 1995; Shi et al. 1998), a global optimization technique. The diversity of a swarm of particle is linked to the quality of the optimization, and thus works studying and proposing diversity measure are numerous (Riget et al. 2002; Krink et al. 2002; Blackwell 2005; Olorunda et al. 2008; Shi et al. 2008; Wang and Han 2009; Cheng et al. 2013). Olorunda et al. (2008) in particular reviews the existing diversity measure used for quantifying swarm diversity, and propose to compute the radius, the diameter of the swarm, or the average distance around the swarm center, normalized or not by the swarm diameter. Also proposed is the swarm coherence, that exploits the velocity of each particle in the swarm. Except the last one, these measure could be straightforwardly applied to our case.

Interestingly, Yen et al. (2006) proposes a swarm optimization method were mutliple swarm are used on a problem with a high number of local minima, and consider exchanges of particles between the swarms during the optimization based on diversity. These ideas share similarity with the methods we will present in the second part.

Diversity is also used to create classifier ensembles (Brown, Wyatt et al. 2005; Tang et al. 2006; Hadjitodorov et al. 2006; Ulaş et al. 2009; Connolly et al. 2012; Krawczyk and Wozniak 2013; Krawczyk and Woźniak 2014; Özöğür-Akyüz et al. 2014). A diversity of classifiers, when also avoiding weak classifiers, has been show to improve accuracy. In that case, diversity is based on disagreement between the different classifiers (Kuncheva 2001; Kuncheva and Whitaker 2003).

Diversity has also been proposed as a regularization metric in the result of search engines (Agrawal et al. 2009). This is not surprising: if the most relevant results are all very similar, less relevant but different results, after a few examplars of the most relevant class are given, are better, since they widen the number of requests that are answered in response to a given query.

And recommender systems, that are used to propose movies—the Netflix prize (Bennett et al. 2007) having largely popularized the concept—, restaurants, research articles, or mates in online dating systems to users, are no exempt either. Researchers have learned that diversity in recommendations, while sacrificing some accuracy, significantly increased user satisfaction (Ziegler et al. 2005; Zhou et al. 2010; Vargas and Castells 2011; Vargas 2014; Alexandridis et al. 2015):

> *An accurate recommendation, however, is not necessarily a useful one: real value is found in the ability to suggest objects users would not readily discover for themselves, that is, in the novelty and diversity of recommendation.*
>
> Zhou et al. (2010, p. 4511)

Here we find the idea that humans will not necessarily be able to discover by themselves information that are relevant and interesting to them in the environment. A recommander system exists precisely because exploring a computerized database is not something humans are intuitively good at, and because the database hide most its information: it does not provide clues to find information incrementally.

In the real world, rational behaviour and deductive reasoning is not sufficient to find information is the environment that is not revealed by indicative clues that something is to be found.

This explains why exploratory behaviour is necessarily intrinsic: because *it cannot be extrinsic*, as significant information is present in the environment but its presence is not detectable. One cannot deduce that a toy giraffe squeaks when pressed from passive observation. The water temperature of a river is hardly betrayed by its appearence, this is why it often contrasts with expectations. Thus, humans must be optimistic about finding information in the environment, they cannot wait for an indication it is there: exploration must be motivated intrinsically.

And because humans have greater capacity to make sense of and use the information they discover in the environment through exploration, exploratory behaviour is more rewarding. It seems then natural that their intrinsic motivational system is stronger, more developed and more complex that some other animals, for instance capuchin monkeys (see section 2.4).

2.8 SLAM algorithms

The expression 'exploration in unknown environments', when used in the context of robotic research, usually designate mobile robots mapping their environment. Stachniss et al. (2003), for instance, present an approach that uses 'coverage maps', and that even uses 'optimal information gain' to decide which areas of the map to explore next. In first approximation, this research and sensorimotor exploration appear to be related.

The fundamental difference between sensorimotor exploration and mapping exploration is that mapping exploration assumes that the agent knows how to move about on the map. The challenge, then, is to map the entire space (for instance, if its an interior environment) as efficiently as possible, with the best possible accuracy. Another difference is that mapping typically only considers 2 or 3-dimensional environment, while in sensorimotor exploration the dimensionality of the sensory space can be arbitrary[18]

The underlying assumption that how to move in the space is not to be learned has led to highly specialized and efficient *SLAM* (Simultaneous-Localization-And-Mapping) algorithms (Smith and Cheeseman 1986; Smith, Self et al. 1990; Thrun 2005, pp. 309-485), which use techniques unfit for sensorimotor exploration.

Discussion

Exploration as a Multidisciplinary Subject, Ripe for Interdisciplinary Research

The pervasiveness of exploratory process and exploratory behaviour across a wide range of scientific field suggests an important potential for interdisciplinary communications and collaboration, as was noted by Gottlieb et al. (2013).

This review of the studies of exploratory behaviours should not be considered exhaustive in any way. The neuroscience account is under-represented (Kang et al. 2009; Düzel et al. 2010; Shohamy 2011; Jepma et al. 2012), as is the literature taking an information theoretic perspective. The study of attention has a major research strand on perceptual exploratory behaviour and information-seeking behaviour, which have many interactions with the theories of motivation (Gottlieb et al. 2013; Laucht et al. 2006; Nocera et al. 2014). We barely mentioned the relation to the creation of diversity through pretend play (Belsky et al. 1981), and only investigated playful behaviour from a specific perspective, in the child-as-scientist paradigm. Biological or artificial creativity (Saunders 2002; Barbot et al. 2012; Mántaras Badia 2013), divergent thinking (Kleibeuker et al. 2012), or even counterfactual thinking were not discussed in relation to exploratory behaviour, and the exposition to, and the production of diversity. We didn't discuss the exploratory behaviour of populations, for instance, how ant and termite self-organize exploration, in both sedentary and army ants types, or how tourists are motivated by novelty (Lee and Crompton 1992). Similarly, exploratory behaviour is present in human organizations (March 1991; Gupta et al. 2006).

[18] This thesis only features low-dimensional environments though.

Nonetheless, a common trend can be observed amongst psychology, intrinsic motivation and evolutionary approaches: the incentive or reward that encodes explicitly a specific objective in the environment is not necessarily the best way to induce an agent to reach that objective, and may even actively prevents it.

The Many Intrinsic Motivations: A Benchmark?

The sheer number of different explanations for intrinsic motivation in psychology, and the correspondingly numerous and diverse models that have been implemented in computational intrinsic motivation hints at the complexity of the issue, and, perhaps, at the relative subjectivity that has accompanied its study so far. Most approaches to motivation will showcase how they can explain or successfully produce specific interesting behaviour.

But the overall field lacks a systematic and a comparative approach. Intrinsic motivation Intrinsic motivations are rarely compared against each other over identical, controlled environments. Santucci et al. (2013) proposed a detailed comparison of knowledge-based versus competence-based approaches, but the task considered, a two dimensional 2-joint arm can hardly be considered complex enough to allow to extrapolate the results to realistic settings[19]. The field lacks a benchmark, a set of *diverse* environments that implementations can measure against. The work of Singh, Lewis, Barto and Sorg (2010) has shown that a set of environments could efficiently filter good motivational drives.

A benchmark would not only allow to compare implementations, but also highlight the strengths and weaknesses of each one, by comparing the performances of one implementation across environments. There are important fundamental and technical difficulties to testing different implementations on the same environments: different approaches make different assumptions and have different requirements. But the set of environments a strategy can be applied to should be considered as one more way to differentiate and characterize approaches. In evolutionary robotics, Lehman and Stanley (2008) introduced two environments to test the diversity approach. Those environments have been reused by Delarboulas et al. (2010) and Mouret (2011).

The goal is not to decide which intrinsic motivation measure is the best—as we highlighted in chapter 1, we have an evaluation problem. Furthermore, the diversity of the field is precisely suggesting that one may not be enough to explain the behaviour of humans (Hughes (1997): 'no single approach has adequate explanatory or predictive power'). Neuroscience tells us that different brain structures, the colliculus and the hippocampus amongst them, have been linked to the origin of intrinsic motivation signals, strongly suggesting that this diversity of intrinsic motivations is inherent—and probably cannot be escaped by a cleverer take on the issue.

[19] Moreover, as we illustrated chapter 0, a random motor babbling strategy provides adequate performances on a 2-joint arm.

Another issue is that as intrinsic motivations help steer the developmental process of infants, they are also naturally part of it: children motivations change as they develop Trevarthen et al. (2003). This can certainly be explained, in part, by the functional dependency of motivations with experiences, but assuming that this is sufficient is not a trivial assumption.

Of course, as happy as we are to provide advices, we'll blatantly ignore them in this thesis, as will be made explicit in the discussion of the next chapter.

Our work on the study of exploration and the production of diversity is not directed at explaining complex behaviour in humans, or to propose algorithms that can compete in terms of performance with the state of the art. Rather, it has been to find some of the most simple mechanisms of exploration, and to modify them every which ways in order to investigate their dynamics, and the relative impact of the submodules that compose them. This will be the focus of the next chapter.

Kids should be allowed to break stuff more often. That's a consequence of exploration. Exploration is what you do when you don't know what you're doing.

Neil deGrasse Tyson

3
Revisiting the Two-Dimensional Arm

In the example of chapter 0, we illustrated that on an idealized two-dimensional arm setup, a goal babbling strategy was able to discover significantly more of the reachable space than a motor babbling one. For the sake of brevity, many details were not investigated. We take a closer look at them now.

3.1 The Exploration Algorithm

Abstract · We formalize the motor and goal babbling algorithm discussed in chapter 0, and provide a quantitative analysis of it, using the diversity measures introduced in chapter 1.

We consider an environment $f : M \mapsto S$, as formalized section 1.3. For each sampling of f, the exploration algorithm does either a random motor babbling action—picks a random point $\mathbf{x}$ in the hyperrectangle M—, or a random goal babbling action, i.e. picks a random point in the bounded sensory space S as a goal for the inverse model, and infers an motor command $\mathbf{x}$ to execute.

In the following sections, we formalize the inverse model and the exploration algorithm.

3.1.1 Inverse Model

Given a goal, the inverse model we used in chapter 0 finds the nearest neighbour in the observed effects and applies a small perturbation on its corresponding motor command.

Formally, M is a closed hyperrectangle of R^m, and as such it is the Cartesian product of m closed intervals:

$$M = \prod_{m=0}^{m-1} [a_i, b_i]$$

Given a motor command $\mathbf{x} = \{x_0, x_1, ..., x_{m-1}\}$ in M, a perturbation of $\mathbf{x}$ is defined by:

$$\textsc{Perturb}_d(\mathbf{x}) = \{\mathit{random}(\mathit{max}(a_j, x_j - d(b_j - a_j)), \mathit{min}(x_j + d(b_j - a_j), b_j))\}_{0 \leq j < m}$$

with the function $\mathit{random}(a, b)$ drawing a random value in the interval $[a, b]$ according to a uniform distribution. d is the *perturbation parameter*, and the only parameter of the inverse model, that we can now express in Algorithm 1.

```
Algorithm 1: $\textsc{Inverse}_d(\mathbf{y}_g, E)$
Input   :
    • $d \in [0, 1]$, a perturbation ratio.
    • $E = \{(\mathbf{x}_t, \mathbf{y}_t)\}_{0 \leq t < N} \in (M \times S)^N$, past observations.
    • $\mathbf{y}_g \in S$, a goal.
Output:
    • $\mathbf{x}_e \in M$ a motor command.

Find $(\mathbf{x}_i, \mathbf{y}_i)$ in $E$ so that $\mathbf{y}_i$ is the nearest neighbour of $\mathbf{y}_g$ in $\{\mathbf{y}_t\}_{0 \leq t < N}$.
$\mathbf{x}_e = \textsc{Perturb}_d(\mathbf{x}_i)$
```

The inverse algorithm is simple, but effective. Its only assumption is that a small perturbation of the motor space produces a comparatively small changes in the sensory feedback. It does not extrapolate, nor does it interpolate observed data. The model is not sensitive to the distance of the goal from its nearest neighbour. Consequently, whole areas of the goal space are strictly equivalent for the inverse model. Additionally, the model has difficulties escaping attractors, and is susceptible to local minima, as illustrated by the arm loops in chapter 0.

Because of this, more powerful models such as Locally Linear Weighted Regression (LWLR) (Cleveland et al. 1988; Atkeson et al. 1997a,b) might obtain better results, in particular when goals are far from the observed data. We'll use such a model in the second part. Yet, in highly dimensional non-linear motor spaces, such models usually need a large amount of observations, concentrated in small neighbourhoods of the motor space to work well. This creates situations where more complex models are worse at exploring under a scarcity of data (for instance, newly discovered areas, or during the beginning of the exploration), and will reward exploring already well sampled areas, just because they are more effective on them.

In practice, for the experimental context we consider in this chapter, the performance and robustness of our model is competitive. Additionally, our model generates precisely the kind of data distribution (concentrated clusters of motor vectors) that more complex forward and inverse models might take advantage of. Let's remark here that this inverse model is not completely unreasonable in biological organisms (Loeb 2012).

Furthermore this model is intuitive, allowing the reader to run the exploration algorithms in his head without abstracting the learning step. And it is computationally frugal, allowing to reproduce most of the experiments in minutes or seconds, thus ensuring that the interested reader can modify and play with the experiments presented in this chapter with minimal commitment.

3.1.2 Motor and Goal Babbling

The basic exploration strategy we will consider throughout this thesis is composed of two distinct phases: a motor babbling phase and a goal babbling phase. Although the implementation we distribute is modular, we present an equivalent monolithic formalization in Algorithm 2.

More complex exploration algorithms will be proposed in this chapter, but this strategy is simple and effective. The K_{boot} parameter articulates the balance between undirected exploration and directed exploration. Our objective is to set K_{boot} to reduce the duration of the random motor babbling phase as much as possible without significantly compromising performance. Let's note that this goal babbling strategy needs S to be bounded, and reasonable. We will address this problem in the section 3.2.

Henceforth, when referring to a random goal babbling strategy—or simply *goal babbling*—, and unless stated otherwise, we will be referring to the Explore algorithm with $K_{\text{boot}} = 10$.

3.1.3 Quantitative Analysis

For the two-dimensional arm environments, we will use, unless otherwise indicated, a Testset-based Average Distance measure introduced section 1.4, based on a lattice restriction to the unit disk, as pictured Figure 3.1. In this section however, we compare

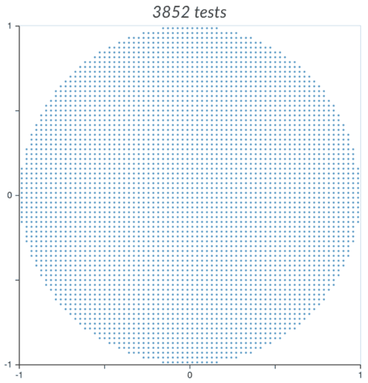


Figure 3.1: The lattice testset for the two dimensional arm can characterize how well the reachable space is covered. Here we rely on an approximation of the reachable space as the unity disk, allowing to use this testset for any two-dimensional arm. [source code]

```
Algorithm 2: EXPLORE($(f, n)$, $K_{\text{boot}}$)
Input:
    • $(f, n)$, environment.
    • $K_{\text{boot}}$, duration of random motor babbling.
Result:
    • $E = \{\mathbf{x}_i, \mathbf{y}_i\}_{0 \leq i \leq n} \in (S \times M)^n$, exploration trajectory.

$E \leftarrow []$
for $t$ from $0$ to $n - 1$ do
    if $t \leq K_{boot}$ then
        $\mathbf{x}_t$ = MOTORBABBLING($M$)
    else
        $\mathbf{x}_t$ = GOALBABBLING($S, E$)
    $\mathbf{y}_t \leftarrow f_A(\mathbf{x}_t)$ // execute the command
    add $(\mathbf{x}_t, \mathbf{y}_t)$ to $E$

MOTORBABBLING($M$)
    choose $\mathbf{x}_t$ randomly in $M$
    return $\mathbf{x}_t$

GOALBABBLING($S, E$)
    choose a goal $\mathbf{g}_t$ randomly in $S$
    $\mathbf{x}_t = $ INVERSE($\mathbf{g}_t, E$)
    return $\mathbf{x}_t$
```

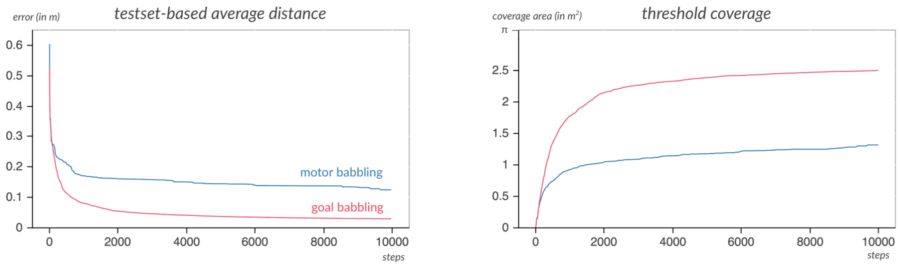


Figure 3.2: Comparison of exploration performances. The experiments are the same as Figure 1.9, over 10000 timesteps. For the coverage performance, $\tau = 0.05$. [source code]

the Testset-based Average Distance to the Threshold Coverage measure (section 1.4), that will be used in all the second part.

The reason for using a testset-based diversity measure is that the reachable space is well defined for the two-dimensional arm environments, and that, as mentioned previously, it is compatible with a learning performance interpretation. In the second part, the reachable space is more difficult to assess, and we use the Threshold Coverage measure because it is more robust.

The experiments are run in the same conditions as chapter 0, on a 20-joint arm. K_{boot} is set to 10, $d = 0.05$, and 10000 steps are run. For the threshold coverage

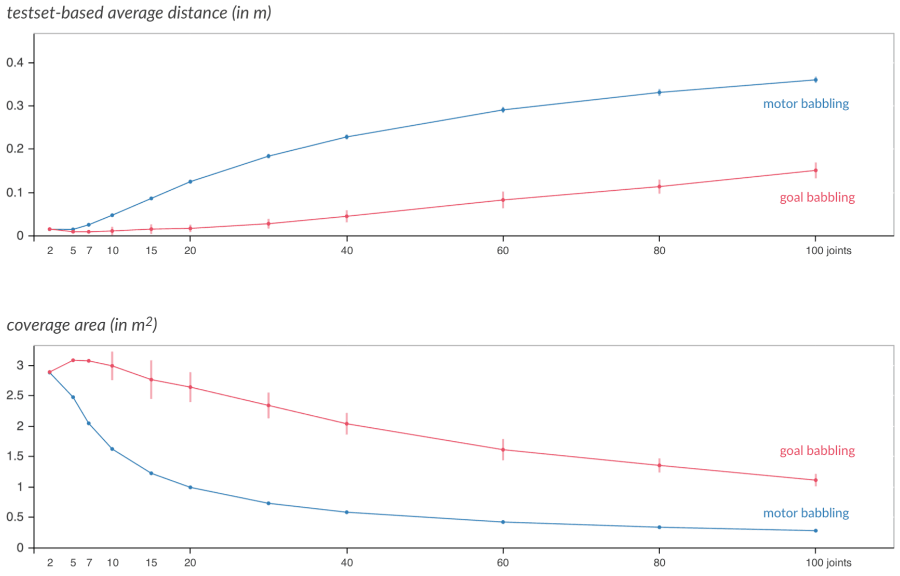


Figure 3.3: Goal babbling is a better strategy when many joints are involved. Performances are shown at the end of the exploration (t = 10000), and experiments are repeated 25 times. Interestingly, in the case of goal babbling, a sharp increase in standard deviation can be observed at dimension 10; this is caused by the looping of the arm in some experiments and not others, generating increased variability in performance. [source code]

measure, $\tau = 0.05$. In Figure 3.2, the two diversity measures are compared on a single experiment[1]. Both show the benefits of the goal babbling strategy over the random motor babbling strategy. The testset measure is more sensitive to the slightly more stochastic performance of the early motor babbling exploration.

In Figure 3.3, the performance of the two strategies are in function of the number of joints of the arms. Both the ratio between the random goal babbling and motor babbling coverage areas and the difference of the average distances stabilize after 40-joint s. Both measures are sensitive to the increase in variability (due to the arm loops, chapter 0) from one run to another after 10-joint for the goal babbling strategy. As such, the two measures convey similar information.

[1] In all performance graphs of this thesis, the diversity measure has been computed for timesteps 1, 2, 3, 4, 5, 10, 15, 20, 25, 50, 75, 100, 125, etc.

3.2 The Distribution of Goals

Abstract · We show that the goal distribution markedly impacts the distribution of effects, which create challenges and opportunities when guiding exploration.

In the setup of chapter 0, the goal space consisted of the axis aligned bounding box of the reachable space[2]. If this information was included in the definition of the problem, that would be fine, because motor babbling makes no use of the information. Yet, knowing the bounding box of the reachable space is an unreasonable assumption in the general case.

That would still be fine if the distribution of goals did not significantly impact the distribution of effects. Alas, this is not the case. To prove this, we consider four different goal distributions besides the 2 meters by 2 meters (*fit* scenario) of chapter 0. Three are centred at the origin and of dimensions 1 m x 1 m (*half-size* scenario), 4 m x 4 m (*double size* scenario), and 10 m x 10 m (*ten times bigger* scenario) respectively. Another is off-centre (*corner case* scenario), and is 0.25 m x 0.25 m. All distributions are depicted in Figure 3.4, as well as the respective distribution of effects they induce on the 2-joint and 20-joint arm – over 10000 timesteps, using the goal babbling strategy of the previous section.

The distribution of goals radically impacts the distribution of effects. In the *half size* scenario, the effects stay concentrated in the centre of the reachable space, and do not reach its outer edge. Inversely, when the goal space is bigger than the reachable space, the effects concentrate on the boundary of the reachable space, to an extent that correlate with how big the goal space is, as the *double size* and *ten times bigger* scenarios illustrate[3].

This phenomenon is also observable, to a lesser extend, in the *fit* scenario of the 2-joint arm as well: the four corners of the goal space do not overlap with the reachable space, and we see increased effect density on the reachable space boundary in those corners. This pooling behaviour has been analysed and explained in chapter 0.

If the goals are concentrated in a small part of the reachable space, as in the *corner case* scenario, so are the effects.

These results show that when goals are drawn randomly, a bad estimation of the goal space can easily lead to a bad distribution of effects. Of course, they also illustrate the flexibility of goal babbling exploration: it can efficiently guide the exploration of the sensory space. If an area of the effect space is deemed more interesting than another, we can manipulate the distribution of goals to concentrate exploration in this area—as the *corner case* scenario illustrates—without changing the other mechanisms

[2] Actually, not exactly. Because of the angle constraints, no posture of the arm reaches a position where $y = -1$. But the imprecision is not significant for our argument here.

[3] A 100 m x 100 m goal space would not have produced a significantly different distribution than the 10 m x 10 m one, as the quasi-totality of the goals are outside the reachable space in both cases, and the inverse model, as it projects each goal to the nearest observed effect, is insensitive to how far the goal is from the effect.

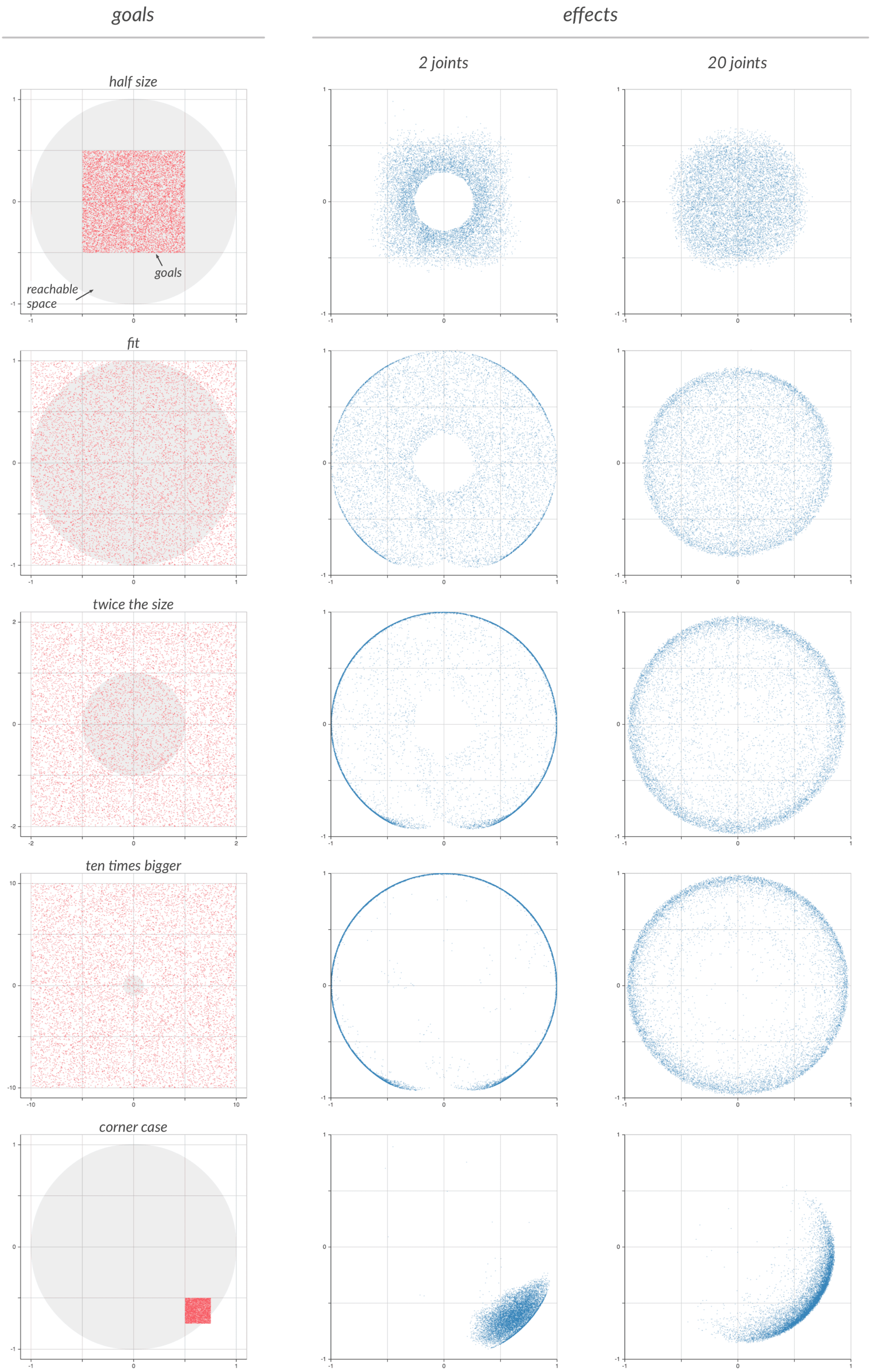


Figure 3.4: By manipulating the goal distribution, we can manipulate the distribution of effects. On a 2-joint and 20-joint arm, we compare five goal distributions (first column), some under-dimensioned and some over-dimensioned compared to the reachable space (grey disk). [source code]

of the exploration. This is the methods proposed by Oudeyer and Kaplan (2007, p. 8); Rolf et al. (2011), Jamone et al. (2011), Baranes and Oudeyer (2013) and Hervouet et al. (2013), where the exploration trajectory is guided by preferences over the goal space. Pushed to the extreme, i.e. considering only one goal, the exploration strategy seamlessly collapses into an optimization one. These characteristics make goal babbling easily interfaceable with an attention mechanism or an interest measure.

With exploration, one objective is to cover the reachable space in a homogeneous manner, producing exemplars of the possibility it offers, and for this the goal space must not be too dissimilar from the reachable space. Since we don't have access to the geometry of the reachable space, we have to estimate it from current observations. This is similar to the problem of density estimation (Rosenblatt 1956; Parzen 1962), where the density of an unknown distribution must be estimated from a discrete number of samples. Here, we are only interested in the *support* of the distribution, i.e. the subset of the space where the density is not null[4]. Furthermore, the sampling available to the agent is not independently and identically distributed, but function of the competence of the agent.

A simple approach to estimate the reachable space is to take the bounding box of the current observations. To make exploration more aggressive, the goal space could be expanded from the current boundaries of the estimation of the reachable by a factor superior to 1. The higher the factor, the more aggressive the exploration.

This approach assumes that the ratio of the bounding box volume to the reachable space volume is low, as is the case for the two-dimensional arm. But it is not efficient for sparse, non-contiguous reachable spaces. To robustly explore those spaces, we need a good estimation of the *reached space*.

[4] In practice, we can relax this by only considering the areas of the space where the density is above a small threshold.

3.3 Exploration on a Grid

Abstract · *We introduce grid partitioning, and an approximation of the reached space based on it which will be instrumental for several exploration algorithms. We describe explorers that define goals inside and outside of the reached space. We show that combining these explorers allows some independence from the dimensions of the goal space.*

Given a partition of the sensory space, we define the *estimated reached space* during exploration as the union of the elements of the partition that contain at least one observed effect. The quality of the exploration depends on how the sensory space is partitioned. In this thesis, we will use a simple, good-enough, computationally efficient partitioning scheme: *grid partitioning*.

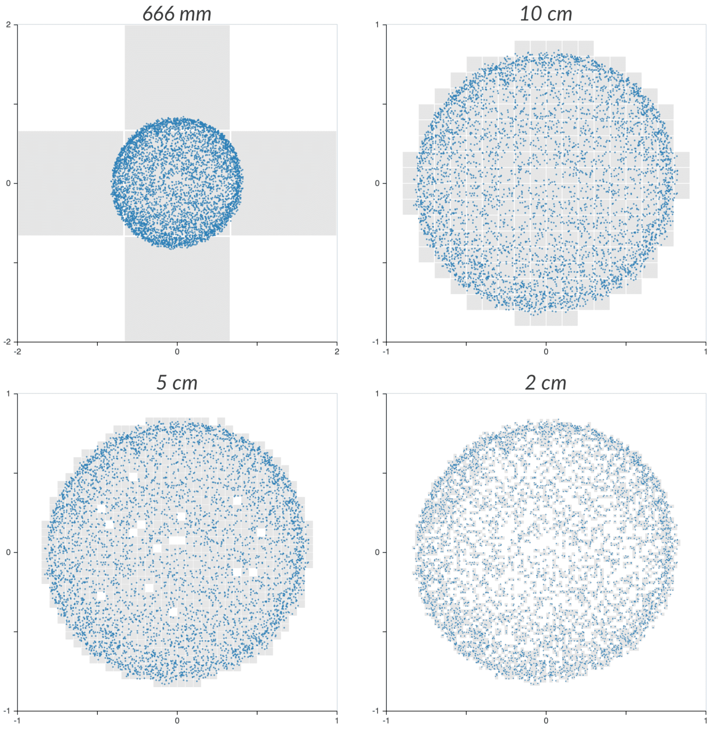


Figure 3.5: The size of the cells has a huge impact on the estimation of the reached space. This figure exhibits examples of grid partitioning that underfit (666 mm), fit (10 and 5 cm) and overfit (2 cm) on the same data. For the two-dimensional arm, we shall mostly use cell widths of 5 and 10 cm. [source code]

3.3.1 Grid Partitioning

Grid partitioning partitions the sensory space into axis-aligned hyperrectangles of identical dimensions, hereafter designated as *cells*, whose centres form a lattice over the sensory space.

Grid partitioning is parametrized by two vectors in $\mathbb{R}^n$: **b**, the coordinates of the centre of the cell which contains the origin of the reference frame—the grid's origin—, and **a**, the size of a cell. Given a point **x** in $\mathbb{R}^n$, the coordinate **c** (in $\mathbb{Z}^n$) of the cell that contains **x** is:

$$c_i = \left\lfloor \frac{x_i - b_i}{a_i} \right\rfloor$$

By varying the cell-size[5], we can obtain large cells which fit the reachable space loosely, or small cells which overfit the current observations, as Figure 3.5 illustrates.

The size of the cell effectively sets an implicit threshold for similarity and saliency.

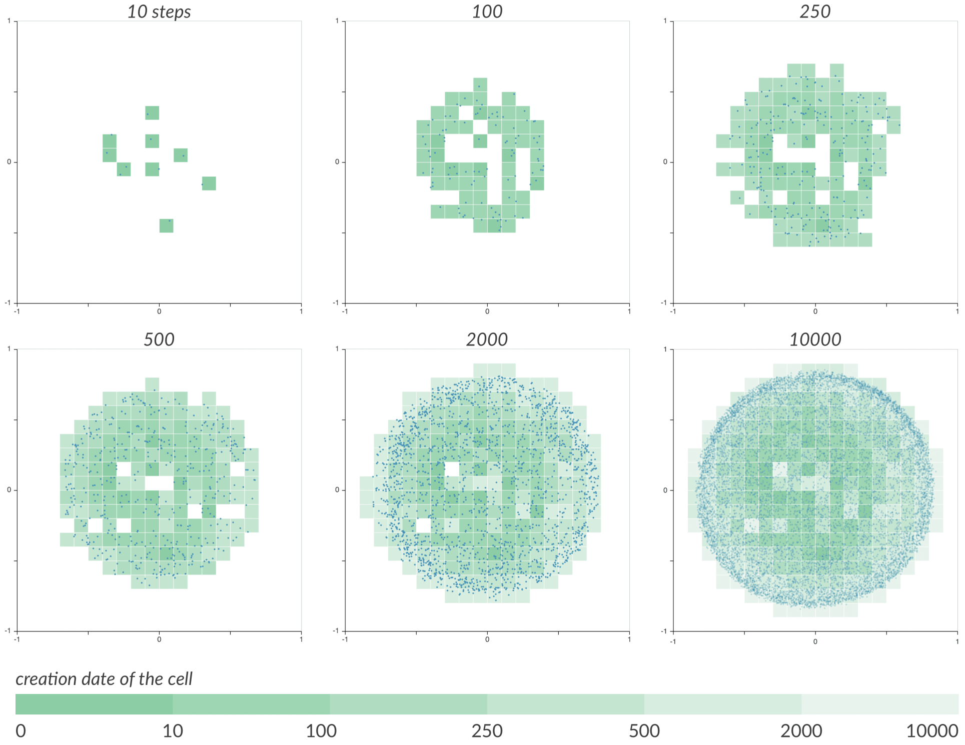


Figure 3.6: The reached space growth slows as exploration progresses. The colour of the cells indicates the time at which they were added to the reached space. Some regions of the reachable space enclaved in the reached space, that would have been discovered early by random motor babbling, are still not explored after 2000 timesteps by the goal babbling strategy. [source code]

[5] The origin of the grid usually has little consequence, although, with large cells, its importance increases, as it is apparent in the leftmost graph of Figure 3.5.

Two effects belonging to the same cells are considered identical with regards to what they tell us about the reachable space. An effect belonging to a new cell is salient, because it represents discovering a new area of the reachable space. Note that, how the grid's origin is defined can have unintended local consequences, since any two effects can be arbitrarily close, yet belong to two different cells, if the grid's origin is chosen appropriately[6].

There are many ways to set the size of the cells of the grid. One could bound the number of occupied cells, and enlarge the cells when needed. This has the advantage of offering the possibility to adaptively match the grid topology to the learning abilities or time resources available to the agent, who may not have the time or the capacity to handle a large number of cells. In this manuscript, the size of the cells is set arbitrarily, to avoid complications. For the arm example, we used cells of size 5 and 10 cm depending on the experiments.

Having now a grid partitioning method, we can estimate the reached space during exploration, as depicted Figure 3.6.

The idea of partitioning continuous sensory spaces for goal exploration has been explored in the context of the *SAGG-RIAC* algorithm (Baranes and Oudeyer 2010), which was derived from the *IAC* algorithm proposed by Oudeyer (2004) and Oudeyer, Kaplan and Hafner (2007) and later improved in more robust versions as *R-IAC* (Baranes and Oudeyer 2009) and *CBIM* (Lee, Walker et al. 2009). *IAC*, *R-IAC* and *CBIM* partition the sensorimotor space into regions. *SAGG-RIAC*, in contrast, only partitions the sensory space. It does so adaptively: regions where many effects are observed are split into smaller regions, in a way that optimizes the difference between the empirically-measured competence progress of the newly created regions. The hope is that it allows to discriminate efficiently between regions of different level of learnability. In practice, the regions it creates are sometimes difficult to explain and random splits would probably work equally well. To avoid unnecessary complexity, we opted for a simpler grid approach in this thesis.

3.3.2 Goals on a Grid

Having an estimation of the reached space allows us to define more complex exploration strategies.

The *reached* exploration strategy only selects goals in the current estimation of the reached space. More specifically, to choose a random goal, one chooses a random, non-empty, cell, and then draws a random point inside it.

Inversely, the *unreached* exploration strategy considers a finite subset of sensory space chooses a random goal amongst the empty cells belonging to this subset, if any

[6] There are ways to avoid those borderline effects, such as making the cells partially overlapping. The added complexity did not seem worth it in the context of the algorithms presented. And we can run in exponential trouble in high dimensions if not done carefully.

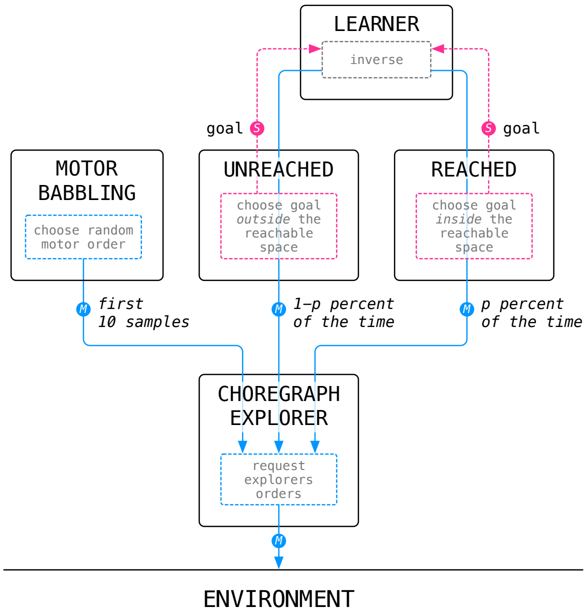


Figure 3.7: Three explorers are combined to form the p-reach strategy.

exists[7]. This exploration strategy drives the exploration towards unexplored regions of the sensory space. Here, we will usually consider the subset of cells as the set of cells contained in an hyperrectangle containing the bounding box of the observed effects.

To illustrate how those two strategies can be employed together, let's consider a

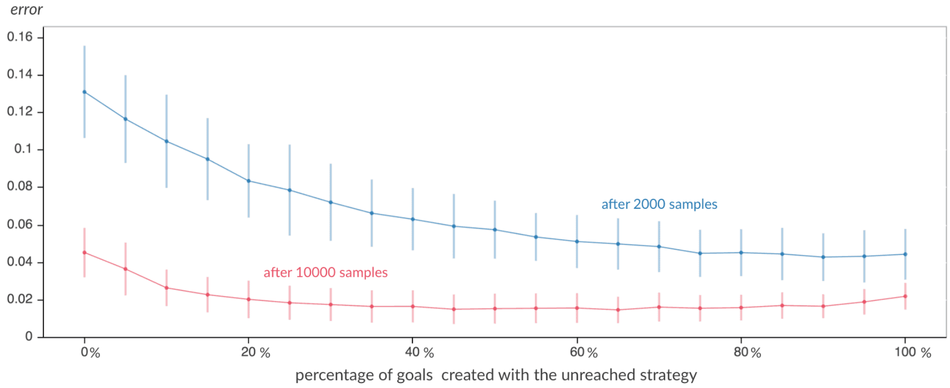


Figure 3.8: The more aggressive the exploration, the better it will do early on, but pursuing the same strategy will be detrimental in the long term. Here we see the error rate of a 20-joint arm. Averaged over 50 runs. [source code]

[7] In our experiments, some cells are always empty because they are unreachable. In our implementation, the strategy defaults to a predefined strategy (for instance, random motor babbling, or the reached strategy) if a strategy proves unable to provide a motor command.

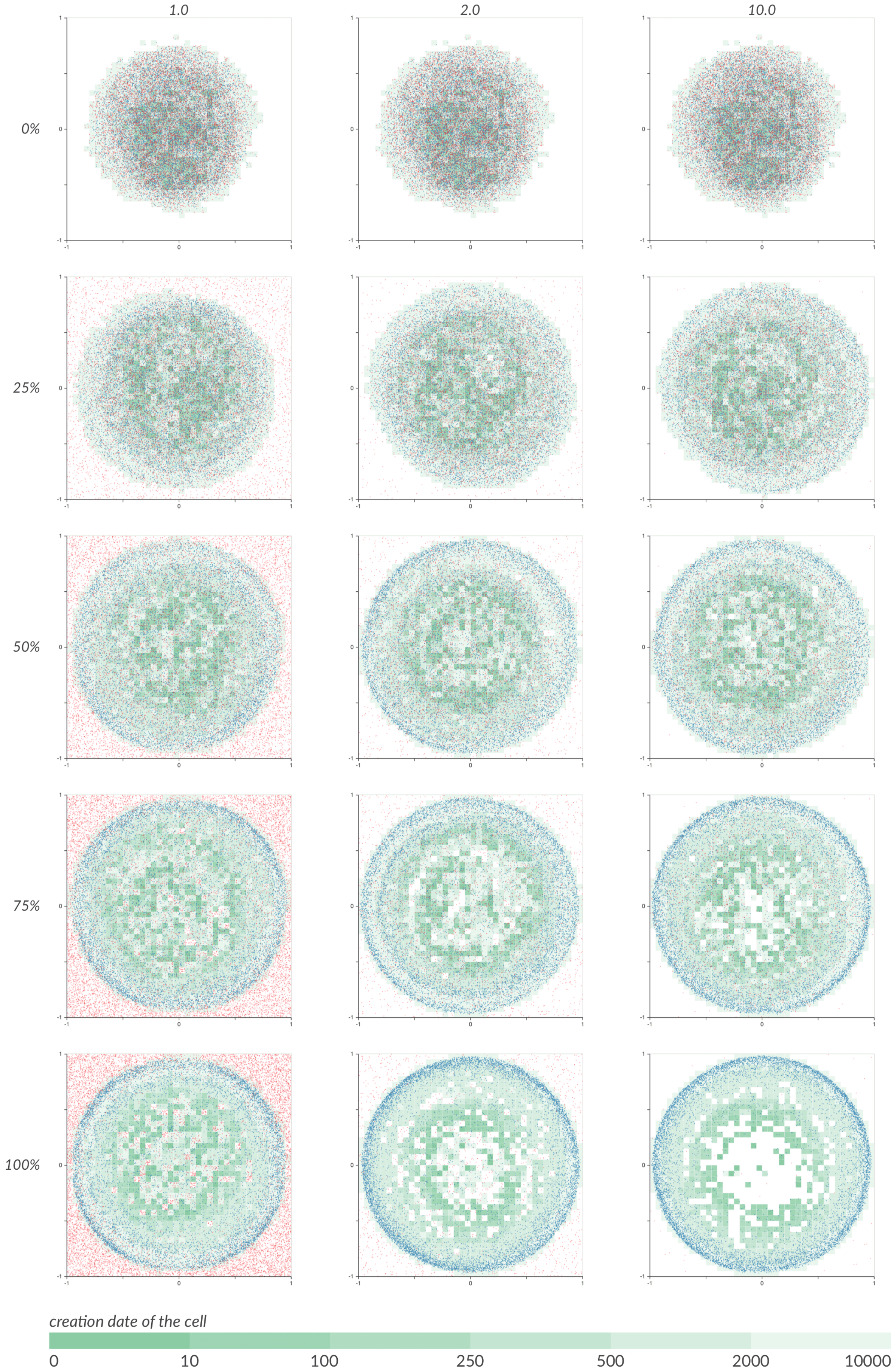


Figure 3.9: A balance ($p = 0.5$) between the *reached* and the *unreached* strategy proves effective, and robust to the goal space dimensions. The percentage of the *reached* versus the *unreached* exploration strategy modulates how aggressive the exploration is. The red dots represent the goals. Done on 10000 samples, with 10 random motor babbling bootstrap samples, on a 20-joint arm, with a 5 cm cell size. [source code]

mixed strategy that picks goals according to the *unreached* strategy p percent of the time, and the *reached* strategy $1 - p$ percent of the cases. As illustrated in Figure 3.7, we modify the EXPLORE algorithm of section 3.1.2, and replace the GOALBABBLING() call by a probabilistic call to the reached and unreached strategies. The inverse model remains the same ($d = 0.05$), as does K_boot, set to 10 timesteps. We will refer to this strategy as the p-reach strategy.

The value of p in the p-reach exploration strategy represents how aggressive the strategy is at trying to reach unexplored cells of the grid. By setting the value of p, the amount of exploration that is done inside and outside the reached space can be explicitly controlled. This makes the goal distribution adaptive to past exploration, and lessens the impact of the geometry of the goal space. This is illustrated Figure 3.9, where the effects of the p-reach strategy on the distribution of effects is displayed for p taking values $0, 25, 50, 75$ and 100.

While the 0-reach strategy produces a spread of effects that does not extend to the limits of the reached space, it does explore the centre exhaustively. As p augments, the numbers of cells located in the centre that are reached late in the exploration (lighter shades) increases. This creates blindspots in the exploration of aggressive strategies, that are exacerbated when the goal space overestimates the reachable space. A balanced strategy with $p = 0.5$ however, consistently provides a good exploration and seems robust to a large goal space.

To verify those results, we perform a quantitative analysis of 21 different p-reach strategies, with p varying from 0 to 1 over 0.05 increments. The results, Figure 3.8, reveal that at the 10000 steps horizon, a large number of values of p (roughly, $0.35 \leq p \leq 0.8$) provide a good performance. However, early in the exploration (t = 2000), a more aggressive strategy is preferable. This suggests that the best exploration strategies may need to make p evolve during exploration. More about that in chapter 4.

3.3.3 The FRONTIER Strategy

The p-reach strategy, for adequate values of p, is efficient and robust to the size of the goal space[8]. The robustness can partly be attributed to the inverse model used. Indeed, the inverse model projects the goal to its nearest neighbour—or, differently said, all the point of the goal space that have the same nearest neighbour in observed effects are equivalent—, setting a goal far from what is possible does not creates problems. When using a different inverse model, this may lead to singularities and inefficiencies.

This is particularly problematic when the reachable space is sparse compared to its axis-aligned bounding box, that we gave as an heuristic (using the *reached* space) for defining the subset of the sensory space the unreached strategy should choose goals in.

[8] When the goal space is larger than the reached space, that is. But this is not a difficult condition to verify.

Additionally, when p is high, the p-reach strategy tends to create unexplored areas in the centre of the reached space.

To solve these issues, we introduce the FRONTIER strategy. The FRONTIER strategy removes the need to explicitly define the boundaries of the goal space: they are consistently updated in function of the reached space.

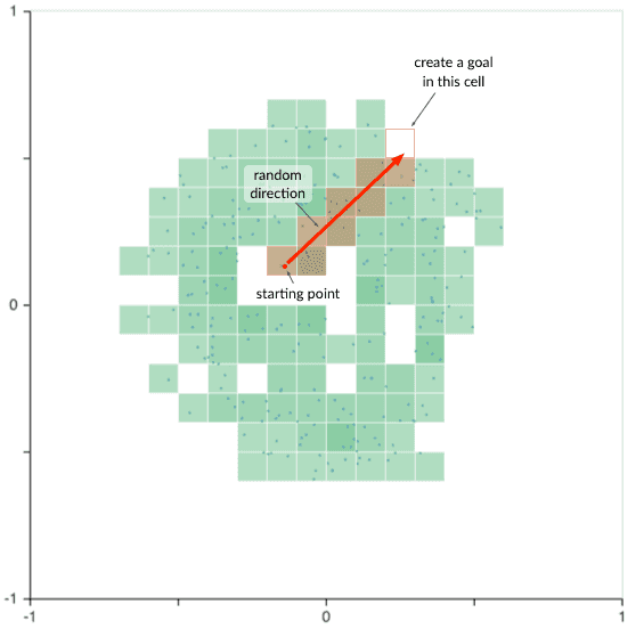


Figure 3.10: Illustrating the Frontier strategy.

The FRONTIER algorithm lays a grid on the goal space. At each timestep, a random existing effect and a random direction are chosen. The grid is then traversed starting at the selected effect, and moving in the chosen direction. The goal is randomly drawn from the first empty cell traversed in this manner. Figure 3.10 illustrates the process.

The idea behind this algorithm is not new. It can be found in the *goal directional sampling* algorithm of Rolf (2013), and previously, in the *SAGG-RIAC* algorithm of Baranes and Oudeyer (2010). A similar idea can also be found much previously in the *shifting setpoint algorithm* of Schaal and Atkeson (1994) (see also Atkeson et al. (1997a,b)). Contrary to those methods, the FRONTIER algorithm does not take multiple steps toward a goal or reevaluate the direction if no sufficient progress is made towards it. Instead, the FRONTIER algorithm chooses a goal, does one step of exploration and then switches to another goal.

In Figure 3.11, five explorations are shown, with the FRONTIER strategy being used

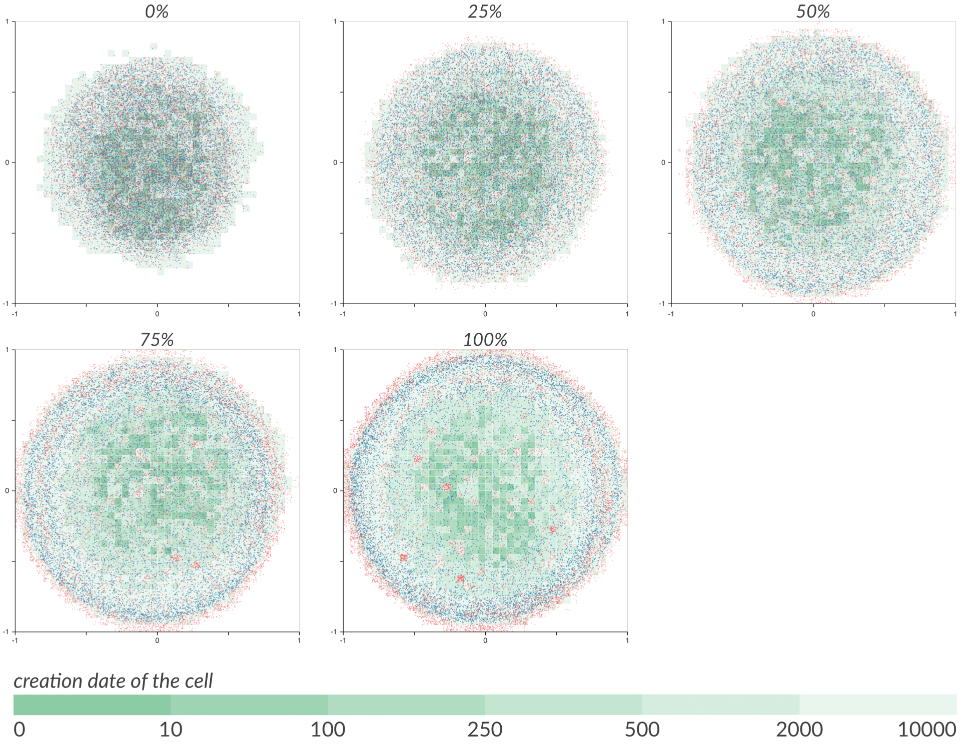


Figure 3.11: The Frontier strategy allows different amounts of aggressiveness while exploring the insides of the reachable space completely. The exploration is driven by a mixed strategy between the Frontier strategy and the reached strategy, over 20000 steps. Cell size is 5 cm. [source code]

0%, 25%, 50%, 75% and 100% of the time respectively[9]. The rest of the time, goals are chosen randomly inside the reached space (i.e. this is the p-reach strategy with the unreached strategy replaced by the Frontier strategy). The Frontier strategy displays exploratory aggressiveness, while exploring the inside of the reached space correctly.

In the 100% case, the last unexplored cells inside of the grid have had a large number of goals set inside them. This is not necessarily a desired behaviour, and the Frontier strategy can be further parametrized by setting a maximum number of goals that can be set per cell, and a minimum number of effects per cells before a cell is ignored when choosing a goal (in our original description, the minimum is equal to 1). Figure 3.12 exemplifies those parameters, using the Frontier strategy 100% of the time after the 10 timesteps of motor babbling, and setting the maximum number of goals at 6 per cell, and the minimum number of effects at 2 per cell.

The Frontier algorithms strikes a balance between conservative and aggressive exploration. By placing each goal near observed effects, yet in unexplored areas, the

[9] In our implementation, we only coded axis-parallels directions. This makes the code simpler, and is estimated to have little impact on the exploration.

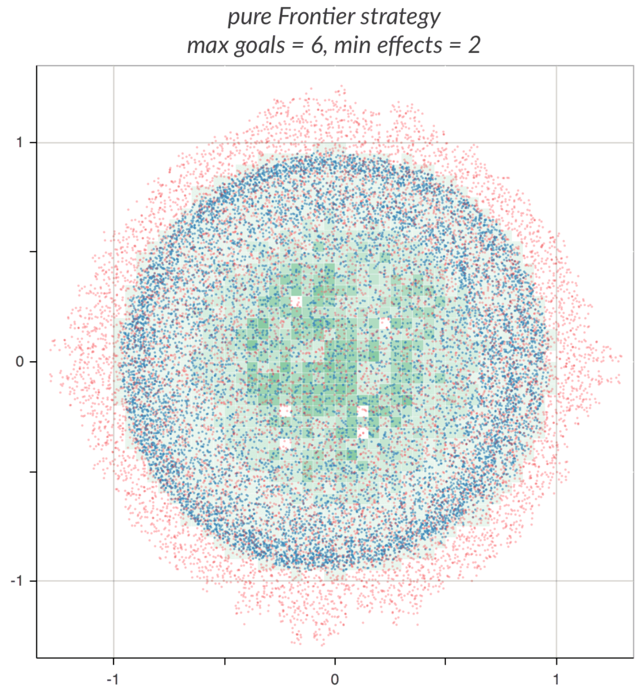


Figure 3.12: The parametrized Frontier strategy provides a balanced exploration. A few cells in the interior are empty but not laden with goals, and the goals have spread out. Cell size is 5 cm. [source code]

algorithms ensures that it can rely on reliable (because near) observations while increasing the potential to reach a new area. To make the Frontier more adaptive, we could consider to set the minimum number of effects per cell as a fraction of the average number of effects per cell. This would balance the exploration between the centre of the reachable space, and the pooling of observations on the edges of it.

Discussion

We established that either partially with the p-reach strategy, or completely with the Frontier strategy, the exploration process can free itself from the necessity of explicit boundaries on the goal space. This is a small yet important step towards self-sufficiency in exploratory behaviour. It makes these exploration strategies more robust because no experiment-specific bias on the geometry of the goal space is needed.

The main criticism one can make here, and this will be valuable for most of our experiments, is that those algorithms were tested one *one*, very simple environment not even involving simulated physics. This casts doubts on the applicability of those results to more complex setups, or to real robots.

We also realize that we did not do a quantitative comparison between the original goal babbling strategy, the p-reach strategy and the Frontier strategy. This will be rectified.

For chronological reasons, the Frontier strategy will not be employed in other experiments of this thesis; its usage would probably marginally affect some results.

3.4 The Inverse Model

Abstract · We investigate the impact of inverse model's quality on the exploration.

Given a goal, the inverse model finds the nearest neighbour in the observed effects and applies a small perturbation on its corresponding motor command (see section 3.1.1).

The only free parameter d in $\textsc{Inverse}_p(g)$ impacts the quality of the model. If d is too low, the perturbation is too small and does not produce enough environmental changes, hindering the progress of the exploration. It also can become indistinguishable from noise in noisy environments. If d is too high, then the inverse model approximates random motor babbling (it is equal to it when $d = 1.0$).

In chapter 0, d was tuned for good performance. What happens to the exploration we modify the value of d? The effects of different values of d can be seen qualitatively Figure 3.13 and quantitatively Figure 3.14. We use the vanilla Explore strategy (see section 3.1.2).

In Figure 3.13, we observe that when the value of d is low ($d = 0.001$, $d = 0.005$), the exploration degenerates into 10 disconnected clusters, corresponding to the random motor babbling commands. Higher values of d ($d = 0.05$, $d = 0.1$, $d = 0.2$) offer good performances, but the loss in exploration performance is noticeable in the 7-joint and 20-joint arm. In the case of the 20-joint arm, $d = 0.2$ represents $\pm 60^\circ$ of random variation for each motor. The possible displacement between an arm posture and its perturbation is almost unconstrained in terms of end-effector position. Still, significant performance is displayed. For high values of d ($d = 0.5$, $d = 1.0$), the behaviour is respectively similar and identical to the behaviour of the random motor babbling strategy.

An interesting thing to notice is that random motor babbling seems preferable to random goal babbling when $d = 0.001$ or even $d = 0.005$ for the 7-joint arm. For the 20-joint arm however, the exploration is definitely better for $d = 0.005$ but not for $d = 0.001$. In any case, this illustrates that goal babbling with a bad learner is not very efficient. In other words, goal babbling is an effective strategy insofar as a good-enough inverse model allows to move around the sensorimotor space with efficiency.

The quantitative analysis displayed Figure 3.14 validates these qualitative remarks; the performance follows a U-shape, with bad performance near the extremes, and $d = 0.001$ is worse than $d = 0.5$ for the 20-joint arm.

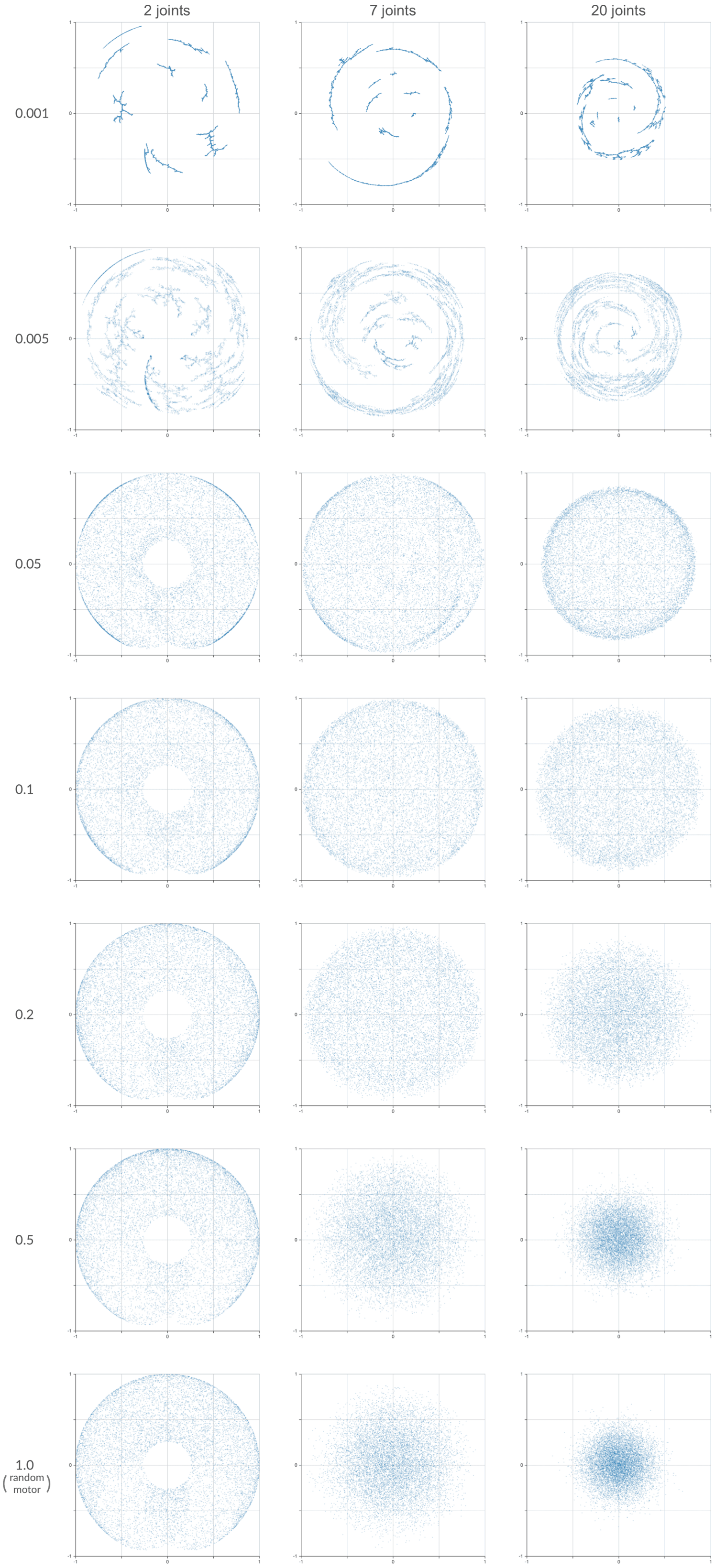


Figure 3.13: The quality of the learner correlates with the quality of the exploration. Yet even with degenerated learners on the 7-segment arm, the exploration is better than motor babbling. [source code]

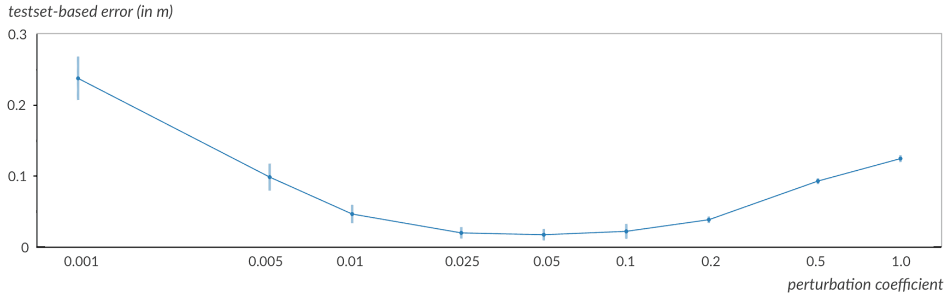


Figure 3.14: The testset average distance describe a U-shape, and take high values around the extremes. Computed for a 20-joint arm on a 10000-step random goal babbling strategy, repeated 25 times. [source code]

3.5 Motor Synergies

From a cursory review of the example of chapter 0, one could conclude that motor babbling becomes less effective as the dimensionality of the motor space increases[10]. As pointed out, this is not the correct interpretation: motor babbling is less effective in sensorimotor spaces where the heterogeneity of the redundancy is high, precisely because motor babbling can be understood as a density estimator of the redundancy.

Musculoskeletal systems in biological systems typically exhibit motor synergies (Holst 1939), i.e. groups of muscles that activate together. An explanation for these synergies put forward by (Bernshteĭn 1967) was that they were reducing the redundancy of the musculoskeletal system, explaining how biological entities were able to control complex, highly redundant limbs. An alternative explanation is that the spinal cord interneurons dramatically increase the number of motor dimensions and give access to hardwired pattern generators that are responsible for motor synergies (Perfiliev et al. 2010; McCrea et al. 2008). The dimensionality increase is accompanied by an important decrease in the heterogeneity of the redundancy(Raphael et al. 2010). In other words, the cerebellum and the spinal cord provide a control interface for the muscles where the density of useful solutions is higher than if muscles were wired independently. This allows to find good behaviour by trying random motor activations, and improving them towards the nearest local minima by trial-and-error:

> *The strategy for learning the repertoire of good-enough motor skills may consist largely of trial-and-error exploration of a high dimensional space that evolution has endowed with an unusually favourable distribution of attractors to desirable behaviours*
>
> Loeb (2012, p. 761)

We illustrate this by showing that we can actually improve the performance of the random motor babbling strategy by increasing the dimensionality of the motor space. In our setup, considering an arm with n joints, there are n motor channels—let's call them $\mathrm{c}_0, \mathrm{c}_1, \ldots, \mathrm{c}_{n-1}$ here—, each sending an angle command to its respective joint. We add $\lfloor \frac{n}{2} \rfloor$ other motor channels $\{\mathrm{syn}_0, \mathrm{syn}_1, \ldots, \mathrm{syn}_{\lfloor \frac{n}{2} \rfloor}\}$, with channel syn_i sending a motor command to joints $2i$ and $2i+1$. Therefore, each joint receives 2 angle commands, except the last one if there is an odd number of joints. The angle command are averaged according to the channel weight, which is equal, for each channel, to the inverse of the number of joints they target. That way, no channel exerts more influence over the final angle configuration that any other. Formally, the value of angle joint i is given by:

$$\mathrm{angle}_i = \frac{2}{3} c_i + \frac{1}{3} syn_{\lfloor \frac{i}{2} \rfloor}$$

[10] In our example, the dimensionality of the motor space is equal to the degrees of freedom of the arm—i.e. the minimum number of independent variables required to define the position of the arm.

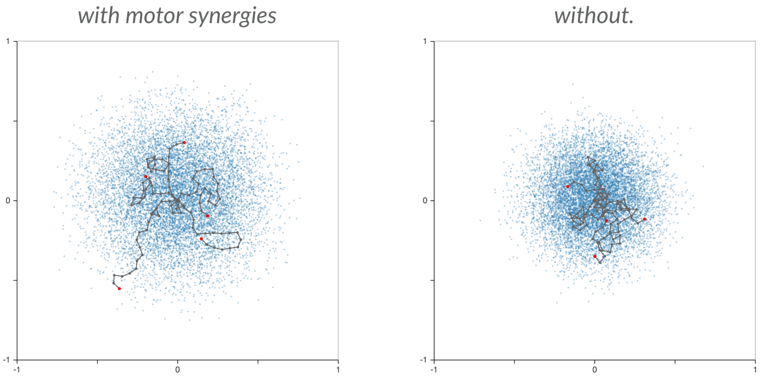


Figure 3.15: By increasing the motor space dimensionality from 20 to 30 by adding motor synergies between neighbouring joints, we improve random motor babbling. The two pure random motor exploration strategies are run on two 20-joint arms, one of which is equipped with motor synergies, for 10000 timesteps. [source code]

We run the random motor babbling strategy on a regular arm and on another with motor synergies, each with 20 joints. The results are available qualitatively Figure 3.15, and quantitatively Figure 3.16. The increase in the exploration performance of random motor babbling is significant.

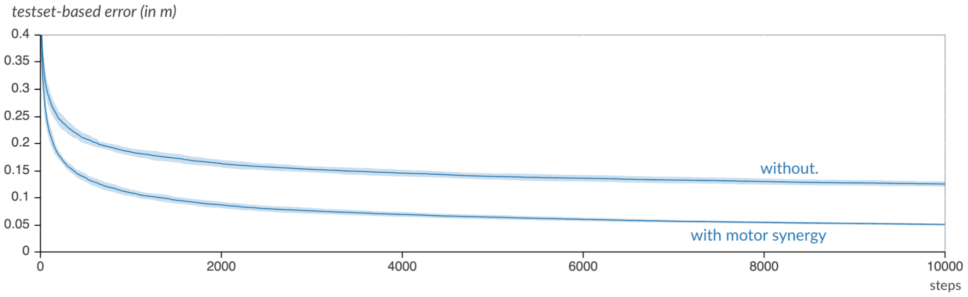


Figure 3.16: Although the asymptotic performance of random motor exploration is zero, in the context of a reasonable timeframe, adding synergies drastically improve the final performance of random motor exploration. [source code]

In order to explain the results, we can observe that the synergy channel between two consecutive joints has a correlation effect: it makes the difference between the value of the joints smaller[11]. This in turn make the arm straighter on average, allowing for greater reach, and better exploration.

We do not claim to have proven any positive results—and certainly our example of motor synergies is simplistic and contrived. Yet we provided a non trivial counterexample to the notion that a greater number of motor dimensions would be detrimental to motor babbling.

[11] If the values of regular the motor channels for joint 1 and 2 are v_1 and v_2, then whatever the value of the synergy channel of both s_{12}, the difference between the angle of the first and second joint is $|(\frac{2}{3}v_1 + \frac{1}{3}s_{12}) - (\frac{2}{3}v_2 + \frac{1}{3}s_{12})| = \frac{2}{3}|v_1 - v_2|$. The difference is reduced to 2/3th of its value in a setup without synergies.

3.6 Developmental constraints

Developmental constraints—or *maturational constraints*—are limitations that are placed on the agent's motor, morphological, sensory or cognitive abilities, and that evolve during development[12]. Long thought to be obstacles to children's development, developmental constraints started to be recognized as an essential component of development in the early 1980's:

> *The widespread use of the term 'immature' to describe the infant and child attests to the dominance of the view that structure and function are best understood in terms of what they will become, rather than in terms of what they are. It is possible, however, and probably meaningful to view the somatic and behavioural capacities of the developing organism as uniquely adapted to that organism's current stage of development.* Turkewitz et al. (1982, p. 358)

In particular, developmental constraints are credited in reducing the size and the complexity of the sensorimotor space available in infancy (Rutkowska 1994; Berthouze et al. 2004). Developmental constraints can be broadly discriminated into cognitive and sensory constraints on one side, and motor and morphological on the other side.

Elman (1993) was among the first to illustrate the importance of constraints in a synthetic setting: using a recurrent neural network, he showed that a network whose

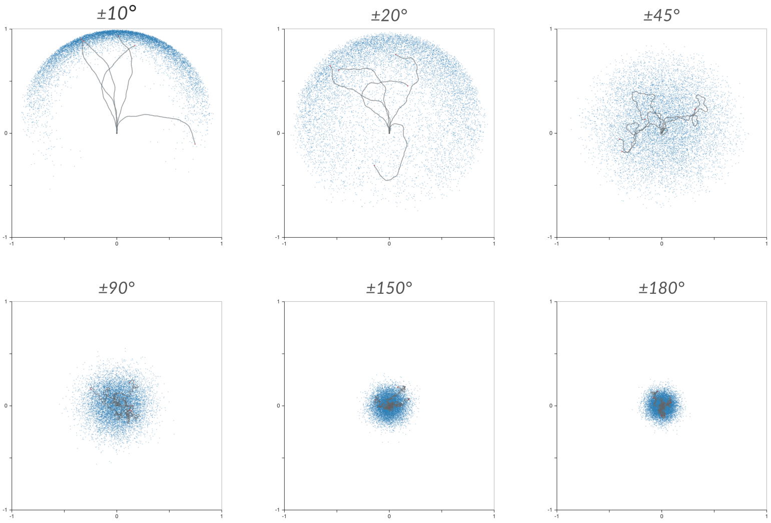


Figure 3.17: A low joint range makes random motor babbling very effective. Here the arm has 100 joints, the strategies are run over 10000 steps, and repeated 25 times. [source code]

[12] Constraints are generally lifted during development in biological systems, but we use *evolve* because the assumption that they are always lifted, never tightened, is too strong to make without justification.

available memory was at first low and then expanded later during the training sequence performed better than the same network with the whole memory made available from the start.

Many works on cognitive and sensory constraints have targeted vision. French et al. (2002) and Dominguez et al. (2003) showed that restricting the sensory frequencies in a vision system improved performance, with fixed and maturational constraints respectively. More recently, Nagai et al. (2006) pioneered a work on maturational constraints and shared attention, showing that restricting visual capacities allowed to learn one aspect of the interaction at a time.

Morphological and motor constraints have received a lot of attention, as they are easier to implement in their simplest instantiation. Bongard (2011) has shown that the development of gait controller for hexapod robots was faster if the robot started with small limbs that grew throughout the experiment. Lee and Meng (2005) and Lee et al. (2007a,b) propose a framework that lift constraints to create staged and organized development of motor coordination.

For the two-dimensional arm, an example of a simple morphological constrain is to reduce the range of available angles during the exploration. Figure 3.17 shows the effect of different angle ranges on the exploration of the random motor babbling strategy: reducing the range of the joint to appropriate levels makes the random motor babbling strategy able to efficiently explore most of the reachable space of a 100-joint arm, as is apparent in the $\pm 20^{\circ}$ and $\pm 46^{\circ}$ cases: good constraints are more efficient than goal babbling in this case.

This can be exploited by reducing the range of the joint range during the early phase of the exploration, and lift the constraint thereafter. We run a goal babbling exploration where during the first 500 steps of the exploration, the joint ranges are limited to $\pm 80^{\circ}$. After the 500 timesteps, the constraints are lifted and the range of the joints returns to $\pm 150^{\circ}$. As can be seen in Figure 3.18, the constrained scenario generates a better exploration than the one without constraints. Additionally, the loops that were observed on the arm in chapter 0 are not present[13]. The early constraints drive the exploration to good attractors in the sensorimotor space.

Once again, this is merely an illustration of an idea, and is not meant at establishing any sort of general result.

[13] This would need a more thorough, quantitative analysis to be claimed as a solid result.

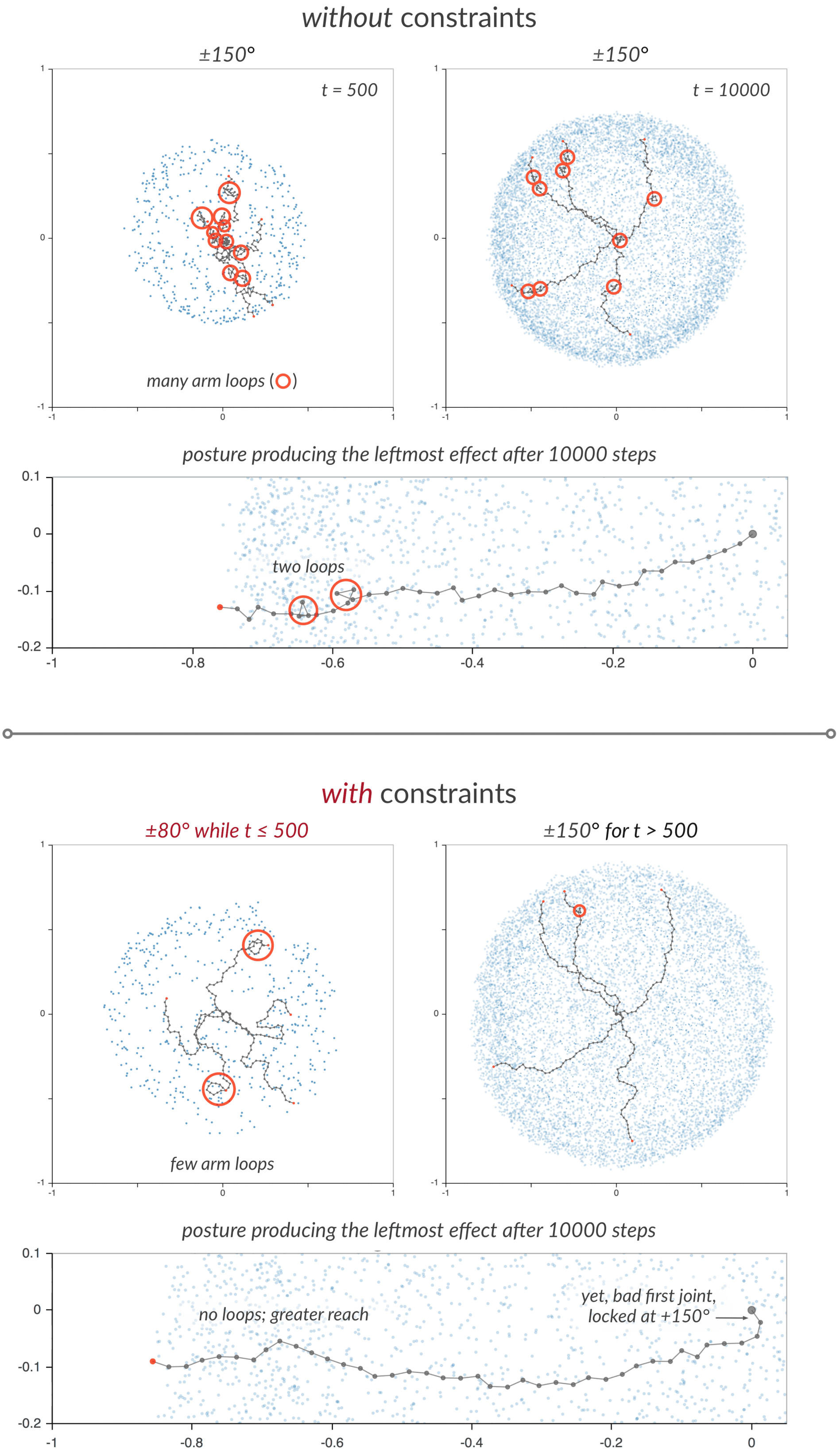


Figure 3.18: Constraints can help exploration. In this scenario with two 40-joint arms, one is constrained to ±80° during the first 500 steps, while the other is not. This seems to significantly decrease the number of loops that are present in the explored postures. A comprehensive quantitative analysis is needed to verify those observations. [source code]

3.7 Demonstrations

Learning from demonstration, also called learning by imitation, has been recognized as an important technique for current robot learning[14] (Schaal 1999; Billard et al. 2008; Calinon 2009; Argall et al. 2009; Lopes, Melo et al. 2010).

In this section, we show that providing a good demonstration can have a dramatic influence on exploration. In Figure 3.19, a single demonstration is provided to the explorer, the zero-posture, where all joint angles are zero, resulting in the effect $x = 0, y = 1$. A normal random goal babbling exploration is then run for 10000 steps on a 100-joint arm.

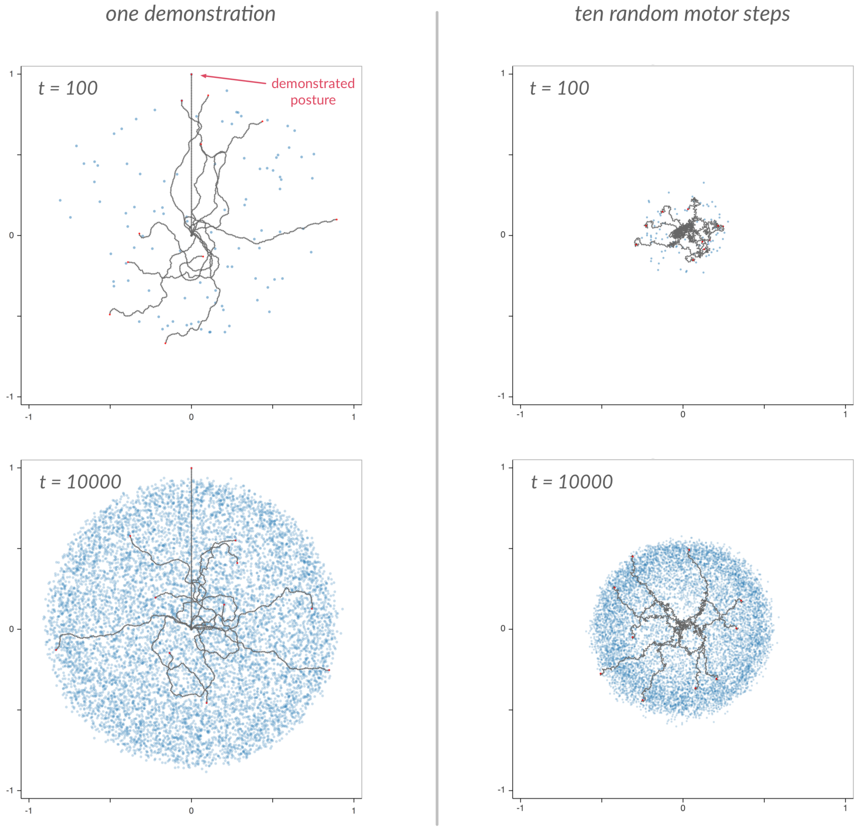


Figure 3.19: Demonstration can have high beneficial influence on exploration. [source code]

The demonstration provided to the robot is not innocent. It places the exploration

[14] As the robotic learning domain is in infancy, learning algorithms are still exceedingly limited. Typically, learning algorithms can derive a solution if given a starting point not too far away from the solution—or with enough guidance to get to it. Learning by demonstration does precisely one or the other. But social learning should not be understood as mandatory for the development of highly intelligent behaviour. Several species of cephalopods (amongst which, octopi) display highly intelligent behaviour yet live short, solitary lives (2-3 years).

is a good attractor where arm loops are absent, from which it can easily explore the reachable space. The difference in coverage with the regular goal babbling strategy is dramatic, and is maintained over the long run.

This result, and the one from the previous section on developmental constraints suggest that simple exploration mechanisms can, in some instances, tackle complex environments, as long as some mechanism, limitation, prior or external influence help them discover good areas of the sensorimotor space.

Discussion

Simple Environments

> *[W]ith a simplified world [...] it is very easy to accidentally build a submodule of the systems which happens to rely on some of those simplified properties...the disease spreads and the complete system depends in a subtle way on the simplified world.*
>
> Brooks (1991b, p. 7)

Given our purposeful exposition of the embodiment concept in the previous chapters, it would behove us to heed Brook's forewarning. Yet, clearly, our work forgoes it.

Our work fails on many fronts. It employs an oversimplified environment, and it employs only one environment. We cannot argue that the mechanisms we expose tell us anything about real robots' exploratory behaviour, nor can we claim even limited domain independence.

Moreover, although we advocated the role of embodiment in chapter 1, the two-dimensional arm is far removed from a context that would allow us to study such a phenomenon in any realistic fashion. There is no noise, motor commands operate over a discretized time; a motor command unambiguously corresponds to one sensory feedback. The sensory signal itself is highly abstracted, and does not correspond to any reasonable self-sufficient sensory hardware (considering an adversarial environment). The environment is perfectly isotropic, and does not feature any events other than the one created by the robot. Brooks advocated starting simple, but in a realistic environment.

By studying exploration strategies in a simplified environment, we run the risk to—or rather, it is certain that we—ignore problems that exist in the real world. In our view, for this chapter, this is a feature. By ignoring many problems that a real robot faces, we obtained a controlled situation where the impact of each component of the exploration could begin to be well understood, and where comparatively fewer different hypotheses can be made to explain the results.

In fact, the investigative method we followed to create those experiments was to simplify them as long as their qualitative illustrative properties remained. Developmental robotics has chosen a difficult path, fraught with theoretical, experimental, and methodological issues. One of them arise when the investigated systems become complex: how to study, then, the contribution of each dimension of the system to the experimental performance? Their sheer number makes a complete experimental analysis quickly unfeasible. And, even when the variation of only one dimension is of interest, the non-linearity of its relationship with the rest of the system still makes the analysis expensive.

This point was perhaps best argued by Richard Lewontin, when criticizing the experimental approach of genetics:

> *To determine the motion of a planet I need know only its mass, distance from the sun, angular velocity, and the path of a couple of nearby bodies. For the development and behavior of an organism I need to know vastly more. But how are we to investigate the network of weakly determining causal pathways except by experimentally (i.e., unnaturally) holding constant as many variables as possible and making (unnaturally) large perturbations in one or two other variables—large enough, that is, to see an unambiguous effect. When we do this, however, we have a problem of extrapolation. Would the effect of the perturbation be the same if we had other values of the controlled variables, or, worse, would the effect of small perturbations be simply scaled-down effects of the unnaturally large ones, or are there threshold effects?*
>
> Oyama (2000, [)

xi] The interesting point of Lewontin arguments that even if an exhaustive study of all the gene perturbations regarding a particular trait is done, nothing is necessarily learned about how the trait is produced in the first place: 'A sufficient explanation of why two things are different may leave out virtually everything needed to explain their nature.' (Oyama 2000, p. ix).

Our situation is different of course. We benefit from knowing the nature of the phenomenon we are studying. But studying a complex phenomenon, in a realistic environment, is not only experimentally challenging, it also reduces the tools we can use to analyse it, and forces us to act like a geneticist, changing one variable at a time to discover which exploration strategy works best. And in any complex-enough scenario, many of those variables, or variable ranges would have to be left unstudied.

But what we most crucially loose, proceeding this way, is the explanation behind the results. We took the deliberate decision in our experimental approach to target a setup that was not completely adverse to an exhaustive study, and that was not the least impenetrable to our comprehension. And we made efforts to go beyond the simple perturbation/observation method, and provided explanations of the results as well as the results themselves. The criticism that one could make is that perhaps we did not do enough in this regard.

A Critical Analysis of Intrinsic Motivations

The motivation for this research was initially to thoroughly study how intrinsic motivations were contributing to the behavioural success of agents compared to simpler goal babbling strategies. Evidently, we only managed to start studying simpler goal babbling strategies. Yet, this analysis can serve to put some results of the literature in perspective.

Baranes and Oudeyer (2013) proposed SAGG-RIAC, a goal babbling algorithm where goals were selected according to a intrinsic motivation measure, based on competence progress. SAGG-RIAC was tested on a setup similar to ours: a 15-joint two-dimensional arm. It is not pointless to say that this work has had a tremendous influence on ours. Besides underscoring the better performance of goal babbling over motor babbling on this task, the main quantitative result was that random goal babbling performs worse than intrinsically motivated goal babbling when the goal space is larger (in the article ~9 or 100 times larger) than the reachable space. When the goal space fits the reachable space, no significant difference can be shown.

Our experiments allows to add to these results. First, the choice of the inverse model understandably affects the performance of goal babbling. Baranes and Oudeyer (2013) used the inverse of the Jacobian estimated from sampled data, which is sensitive to the goal distance, and therefore is susceptible of behaving badly when asked to find solutions to impossible goals. Of course, SAGG-RIAC is precisely intended to deal with this problem, by monitoring competence, hence the performance of the inverse model. But the experiment does not disambiguate between the loss of performance due to the inverse model misbehaving faced with impossible objectives, and the loss of performance due to the distribution of goals driving the exploration towards the edges of the reachable space rather than being homogeneous over it.

Second, we provided a method to adapt the goal distribution to the reached space using the Frontier algorithm. Since a correctly sized goal space does not produce any difference in performance between random goal babbling and intrinsically motivated babbling in SAGG-RIAC, a Frontier-backed exploration strategy would probably produce competitive performance even with an oversized goal space. As such, the performance advantage of intrinsic motivation over a (subjectively) simpler method, the Frontier algorithm, has not been clearly empirically demonstrated[15]. Of course,

[15] Baranes and Oudeyer (2013) proposed another experiment using a quadruped robot, but only considered a larger space that the reachable space in that case.

this discussion would have been more substantial if we had actually done the actual experimental comparison.

But our intention with this discussion is not to say that intrinsic motivation is not useful. It is evident that an intrinsic motivation measure such as competence progress used in SAGG-RIAC or others proposed in the literature affords the agent adaptive capabilities that have the potential to discriminate his performance from other agents in significant ways in complex environments. Our point is: finding complex environments where the improvement that intrinsic motivation brings can be unambiguously established by eliminating all other hypotheses is a still a major research challenge[16].

Proving such a point convincingly would further the evolutionary reflection on intrinsic motivation. If we assume that intrinsic motivations require significant cognitive resources in organisms (not necessarily a trivial point to make), justifying intrinsic motivation from an evolutionary standpoint must account for fitness advantages that simpler cognitive processes cannot provide. The same point can be adapted to the computational realm.

As a final remark on this subject, let's note that discriminating between intrinsically-motivated agents and non-intrinsically motivated agents is not trivial. As most agents have no clear physiological need (that they are aware of), most could actually be considered intrinsically motivated *by default* to do their tasks for the task's sake, and not an outcome that they may not always observe, much less 'understand'. More precisely, we have seen in section 2.5 that R-max could be considered as novelty-driven. In the same fashion, our Frontier algorithm could be viewed as novelty-driven, because it chooses goals according to a mechanism that targets unexplored areas. We pitched, in essence, a competence-driven motivation against a novelty one. This underscores the need to qualify our use of 'simpler strategies'. A more systematic approach may be needed, qualifying exploration strategy's simplicity by their algorithmic complexity, in time and space, which we may also use to estimate the cognitive cost they may represent.

Cheap Design

Another danger of using simplified environments is to study problems that do not exist in the real world. This is a criticism that is difficult to address directly. Yet, in this chapter, our aims were not to simulate reality, but to show how different phenomena could impact exploration. We studied the impact of goal distribution, of the inverse model performance, of motor synergies, of developmental constraints and of demonstrations. Each time, the investigation was rarely comprehensive. Yet, the set of experiments shows that the solution to efficient exploration processes is a multifaceted approach. Putting the best learning algorithm behind the exploration will im-

[16] The irony of using a set of experiments on a simplified environment to make that point is not lost to us.

prove performances, perhaps tremendously, but this is not cheap design (Pfeifer and Bongard 2006, p. 107), nor does it respect ecological balance (Pfeifer and Bongard 2006, p. 123).

The learning algorithm is probably going to be complex. And learning is not the best place to solve *all* the challenges that the environment raises. Our experiments suggest that an equivalent (or possibly superior) performance can be obtained by combining different, loosely coupled (Pfeifer and Bongard 2006, p. 134) approaches.

Breeding Arm Postures

Our environment, and our focus on diversity, lend themselves to an analogy with evolutionary algorithms. Let's consider that our objective is to breed a population of arm postures. Each arm posture represents the genetic code, and when translated into their phenotype, they produce an end-effector position. We do not have sexual reproduction in our world, everything is done by random mutation. Starting from an initial population of random genetic codes—random arm postures—, we evolve, one a time, their offspring, that produce new end-effector position. Rather than selecting arm postures over environment-specific fitnesses, they are selected to foster diversity in the population: nobody dies ever, so the most *edgy* members are selected to produce offsprings that can venture into yet-unexplored areas.

Under such an analogy, the similarity between our exploration algorithms and the evolutionary robotics approaches based on diversity of section 2.6 is evident. Of course, there are differences, and our analogy works well due to the specific nature of the inverse model we used. But this suggests that both domains can probably get insights from the methods of the other.

Celui qui diffère de moi loin de me léser m'enrichit.
[He who is different from me does not impoverish me—he enriches me]
Antoine de Saint-Exupéry

4
Diversity-Driven Selection of Exploration Strategies

In the previous chapter, we have been looking at the impact of several variations of the experiment in chapter 0. While we investigated strategies that were themselves composed of different strategies, *when* and *how much* to use each strategy was always fixed. In this chapter, we consider situations where several exploration strategies are available and the agent must choose dynamically at each timestep which one to use to generate the next motor command.

4.1 How Much Motor Babbling?

Abstract · The quality of the learners impacts how much motor babbling should be performed: when the learner is bad, more motor babbling is preferable.

In section 3.4, we investigated the impact of the quality of the learner on the effectiveness of the goal babbling strategy. It was noted, in particular, that when the perturbation parameter d of the inverse model—the amount of perturbation that the motor command corresponding to the observed effect nearest to the goal is subjected to—is low, the goal babbling strategy performance is hindered in such a way that the motor babbling strategy becomes preferable.

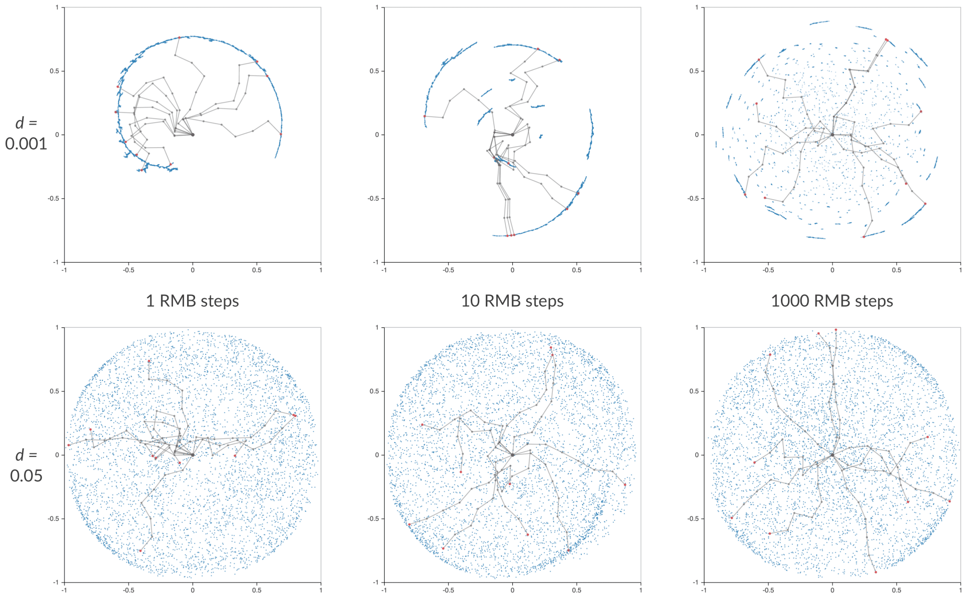


Figure 4.1: The number of random motor babbling steps has a significant impact when the inverse model of the goal babbling strategy is bad ($d = 0.001$), but little effect when it is good ($d = 0.05$). The results depicts the exploration trajectories of a 7-joint arm, over 5000 timesteps, with, at the beginning, 1, 10, or 1000 random motor babbling steps before starting a pure random goal babbling strategy. [source code]

Yet, because the random motor babbling is unable to take advantage of past observations, and therefore to reach the edges of the reachable space of the 20-joint arm, goal babbling, even backed by a poor inverse model, is still useful during the later phases of the exploration. In Figure 4.1, goal babbling strategies featuring motor babbling phases of 1, 10 and 1000 timesteps are compared, for a good ($d = 0.05$) and bad ($d = 0.001$) learner. When the learner is good, the length of the initial motor babbling phase has no significant impact on the quality of the exploration over the long term (5000 timesteps). But when the learner is bad, the longer motor babbling phase allows for a well-explored centre, and a goal-babbling-backed exploration of the edges of the reached space. *A contrario*, a short motor babbling phase creates degenerated clusters.

This leads us to investigate which percentages of motor babbling give good performances for a given learner. We consider three different scenarios: when the learner is bad, with a small perturbation parameter ($d = 0.001$), the learner is good, with $d = 0.05$, and when the learner is bad, this time with a large perturbation parameter ($d = 0.5$).

We use mixed strategies to analyse this. Rather than having an initial phase of motor babbling, followed by goal babbling, the motor babbling strategy is chosen with probability p at each step, and the goal babbling strategy with probability $1 - p$.

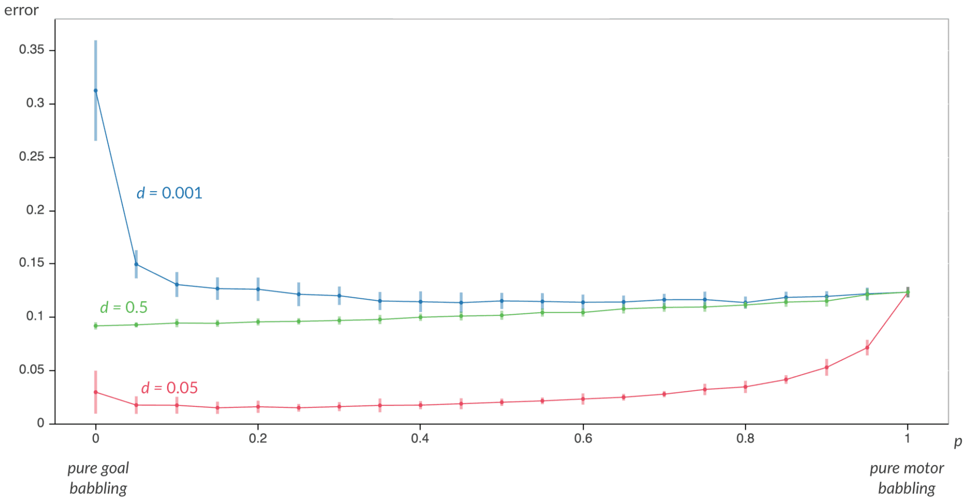


Figure 4.2: Extreme values of the perturbation parameter of the inverse model have undesirable effects on the exploration of a 20-joint arm. Run over 10000 timesteps. When $d = 0.001$, pure goal babbling is the worst strategy, more than twice as bad as a strategy with 5% of motor babbling. Inversely, when $d = 0.05$, a strategy with low—but not null—motor babbling works well. If the inverse model approximates randomness ($d = 0.5$), the difference is less marked, but perceptible, and goal babbling is always preferred to motor babbling in this case. Naturally, when random motor babbling is used 100% of the time, the performance of each scenario is identical. Averaged over 25 runs. [source code]

The first step is always motor babbling[1]. We consider all values of p from 0.0 to 1.0 by increments of 0.05, and run the exploration strategy for 10000 steps on a 20-joint arm.

The results Figure 4.2 show that different learning capabilities call for different exploration strategies[2]. When the learner does not produce enough diversity ($d = 0.001$), performance is best when random motor babbling is used between 35 and 80% of the time. When the learner is good, performance benefits from a small amount of motor babbling (15%), but starts being penalized if the proportion is more than 40%. And when the learner behaviour is only slightly better than random ($d = 0.5$), goal babbling completely dominates the random motor babbling strategy, because the goal babbling strategy is able to exploit the slight edge that the learner provides while producing enough variability. If d had been equal to 1.0, motor and goal babbling would have been indistinguishable.

Using motor and goal babbling in equal amounts ($p = 0.5$), the average performance is good in all situations. But it is not the best choice for the $d = 0.5$ case. No static strategy fits all situations.

[1] This would be the case anyway, as the implementation of the inverse model returns a random motor command if no observation is available in memory.

[2] This may seem self-evident here, but it is worthy of consideration for any self-sufficient agent: are the agent learning capabilities correctly sized-up for its environment? And when an agent is faced with a situation too complex to learn, does he have the capability to adapt its behaviour and learning strategy?

Since different learning capabilities call for different exploration strategies, and that the degree to which the learning capabilities match the challenges of the environment cannot be anticipated in a self-sufficient context, the choice of the strategy must reside with the actor, not the architect, and should be dynamically decided and refined during exploration. We introduce an algorithm that produces such an adaptive behaviour.

4.2 An Adaptive Strategy

Abstract · We introduce a strategy that dynamically selects other exploration strategies with respect to the diversity they respectively produce. The method is shown to successfully adapt to bad learners. We discuss it in the context of the Multi-Armed Bandits and the Strategic Student problems.

In section 2.5, we discussed how intrinsic motivation has been used successfully to guide exploration and learning over sensorimotor spaces. In the literature, intrinsic motivations have mainly been used for deciding what to learn, or, in our case, what to explore. Here we use intrinsic motivations to decide *how* to explore, namely, which strategy to use during the exploration.

Choosing which strategy to employ at each step of the exploration faces three main challenges:

1. **Interdependence**: an exploration strategy effectiveness may depend on another strategy; goal babbling relies on motor babbling to bootstrap the exploration. Given the inverse model currently used, this is even more true, as goal babbling's performance depends heavily the sensorimotor attractors in which it expands, and thus on the location of the observations produced early in exploration by motor babbling.

2. **Dynamical Value**: the usefulness of a strategy may change rapidly. Motor babbling is useful in the beginning of the exploration, but its usefulness drops quickly.

3. **Agnosticity**: since an explorer algorithm might be arbitrarily complex, and possibly involve, in turn, other explorers, an adaptive strategy should not rely on knowledge of the internal workings of the strategies amongst which it must choose.

Interdependence does not have to be handled directly, but suggests that even strategies that did poorly in the past must be re-evaluated regularly as the exploration progresses. The dynamical nature of the contribution of each strategy means that performance data becomes obsolete quickly, and encourage evaluations over short-term time windows. Agnosticity implies that the contributions of the strategies have to be evaluated only from the observations the strategies produce. We introduce a measure that matches those constraints now.

4.2.1 Effect Diversity

A strategy that produces effects over areas that have already been explored is of little use for exploration. We introduce an online *diversity measure* that evaluates, each time a strategy is used, how much diversity is created, with regards to already observed effects.

In order to do that, we rely on the diversity measure introduced in section 1.4, based on the union of disks centred on observed effects, and adapt it to evaluate a single effect: the diversity of a new observed effect is the increase in diversity, i.e., the increase in the covered area.

Definition 5. *Given a set of effects $E = \{\mathbf{y}_0, \mathbf{y}_1, ..., \mathbf{y}_{n-1}\}$, and a coverage threshold τ in $\mathbb{R}^+$, the diversity of a new effect $\mathbf{y}_n$ relative to E is defined as:*

$$diversity_\tau(\mathbf{y}_n, E) = volume\left(\bigcup_{i=0}^{n} B(\mathbf{y}_i, \tau)\right) - volume\left(\bigcup_{i=0}^{n-1} B(\mathbf{y}_i, \tau)\right)$$

with $B(\mathbf{y}_i, \tau)$ the hyperball with radius τ and centre $\mathbf{y}_i$.

The diversity of a strategy, in turn, is the averaged diversity of the effects it produced, over a given time window.

Definition 6. *Given a set of strategies $s_0, s_1, ..., s_{q-1}$, and a set of observed effects $E = \{\mathbf{y}_0, \mathbf{y}_1, ..., \mathbf{y}_n\}$, we have for a given strategy s_j a subsequence $\mathbf{y}^j_0, \mathbf{y}^j_1, ..., \mathbf{y}^j_{n_j}$ of the effects produced by motor commands emanating from the strategy. Given a time window w in $\mathbb{N}^+$, we define the diversity of strategy s_j as:*

$$diversity_{\tau,w}(s_j, E) = \begin{cases} \dfrac{1}{w'} \displaystyle\sum_{i=0}^{w'} diversity_\tau(\mathbf{y}^j_{n_j - i}, E) & \text{if } n_j > 0 \\ 0 & \text{otherwise} \end{cases}$$

with $w' = min(w, n_j)$.

4.2.2 Multi-Armed Bandits and Strategic Students

Using the diversity measure, we can now evaluate the contribution of each strategy to the exploration. Our problem is similar to—although not the same as—the Multi-Armed Bandit problem (MAB) (Robbins 1952): we have to choose between a finite number of different strategies with different diversity scores, and after choosing one we receive a feedback signal from which we compute an updated score.

The classic MAB problem considers only bandits that are independent from one another (choosing one does not affect the value of the others), and stationary (the distribution of rewards of the bandit does not change). A variation of the problem, the *adversarial* (also called *non-stochastic* or *non-stationary*) MAB, removes the stationary and interdependence assumptions: an adversary is free to choose arbitrary rewards for each bandit at each timestep.

In practice, a significant portion of the published literature on the adversarial MAB problem only removes the stationary assumption. In other words, the problem takes place in the *oblivious* opponent model: the actions of the adversary, i.e. the rewards for each bandit at each timestep, are decided before the game starts. This is the case in Whittle (1988) and Auer et al. (2002), who investigate rewards that can arbitrarily change. Garivier et al. (2008) presents *abruptly changing environments*, where all bandits' reward distributions change at specified timesteps. Cesa-Bianchi et al. (2006, pp. 156–169) provides a treatment of the nonoblivious case.

One shall remark that an arbitrary sequence of rewards generated in the oblivious opponent model is indistinguishable from one generated in the nonoblivious opponent model if the game is played once—which is the case in the real world (an agent has access to only one timeline). As a result, oblivious MAB algorithms usually perform well in practical nonoblivious scenarios, where the opponent is not *malicious* (i.e. not actively trying to minimize rewards by modelling the agent's behaviour; the nonoblivious property only comes from environmental dependencies between tasks.). Still, the definition of the opponent model has important ramifications for the proofs of optimality that are established in the previously mentioned studies. One difficulty of the unconstrained nonoblivious opponent model is that the best strategy is usually computationally intractable, as all interactions between choices have to be taken into account.

Recently, Lopes and Oudeyer (2012) introduced the *Strategic Student Problem* that tries to capture the issues involved when learning multiple tasks at the same time. A student has to learn multiple topics (maths, chemistry, history, etc.), and has limited resources (time) to do so. How should he allocate his study time between topics in order to maximize its mean grade at the end of the semester? A possibility is to consider the problem as a MAB problem where the bandits are learning tasks. Interestingly, the works of Baranes and Oudeyer (2010), discussed in section 2.5, can be understood in this perspective: each region of the goal space is a different topic, whose improvement is empirically measured through competence progress during learning, and the exploration strategy must decide how to distribute its action given those learning feedback signals.

The strategic student problem also considers another related problem: a student has one topic to learn, but several possible learning strategies. Which one should he choose? Is a mixture of several strategies better than employing the best one all the time? This is the problem of learning *how* to learn (Schmidhuber 1994). Baram et al. (2004) explored such a problem and showed that a dynamically selected mixture of three active learning strategies outperformed any pure strategy. Konidaris et al. (2008) demonstrated that empirically evaluating and selecting amongst different small state space representations specific to a task during learning was effective and avoided a large task space where learning was unfeasible. The work of Nguyen and Oudeyer (2012) investigates robots dynamically choosing between asking a teacher for a demonstration or doing self-exploration on their own. Jauffret et al. (2013) proposes a method where a robot can self-assess, and has a frustration drive. When frustrated, the robot can opt to choose social help to improve its performance. In the context of reinforcement learning, Hester et al. (2013) develops an algorithm that can evaluate dynamically which exploration strategy brings the most rewards. These exploration strategies are driven by extrinsic and intrinsic motivations: maximizing rewards, reducing variance, seeking novelty, seeking unexplored states (a binary novelty), and seeking or avoiding particular features of the state representation. Clement et al. (2015) uses the framework of the Strategic Student Problem to create a tutoring system that actively personalizes the sequence of activities to each student, by tracking their performance and identifying which exercises and modalities make the student progress faster. The works of (Baram et al. 2004), (Nguyen and Oudeyer 2012) and (Hester et al. 2013) are singular because they combine deciding *how* to learn, and deciding *what* to learn, using a hierarchical approach. The learning strategy is selected first (how), and then it chooses what input to sample (what).

Learning performance typically exhibits *diminishing returns*. A student can hope to greatly improve its grade by studying a subject he knowns little about, but can only hope modest improvement if its expected grade is already near the top. This characteristic of learning should inform the strategy the student employs. For this purpose, Lopes and Oudeyer (2012) consider *submodular functions* (Krause and Golovin 2014). Submodular functions are set functions[3] that are defined around diminishing returns: improvement that a new observation can bring is always greater early in the observation. Equivalently, considering a specific unobserved input, additional observations will not increase the input's expected improvement in performance. Mathematically:

$$\begin{gathered} f \text{ is submodular iff for every } A \subseteq B, \text{ and } \mathbf{y} \notin B, \\ f(A \cup \{\mathbf{y}\}) - f(A) \geq f(B \cup \{\mathbf{y}\}) - f(B) \end{gathered}$$

$f(A \cup \{\mathbf{y}\}) - f(A)$ is the improvement that $\mathbf{y}$ brings, having observed A. This correspond exactly to $\mathit{diversity}_{\tau}(\mathbf{y}, E)$, the effect diversity we have previously defined.

While submodular maximization is NP-hard (Feige 1998; Krause and Guestrin 2005), the greedy strategy is guaranteed to be no worse than $1 - \frac{1}{e} \approx 0.63$ times the optimal solution (Nemhauser et al. 1978), with e the base of the natural logarithm, in the case of non-decreasing submodular functions.

Of course, not all set of learning tasks exhibit a submodular structure. Still, it suggests that a good-enough performance might be obtained through simple-enough algorithms in practice. Lopes and Oudeyer (2012) and Hester et al. (2013) advocate the use of the Exp4 algorithm (Auer et al. 2002) rather than a greedy algorithm, as a more robust approach.

[3] A *set function* operates on sets. Given a set V, the set of all subsets of V (sometimes called the *power set* of V) is noted 2^V, and the set function $f : 2^V \mapsto \mathbb{R}$ assigns a value $f(S)$ to each subset $S \subseteq V$.

Compared to these works, our approach distinguishes itself in its objective: we are selecting exploration strategies to improve exploration, rather than exploration or learning strategies to improve learning. The resulting strategy is another exploration strategy, that can be used to replace any other exploration strategy in an exploration architecture. We provide an algorithm as simple as possible, based on selecting exploration strategies proportionally to their empirically estimated diversity.

4.2.3 The Adapt Algorithm

The Adapt algorithm chooses strategies proportionally to their diversity. To allow for constant reevaluation of the strategies, even those with low diversity, the algorithm chooses a strategy at random α percent of the time, with $\alpha > 0$. Algorithm 3 formally describes this.

Additionally, in order to foster initial experimentation with each strategy, the diversity measure is overestimated at the beginning of the exploration. For a given strategy s_j, instead of considering the set $E_j = \{\mathbf{y}_0, \mathbf{y}_1, ..., \mathbf{y}_{n_j}\}$, we consider the set $E' = \{\mathbf{y}_{-k}, \mathbf{y}_{-k+1}, ..., \mathbf{y}_0, ..., \mathbf{y}_n\}$, with k in $\mathbb{N}^+$. The set $\{\mathbf{y}_{-k}, \mathbf{y}_{-k+1}, ..., \mathbf{y}_{-1}\}$ is composed of fictitious points only available to the selecting strategy, that generate hyperballs that do not overlap with the observed effects. That way, the diversity of the strategy is overestimated during the $w + k$ first times it is selected. This also avoids having the first strategy selected unfairly preferred because it created the first observation, thus receiving the diversity of a full hyperball volume. We will use $k = 1$ in all strategies.

```
Algorithm 3: Adapt($w$, $\tau$)
Input:
    • $s_0, s_1, ..., s_{q-1}$, strategies.
    • $E = \{\mathbf{y}_0, \mathbf{y}_1, ..., \mathbf{y}_{n-1}\}$, a set of effects.
    • $\tau$, coverage threshold.
    • $w$, time window.
    • $\alpha$, ratio of random choice.
Result:
    • $s_j$, chosen strategy

if Random() $< \alpha$ then
    choose a random strategy.
else
    choose a strategy $s_j$ proportionally to its diversity $\mathit{diversity}_{\tau,w}(s_j, E)$.
```

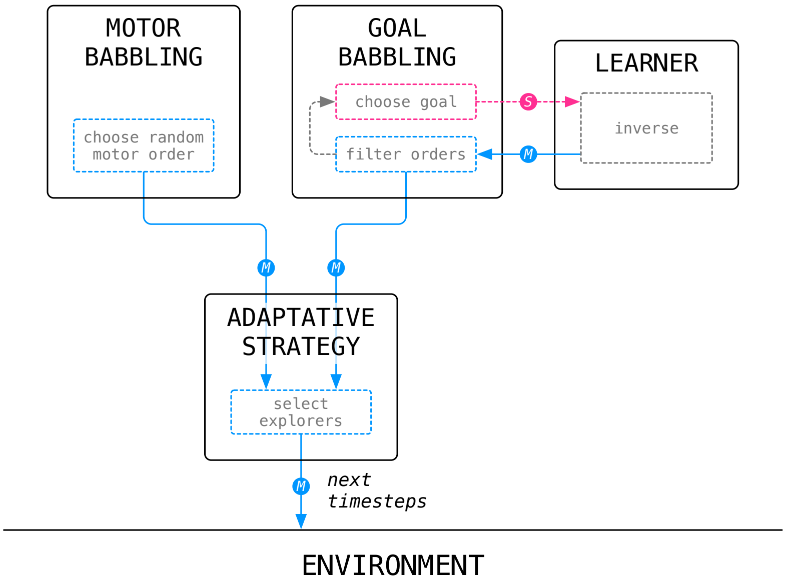


Figure 4.3: The architecture of the adaptive strategy with two explorers.

4.2.4 Experiment

We create an exploration strategy with the Adapt algorithm having the possibility to select between the random motor babbling strategy and the random goal babbling strategy of section 4.1. The explorer architecture is described Figure 4.3.

We run the adaptive strategy on a 20-joint arm, with each of the three learners studied in section 4.1. α is set to 0.1, the time window for the diversity evaluation to 50 timesteps and the coverage threshold[4] τ at 0.02.

In Figure 4.4, the results of the strategy are displayed. In all three learner configurations, the Adapt algorithm identifies and uses the correct strategies. When $d = 0.001$, the goal babbling strategy is inefficient in the beginning, and motor babbling is overwhelmingly used. Motor babbling diversity declines continually during the exploration, and in the later stage, is comparable to goal babbling. As a result, after 4000 timesteps, the two strategies are used roughly equally.

When $d = 0.05$, goal babbling and motor babbling produce the same diversity at the beginning, but goal babbling declines more slowly than motor babbling. As a result, goal babbling is used more and more as the exploration progresses, as it should be.

When $d = 0.5$, motor and goal babbling behave similarly—if d had been equal to 1.0, they would be the same strategy. During the early phase of the exploration, the

[4] It is both unsurprising and ironic that the adaptive strategy, whose purpose is to remove one parameter, the ratio of usage of motor babbling over goal babbling, in turns needs three parameters. However, those parameters are slightly easier to set at reasonable values.

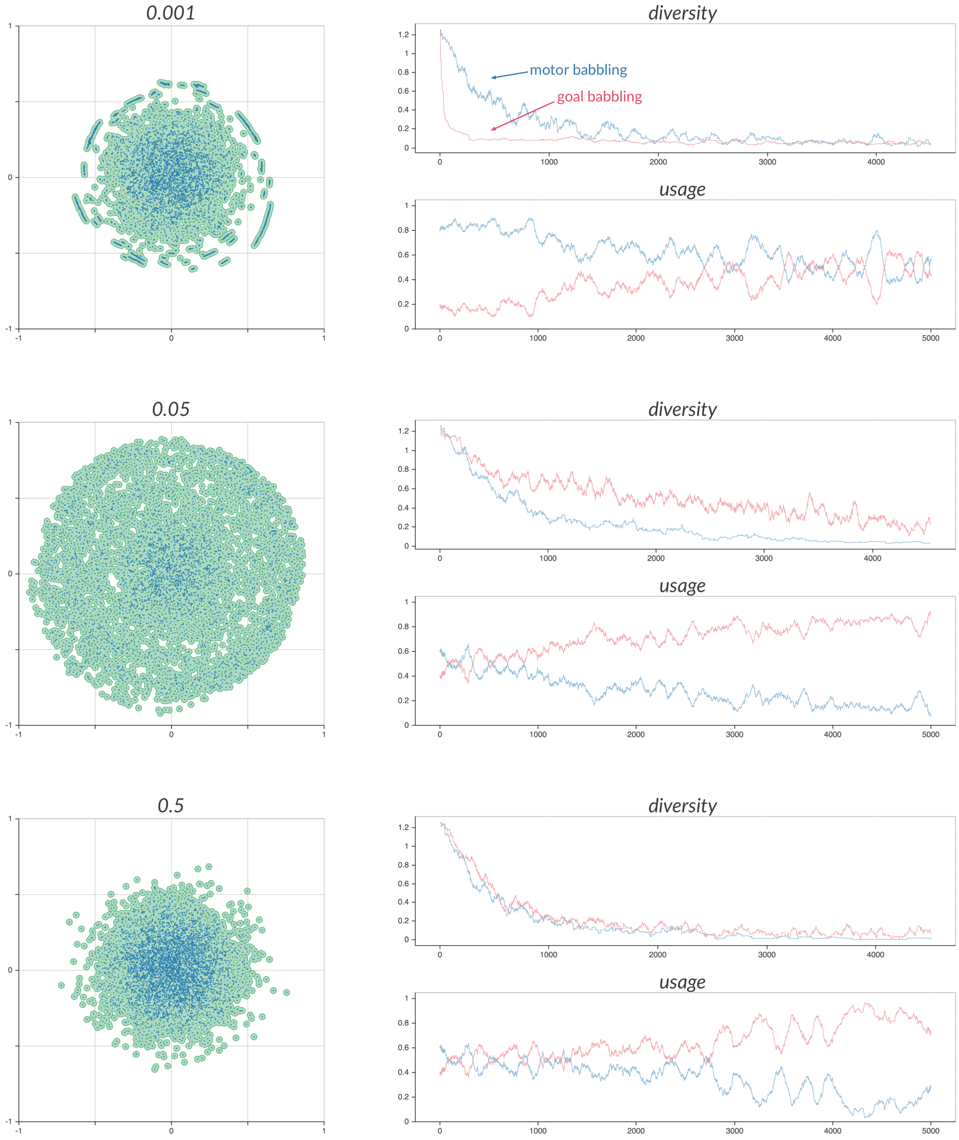


Figure 4.4: The Adapt algorithm correctly selects the best strategy in all three contexts. For each learner, three graphs are shown: the spread graph with the coverage area ($\tau = 0.02$), the diversity graph giving the diversity measure of each strategy in function of the timesteps, and the usage graph, showing how the strategies are effectively used. For the usage graph, the data at time t shows the percentage of use averaged over the surrounding 100 timesteps (50 before, 50 after). [source code]

Adapt algorithm does not distinguish between the two strategies. But in the later phase, goal babbling is able to provide an edge, however small, that is detectable by the Adapt algorithm. Goal babbling usage dominate after 1500 timesteps, and is used 80% of the time after 4000 timesteps.

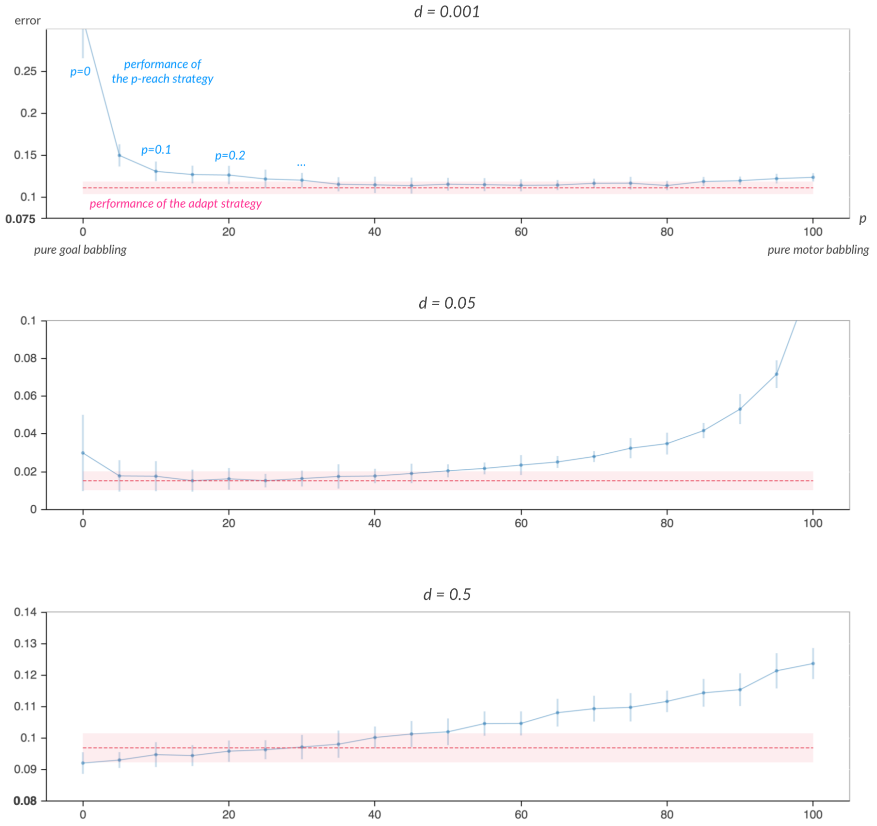


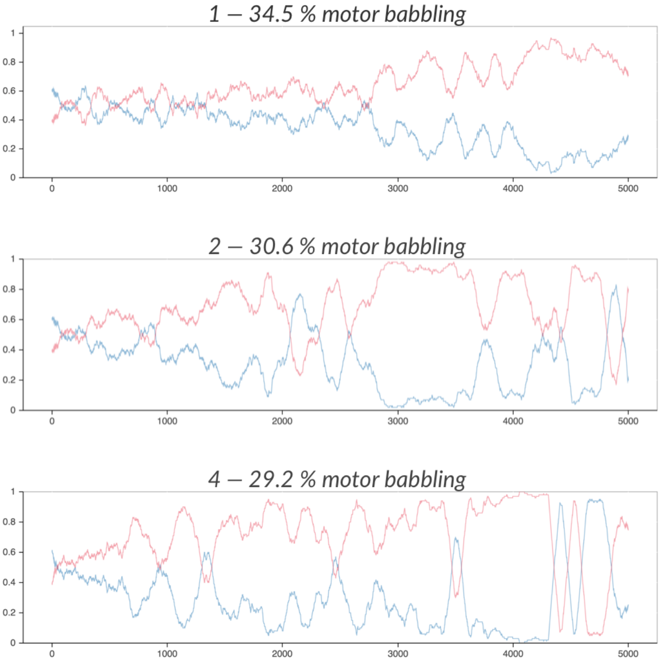
Figure 4.5: The Adapt algorithm performs well when strategies behave distinctly, and better than random with similar strategies. Each graph displays the performances showed Figure 4.2, with the performance of the adaptive strategy added as a dotted line (its standard deviations in displayed in light colour as well). Experiments were repeated 25 times. Note that not all the y-axis of the graphs begin at zero.

While the algorithm works qualitatively, it remains to be seen if this translates quantitatively. Figure 4.5 compares the error of the Adapt algorithm with the error of the fixed mixed strategies of section 4.1.

When goal babbling is much worse than motor babbling ($d = 0.001$) or when it is much better ($d = 0.05$), the Adapt algorithm manages performance on par with the best fixed mixture of strategies. When goal and motor strategy behave similarly, the adapt strategy is more conservative than the best case. This stems from the early stage of the exploration, when the motor babbling and goal babbling strategy are both effective, and hence both significantly used.

4.2.5 A Power Variation

A possible solution would be to increase the impact of the differences in diversity, by considering, for instance, the square of the diversity of a strategy instead. When $d = 0.5$, this approach does not work well, and only increases the usage instability of the strategies, as shown in Figure 4.6, where the strategies where chosen proportionally to their diversity, the square of the diversity, and the diversity to the fourth power.


1 — 34.5 % motor babbling

2 — 30.6 % motor babbling

4 — 29.2 % motor babbling

Figure 4.6: The modest decrease in motor babbling usage is accompanied by an increase in instability - strategy usage shifts suddenly, in a context where the exploration has mostly stabilized. The graphs represent the usage of the motor and goal babbling strategy (with $d = 0.5$), when they are chosen proportionally to their diversity value, the square of the diversity, and the diversity to the fourth power respectively. [source code]

As the proportionality of the selection shows weaknesses, a better method would probably be to use a simple soft-max selection rule, or using the more version offered by the Exp4 algorithm from Auer et al. (2002). Still, the performance is good-enough for now.

4.2.6 Grid Diversity

Computing the area of the union of hyperballs is not trivial ($\mathcal{O}(n \log n)$ in dimension 2, see appendix A for details), and computing the diversity requires to do it n times, with n the number of timesteps. We developed an alternative effect diversity measure based on a grid, computed in $\mathcal{O}(1)$, that produces similar results. Details and graphs are available in appendix B.

4.3 Adapting Reach

Abstract · We present a slightly more complex exploration architecture were some part are fixed and other adaptable. We show that the Adapt algorithm can balance different exploration strategies beyond motor and goal babbling.

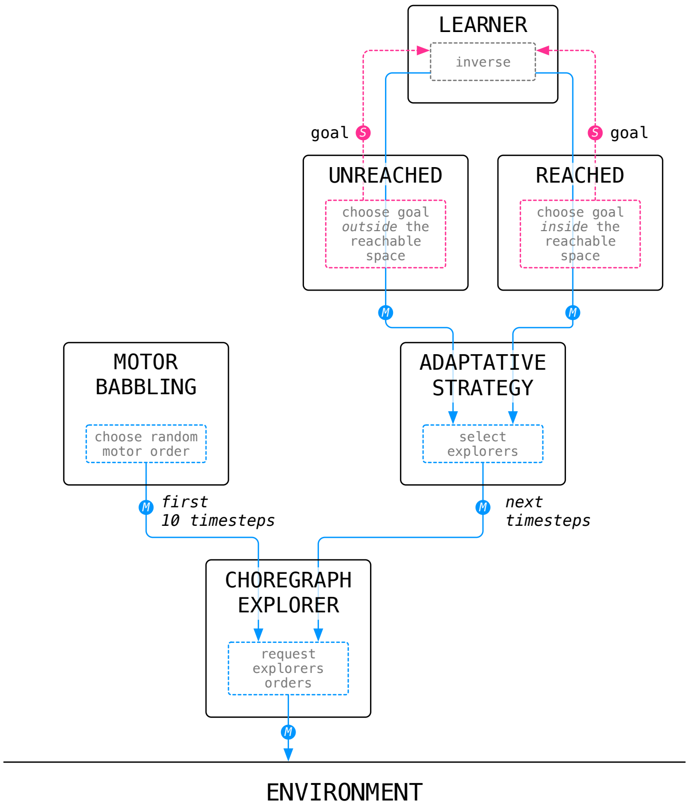


Figure 4.7: The architecture of the p-unreach strategy run by an adaptive strategy is an example of the hierarchical expressiveness of the explorers framework.

In section 3.3.2, we investigated the p-unreach strategy, and how a combination of goal set outside and inside the goal space can allow to balance the aggressiveness of the exploration. We now use the Adapt strategy to adjust the balance dynamically. The resulting architecture is described Figure 4.7. We force 10 initial random motor babbling steps, after which the Adapt strategy can choose between the unreach and the reached strategy. The learner is configured with a perturbation parameter $d = 0.05$.

We study this time the impact of considering different coverage threshold for the adapt strategy. Specifically, we consider $\tau = 0.02$, as we did in section 4.2, and $tau = 0.05$ and $tau = 0.1$. All the other parameters of the adapt strategy remains as they were set in section 4.2: $\alpha = 0.1$ and $w = 50$. The width of the cells the grid of both grid strategies is set to 0.05 along both dimensions.

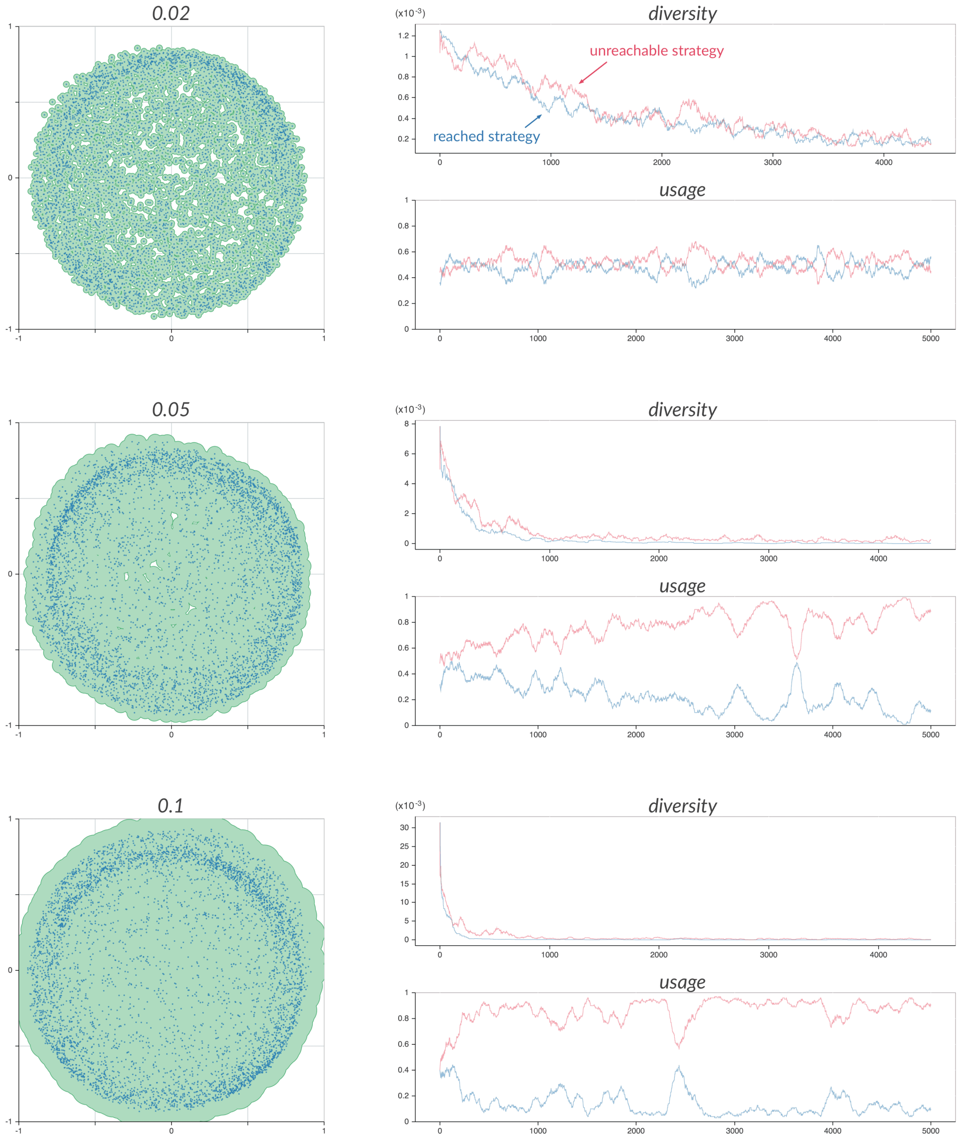


Figure 4.8: The coverage threshold parameter has a huge impact on the diversity measure, and, consequently, on the usage of the strategies. [source code]

In the results presented Figure 4.8, the impact of the coverage threshold parameter is clear: when it is high, coverage of the centre is quickly complete; the only sources of diversity then are found on the edges of the reached space, which favours the unreach strategy. Observations are clustered along the edges of the reachable space. When the threshold is low, the coverage of the centre takes much more timesteps, and the perturbations induced by the inverse model make the reached strategy competitive, yielding a much more balanced usage of the two strategies.

In other words, a high threshold favours aggressive exploration but yield poor diversity of observation, and a low threshold provides high diversity of observations, and favours a less aggressive exploration.

One could possibly obtain the best of both, a high diversity and an aggressive exploration—by implementing a developmental constraint that makes the threshold begin with a high value and lowers it during exploration.

Discussion

The ADAPT algorithm we presented, and the corresponding adaptive strategies we implemented demonstrate the reusability of simple exploration strategies to make better, more flexible ones. The diversity measure is, in many ways, rather crude, but it shows that discriminating between exploration strategy is definitely possible, and, advantageous. This work is related to previous works, and the general idea is not particularly new. It's application to exploration problem, and to a diversity measure, is, however.

In the experiments, we modified the exploration strategy of the agent. It would suggest that a same strategy, then, can adapt to different environments, with different complexities. This is the more important point, but this is not what the experiments established—further work is needed to establish environment independence directly and empirically. The algorithms could also benefit from being tried in different domains. As we argued in last chapter's discussion, our simplified two-dimensional work is hardly convincing of anything else than itself.

Additionally, From the experiments we conducted, it is unclear how the ADAPT algorithm will scale with the number of strategies. As more strategies are available, either more time will have to be devoted to exploratory sampling of bad strategies, or strategies will be less accurately evaluated overall. This is the classic exploration/exploitation trade-off.

We imposed a constraint of agnosticity over the internal working of the selected strategies. However, the usefulness of some strategies typically decreases with time, such as motor babbling. Taking into account on how well each exploration strategy usually performs—perhaps from a prior derived from the experience gathered from exploring similar environments—could improve the performances of the ADAPT algorithm, and avoid to rediscover everything all the time. Coincidentally, we will be attacking that very subject in the next chapter.

Part Two

REUSE

A young corvide bird, confronted with an object it has never seen, runs through practically all of its behavioral patterns, except social and sexual ones.

Konard Lorenz

5
Reuse: The Basic Idea

To illustrate this, let's consider a pair of two-dimensional arms with the same number of joints. The first arm has same-length links totalling one meter, and the environment returns the Cartesian position of the end-effector, as in chapter 0. The second arm has links such that, going from the base to the end-effector, each link is 0.9 times smaller than the previous one, while the total length of the arm remains one meter; this arm also returns the position of the end-effector, but using *polar* coordinates.

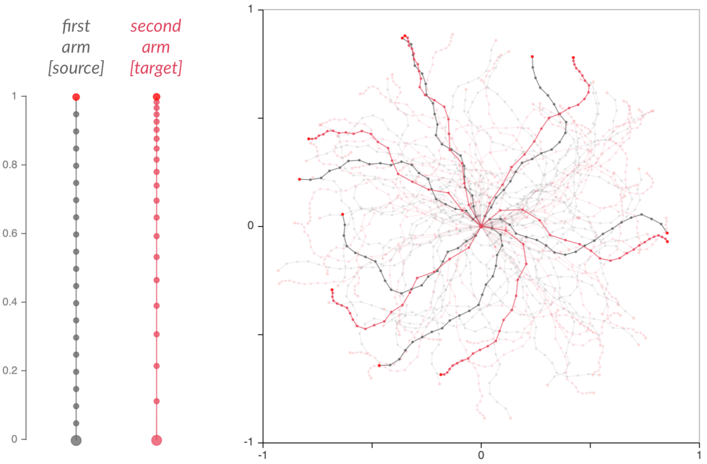


Figure 5.1: When executing the same command on both arms, the position of the end-effector is significantly different most of the time. Here depicted are 50 pairs of executions of the same motor command on the two 20-joint arms, five of which that are highlighted. [source code]

The two arms share the same number of joints with the same available ranges ($\pm 150°$): they have the same motor space. However, because the lengths of the links are different, most motor commands will results in a different position for the end-effector, as shown in Figure 5.1. And because the positions are expressed in two different coordinate systems, the inverse model of one arm is difficult to exploit on the other arm, without having, or learning, a mapping between the coordinate systems.

As in the first part, the agent views the two arms as black-boxes, and has no information about the relation between them. In fact, because the sensory feedback channels are not labelled, the two arms are indistinguishable from one another before any interaction is performed.

The Basic Idea

Let's assume that the first arm has been explored. The idea behind the reuse method is to bootstrap the exploration of the second arm using the exploration history of the first arm.

In all the exploration strategies that we have considered so far, the initial observations were generated through random motor babbling. When using the reuse method, instead of generating the initial motor commands randomly, we instead choose motor commands that were executed during the exploration of the first arm. This is possible since both arms share the same motor space: the motor commands are compatible.

Since the reused commands are executed on the second arm early in exploration, during the motor babbling phase, we cannot rely on acquired knowledge about the second arm in order to choose which motor commands to reuse. Instead, we rely—unsurprisingly—on an intrinsic characteristic of the exploration history of the first arm, one that occupied us already during much of the first part of this thesis: its *diversity*.

We choose a set of motor commands from the first arm that produced a set of effects that has a high diversity, and execute them on the second arm. Because the internal dynamic of the two systems are not too dissimilar, this is likely to create a diversity of effects early in the exploration of the second arm. In other words, the second arm leverages the structure of the exploration of the first arm. In particular, it increases the probability to produce observations in low-redundancy area of the sensorimotor space, that required extensive exploration to be discovered during the exploration of the first arm.

The main condition for the reuse method to be applicable is that the two environments share the same motor space—or at least that the intersection of their motor spaces is not empty. The reuse method does not impose any other condition on the relation between the two tasks. In particular, it is not constrained by differences in sensory modalities, or differences in learning algorithms. And because the number of random motor babbling step that are replaced by reuse steps is configurable, the impact reuse has on the exploration of the second arm can be regulated as necessary.

The condition on the motor spaces is not particularly hindering either. It is already verified in many platforms at the lowest level: the actuation interface is usually as stable as the body and wiring of the robot (an the same goes for biological organisms). When considering higher level of abstraction for motor commands, a higher discrepancy can be expected between tasks's action space. Even in those cases, whether the overlap between the action spaces of the tasks is total, high, low or null, detecting the applicability of reuse is immediate, and reuse can be used opportunistically in conjunction with other exploration strategies.

Experiment

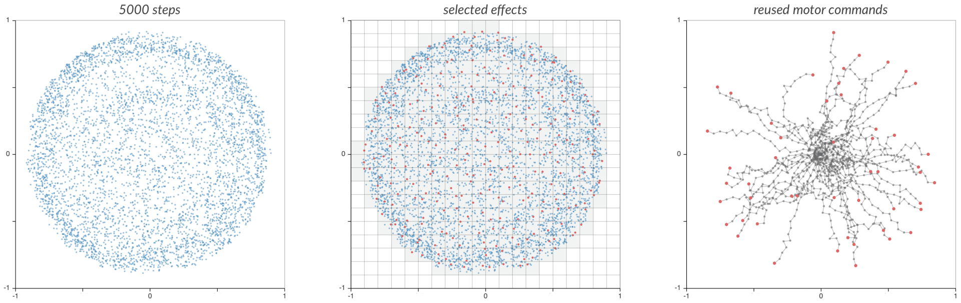


Figure 5.2: After the 5000 steps of the exploration on the first arm have been carried out, a grid is applied on the sensory space. Here, we choose (at random) one effect per cell (in red). Out of those effects, only 50 will be selected and their motor command reused: this is the rightmost graph. [source code]

The exploration on the first arm is conducted over 5000 steps, using the random motor babbling strategy for the first 50 steps, and then using random goal babbling. Both arms have 20 joints.

We implement the reuse method by laying a grid over the sensory space of the first arm at the end of the exploration. The set of reused motor commands is selected by repeatedly choosing a non-empty cell at random and drawing without replacement an effect from that cell; the chosen motor commands are the ones that produced the selected effects. This process is illustrated Figure 5.2.

As can be seen Figure 5.3, Reuse-backed exploration access low-redundancy areas of the sensorimotor space early. The exploration of the second arm environment begins by reexecuting 50 motor commands from the first arm exploration trajectory. At the 400 timesteps mark, the difference between reuse and a classical goal babbling strategy is significant. In particular, the reuse exploration has spread near the edges of the reachable space while goal babbling is still far from them. 5000 steps into the

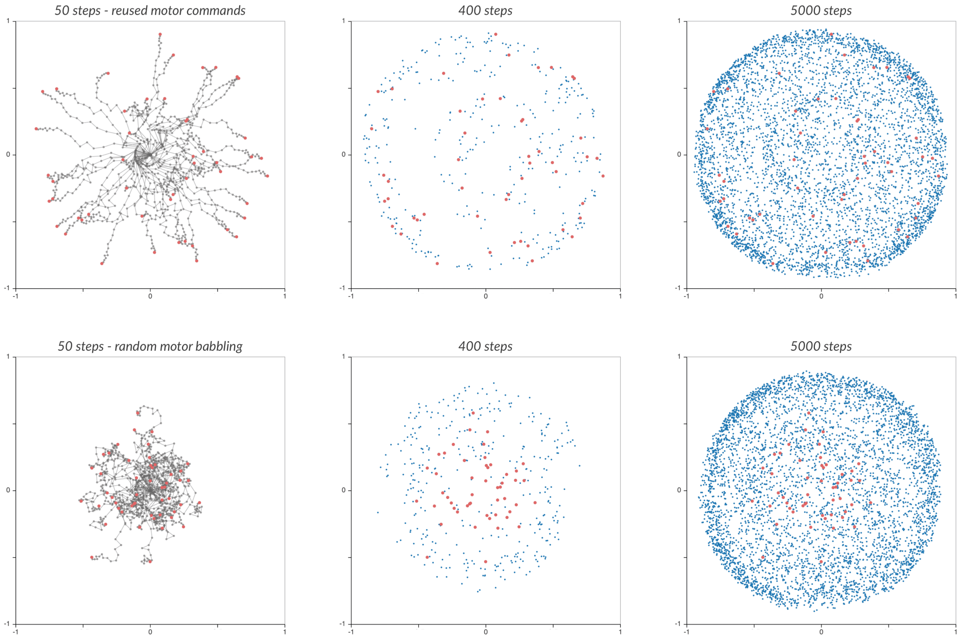


Figure 5.3: The benefits of the reuse exploration manifest mostly early in the exploration. Here, the top-most row shows an goal babbling exploration backed by reuse, while the bottom row shows a regular goal babbling exploration. In red, in the case of reuse, the effects produced by the reused motor commands, and in the case of regular random goal babbling, the effects produced during the 50-steps random motor babbling initial phase. [source code]

exploration, the differences vanish—the better final coverage of the reuse strategy on this specific example is not indicative of a general tendency.

Randomness Or Diversity?

So far, the presentation we have made of the reuse method—although perfectly correct—has been somewhat disingenuous. Indeed, by focusing the attention on how the selection of motor commands was driven by effect diversity, we have implied that it played a major role in the performance of the algorithm. What about choosing motor commands to reuse at random?

In Figure 5.4, the pattern of exploration offers no discernible difference to the one of the top row of Figure 5.3, where the set of reused motor commands was explicitly crafted to contain a diversity of produced effects on the first arm.

The explanation, of course, is simple. The goal babbling algorithm used in the exploration of the first arm already produced a distribution of motor commands that produced an approximately uniform distribution of effects on the reachable space. This was how we motivated goal babbling over motor babbling in chapter 0. Explicitly ensuring diversity is then redundant, and does not provide any advantage. This allows

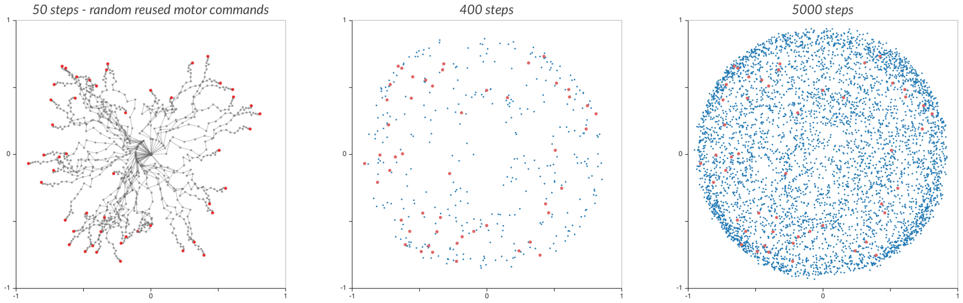


Figure 5.4: Reusing random motor commands seems as efficient as using commands selected for their diversity. In this exploration, the motor commands were chosen randomly from the first exploration, without taking into account how the effects they produced relate to one another. [source code]

to understand better how the *reuse* method works, whether the selection of reused commands in random or not. Reuse leverages the reduction in the heterogeneity of the sensorimotor redundancy of the set of observations of the first exploration.

What happens, then, if the exploration of the first task does not reduce the heterogeneity of the redundancy? What if, for instance, the first exploration is driven by a pure random motor babbling strategy? Surely, we can't expect any advantage from using random reuse then. But what about diversity reuse? *A contrario*, can diversity reuse still be justified, when the exploration of the first arm is good-enough? All these questions will be answered in the next chapters.

For now, let's conclude that the reuse method is simple, requires a condition often already verified in existing robots, and improve significantly early exploration.

6
The Reuse Framework

In this chapter we review the existing literature on transfer learning, motivate our approach and formalize the reuse method.

6.1 Transfer Learning

Classical machine learning considers scenarios where a learner is trained to make predictions from data on a specific problem. Transfer learning (Thrun and Pratt 1998; Taylor and Stone 2009; Pan and Yang 2010) considers how the experience gained in one learning scenario can be used in another scenario to improve performance. The reuse method is an instance of transfer learning.

6.1.1 A Short Motivational Overview

Transfer learning was originally motivated by the cost of labelling instances for classification tasks: for instance, a classifier would be created to detect horses in pictures, and then the need to create a classifier to detect zebras would arise. To train the classifier, thousands of zebra pictures would have to be gathered and manually labelled. If somehow the knowledge contained in the horse classifier, or the labelled horse pictures, could be leveraged to create the zebra classifier, given their obvious similarities, then the number of pictures of zebras that would need to be labelled to achieve a given performance would be reduced. Or rather than a different animal, we may want to create classifiers for a different medium. Can the horse picture classifier help train a classifier detecting horse in videos? It may also be that this is not so much the cost of labels that makes transfer learning desirable, but the scarcity of the data: there might not be enough pictures of zebras or videos of horses available to create a good-enough classifier[1].

Likewise, robots can benefit from transfer methods. As each interaction is costly, any interaction that can be avoided by reusing previously acquired experience represent a significant information gain. Moreover, sometimes the data necessary to solve a problem is not present in the environment, and must be marshalled from past experiences. In other contexts, there is no time to learn: a useful behaviour is expected immediately. Additionally, complex tasks often require the acquisition of a number of subskills before being able to deal with them. Those subskills may be more easily learned in simpler contexts, from which the acquired behaviour must be transferred on the complex task.

Finally, current robots suffer from an absence of good priors when they start learning. Without any knowledge of the world, without any common sense, tasks that are trivial for humans become frustratingly difficult to implement in robots. Transfer learning is part of the answer to this, and, in particular, to the question of the origins of Bayesian priors.

Another important instance of transfer learning is *dataset shift* (Quiñonero-Candela et al. 2008). Dataset shift happens when one cannot assume that the testing data for a learning algorithm will have the same distribution as the training data. For instance, after six-months of operation, a bike-sharing operator wants to build a model to predict how the bikes will be borrowed next month, and where empty and full stations will be located. But bike usages change with the seasons, and no record exists for this month last year. And, the new bike lanes opened by the city last week and the increased popularity of the program have probably to be taken into account to. In this case, we talk about *dataset shift*. The task is the same between the training and the application, but the training data is sampled from a situation different than the one we want to apply it to. Another example is an airport trying to build a classifier to detect smugglers amongst the passengers. Evolution of the value of smuggled goods or changes in repression will modify the behaviour and distribution of smugglers: an increase in the value of the smuggled goods accompanied with more lenient laws will make a portion of the normal population become smugglers. When building a classifier based on historical data, one must take these factors into account.

Online, incremental learning algorithms make robots less affected by dataset shift, which manifests itself when the learning data is divorced in significant ways from the

[1] Interestingly, the sensorimotor loop of a robot can be approximated to the labelling process: motor commands are labelled by the environment: the labels are the feedback sensory signals (this is assuming that the motor commands and sensory stimulation are discrete and can be unambiguously matched with one another). Yet, the fundamental differences between the two highlighted in chapter 1 make most transfer learning methods for classification incompatible with an embodied context.

current context. Robots should expect continuous dataset shift, and constantly update their behaviour to changing conditions. Nevertheless, abrupt changes do happen (being unboxed in someone's home from the factory for instance). And learning a task does not necessarily happen continuously. Being able to deal with dataset shift means being able to resume the learning of a task that was begun in the past, by taking into account that the observation distribution may have changed.

One has to acknowledge that transfer in humans is not a settled topic (Billing 2007). Thorndike and Woodworth (1901) was one of the first to study the phenomenon, and in a 1923 study (Thorndike 1923) famously failed to find a strong causal link between learning Latin and improving one's mastery of the English language. Other studies have reported similar results, for instance on the benefits of learning programming (Pea et al. 1984; Salomon et al. 1987) (although other studies did find instances of positive transfer (Lehrer et al. 1988; Clements et al. 1984)), or even learning how to read and write (Scribner 1981). As Billing (2007) points out, the studies, re-examined today, are usually regarded as not providing good evidence against transfer. Overall, unless considering very narrow definitions of transfer, evidence seems in favour of transfer. In humans, transfer takes multiple forms, but is not systematic; in particular, it is highly contingent on the environmental conditions of learning.

6.1.2 A Computational Definition

In Pan and Yang (2010), a learning scenario is defined as the combination of a domain $\mathcal{D} = \{\mathcal{X}, P(X)\}$ and a learning task $\mathcal{T} = \{\mathcal{Y}, f(\cdot)\}$. The domain is composed of a feature space $\mathcal{X}$ and a marginal probability distribution $P(X)$ with $X \in \mathcal{X}$. In the case of our sensorimotor scenario, $\mathcal{X}$ would be the set of motor features, and $P(X)$ the uniform distribution, since we can choose and execute any motor command we want. The learning task is composed of a label space $\mathcal{Y}$ and an objective predictive function $f(\cdot) : X \mapsto \mathcal{Y}$. $\mathcal{Y}$ corresponds to the sensory space, and $f(\cdot)$ to environmental feedback. Given these notations, Pan and Yang (2010) defines transfer learning as:

> *Given a source domain $\mathcal{D}_S$ and learning task $\mathcal{T}_S$, a target domain $\mathcal{D}_T$ and learning task $\mathcal{T}_T$, transfer learning aims to help improve the learning of the target predictive function $f_T(\cdot)$ in $\mathcal{D}_T$ using the knowledge in $\mathcal{D}_S$ and $\mathcal{T}_S$, where $\mathcal{D}_S \neq \mathcal{D}_T$, or $\mathcal{T}_S \neq \mathcal{T}_T$.*

In the example of the previous chapter, the motor spaces, and their uniform marginal probability distribution are identical. The tasks, however are different; we are in the $\mathcal{D}_S = \mathcal{D}_T$ but $\mathcal{T}_S \neq \mathcal{T}_T$ case[2].

[2] In fact, the algorithms we present straightforwardly extend to more general case where the intersection of the sets of motor commands from the two environments in not null (i.e., when the Bhattacharyya coefficient of $P(X_S)$ and $P(X_T)$ is non-zero; Bhattacharyya (1943)). Still, we will assume $\mathcal{D}_S = \mathcal{D}_T$ unless otherwise stated.

Notice that, while the domain and the task are easily distinguished conceptually, in practice it is often much less clear, and it may entirely depends on how the problem is formalized; the concept of domain and task usually depends on the learning framework. In reinforcement learning for instance, Fernández et al. (2006b) distinguishes the Markov decision process (MDP) as the domain, while the task corresponds to the reward function. As a result, many transfer learning techniques are specific to a given learning framework.

More simply, transfer learning happens as soon as *experience acquired while learning a task can influence how another is learned*. This does not have to be a direct influence. If any of the cognitive changes that have been created while learning a task influence how another task is learned, transfer learning happened. For this reason, transfer learning is present in human and animal learning not only in conscious and specific cases, but all the time, as a intrinsic property of neural learning. Furthermore, any reasonable cognitive architecture for cumulative learning probably implicitly features transfer learning in one form or another, because not doing so would impose strict cognitive isolation between tasks.

Of course, and again, transfer learning is only defined insofar as *tasks* are clearly defined. While most transfer learning experiments provide tasks, and assume that they are related, the problem of recognizing and discovering tasks in an environment is a non-trivial issue for transfer learning that is mostly unaddressed in the literature.

6.1.3 Benefits of Transfer

There are several expected benefits of transfer on learning performance (Taylor and Stone 2009), that acts as many ways to evaluate its impact.

- A improved initial performance. If the transfer occurs before learning in the second task has started, a jumpstart, i.e. a difference in initial performance might be observed between the target task with and without transfer.

- In practical settings, a behaviour might become useful if the performance reaches a specific threshold (for instance, the positional precision of an end effector), and transfer learning may help reach this threshold faster.

- Closely related, the performance might be evaluated at a given time after learning started, and transfer learning may improve the evaluated performance.

- The average performance over a time window may also be increased. In a reinforcement learning context, if the time window extends to the whole learning duration, this represents the total cumulated reward.

- Finally, transfer learning might change the asymptotic performance of the learning algorithm.

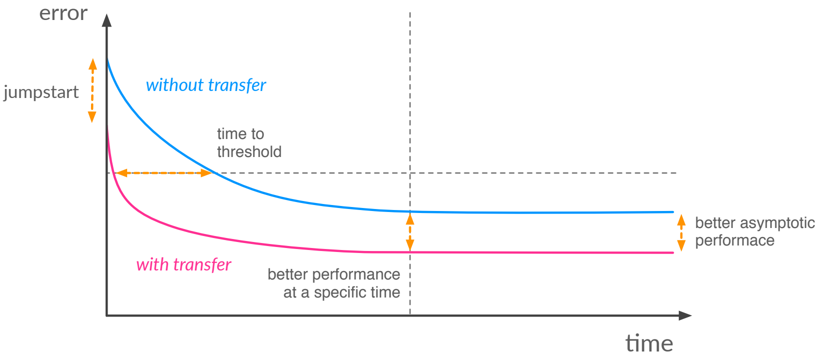


Figure 6.1: Four potential benefits of reuse.

These are some performance-related impacts of reuse, that are mainly valid in the case of monotonically increasing performance. In complex scenarios, the impact of transfer learning might be more complex to measure.

Transfer learning is not necessarily beneficial. If the tasks are too dissimilar, the bias introduced by the source task might decrease performance: this is *negative transfer*. For instance, a source tasks might direct learning toward a familiar region of the learning space in the target task. This would provide a positive jumpstart in performance, but might produce a worse asymptotic performance if this region of the learning space limits the quality of the solutions.

6.1.4 Ways and Means

To conduct transfer from one task to another, one must identify aspects of the source task that might benefit the learning of the target task, and devise a way to transfer them between tasks. Additionally, one must consider that transfer is not always beneficial, and thus decide, in a specific context, if transfer should be carried out. In a cumulative learning setting, one can furthermore expect that more than one source task is available: choosing from which source task to transfer also becomes an issue. This corresponds to the three questions identified by Pan and Yang (2010): *what* to transfer, *how* to transfer and *when* to transfer, and we add *from where* to transfer.

Different methods of transfer have been proposed. Some share instances of the dataset across tasks (Shimodaira 2000; Quiñonero-Candela et al. 2008; Fan et al. 2005; Liao et al. 2005; Huang, Gretton et al. 2006; Dai, Yang et al. 2007; Jiang et al. 2007), in particular in the case of dataset shift. These methods typically use importance sampling and instance reweighting to adapt the dataset. Liao et al. (2005) in particular propose, in the context of logistic regression, an active learning algorithm that chooses which elements of the target task to label, if no label are provided at the start. The active choice is driven by reducing the uncertainty (the variance) on the classifier parameters.

Others share or create common feature representations (Blitzer et al. 2006; Daumé 2007; Dai, Xue et al. 2007; Xing et al. 2007; Wang, Song et al. 2008; Pan, Kwok et al. 2008; Zeng et al. 2012). Zeng et al. (2012) uses kernel learning methods on related tasks, and then applies the low-dimension representations thus created on the target task, where the scarcity of the data does not allow for such methods to apply. In contrast, Daumé (2007) augments the target data with the learned source task features.

Another class of methods share model parameters or models across tasks (Lawrence et al. 2004; Raina et al. 2006; Bonilla et al. 2008; Gao et al. 2008; Chai et al. 2009). As Raina et al. (2006) points out, such methods are in particular employed to set the priors in a Bayesian setting. Gao et al. (2008) proposes to combine multiple models into a locally weighted ensemble of model for the target tasks. Even conflicting models can be combined, the relevance of each of a local area of the target dataset is evaluated, and the more relevant models receive a greater weight over that area.

Most methods we have discussed so far deal with classification and regression. Our account is far from exhaustive, and the interested reader can consult Thrun and Pratt (1998) and Pan and Yang (2010) for reviews, and Quiñonero-Candela et al. (2008) more specifically for dataset shift. In the next section, we discuss another class of transfer learning methods: transfer in reinforcement learning.

6.1.5 Transfer in Reinforcement Learning

In the reinforcement learning framework, the Markov Decision Process representation shared by the different algorithms lays out clear distinctions between the transfer methods. In particular, we can distinguish between methods that assume that the state-action space is shared across the source and target tasks, and those that don't. Among those two groups, we can usually find distinctions of instance transfer, representation transfer and parameter transfer to classify their respective methods.

Among the methods that share the state-space between source and target task, a common technique is to learn and discover *options* (Sutton et al. 1999; Precup 2000) in the source task, and to transfers those policies to the target task. Options are useful for navigating the state space. Learning them in the source task provides the agent in the second task automatized ways to different areas of the state space (Perkins et al. 1999; Bernstein 1999; Şimşek et al. 2005). Asadi et al. (2007) propose to identify bottleneck states in the state-space, and to construct subgoals (and correspond partial policies), based on the structure of the task space rather than the reward.

An interesting method comes from Sherstov et al. (2005), that create a set of task from a source task, and prune the action space from any action that is not optimal in at least one task of the set. The diversity of the set of tasks creates a filter that is used to reduce diversity in the set of actions.

Other methods do not consider the state-space necessarily fixed, but usually require that an expert mapping between the tasks. For instance, the work of Fernández et al. (2006b), *policy reuse*, which upholds ideas that are very close to the reuse algorithm, constrains the state and action space to be similar Fernández et al. (2006b), or that a mapping between the state and action be provided by an expert (Fernández et al. 2006a).

However, the reuse method does too for the action space: it requires that the motor space remains the same. And because our environment is one-step, episodic, the constraints that the state space stays the same is always guaranteed.

The idea behind *policy reuse* is to reuse policies across tasks, in the context of the RL framework. The tasks are composed of an MDP and a reward function. When a task is sufficiently novel, it is stored in a policy library, ready to be reused on new compatible tasks. When the policy learned from a task is novel compared to existing policies in the library, it is added to it.

However, policy reuse by itself does not provide a mechanism to generate diversity. The creation of new tasks is not the prerogative of the algorithm. As with reinforcement learning, policy reuse has only been applied to discrete or discretized spaces. And the similarity between two strategies is tied to the reward they bring. For all these reasons, our work is singularly different from Fernández et al. (2006b,a).

Again, our account is illustrative rather than exhaustive. The readers will find detailed surveys in Taylor and Stone (2009) and Lazaric (2012).

6.1.6 Brief Motivation

Many transfer methods try to find the mapping between the source and target task. Even in simple examples, such as our arm perceiving the world in polar coordinates, this represent a difficult challenge. One that is probably not often necessary to solve.

Rather than considering the functional mapping between the two tasks, we consider the diversity mapping, which is much more robust and much simpler. We assume that a set of motor commands that produce a diversity of effect in one task has a higher probability to generate a diversity of effect in another task, than a less diverse set of effects. Although obviously one can find counterexamples, this assumption is verified in many practical situations.

An interesting application for the reuse method would be team of identical robots (Waibel et al. 2011). Those teams of robots are projected to be connected to one another and share experience data amongst the population. Reusing motor commands between robots of a same population is possible since they share the same body, and is desirable, because it introduces no bias on the target robot: it avoids the representation trappings, and does not try to force the same ontology on all elements of the team, which is both a significant loss of diversity, and can disconnect the knowledge of the robot from its direct experience.

6.2 The Reuse Algorithm

Our method is organized around three algorithms. The first, Explore(), describes the learning and exploration of the source task and has been described in section 3.1. The second, Transfer(), is applied at the end of the learning of the source task, and produces the motor commands to be transferred to the target task. The third, Reuse(), controls how the transferred data impacts the exploration algorithm in the target task.

6.2.1 Processing the Trajectory

For each interaction in the second task, the learning algorithm can request reusing a motor command from the first task rather than doing random motor babbling. Goal babbling behaviour is unaffected. Our reuse algorithm defines which motor command is transferred when such a request is made.

The whole assumption behind reexecuting a set of motor commands from a previous task that generated a diverse set of effects in the past task, is that they might generate a variety of effects in the current task as well, hence bootstrapping the model with good observations. Of course, this assumption hinges on the fact that the two tasks are sufficiently similar.

In order to generate a sequence of motor commands that generated a diverse set of effects, we reuse the grid of the goal babbling algorithm, and assign each cell with a bin. In this bin, we put the motor command of every effect that belong to the cell. When a motor command is requested by the exploration algorithm, we choose a random, non-empty, bin and draw, without replacement, a random motor command from the bin. This procedure is codified in Algorithm Transfer.

This procedure has a low computational cost, and only transfer structured set of motor commands. No sensory data is shared across tasks, hence the target task never tries to use the forward or inverse model of the source task. This particular method as the added advantage that the structured set of motor command can be computed before knowing about the second task, and be used even if the first task has been forgotten.

6.2.2 Target Exploration

We modify Algorithm 2 to replace the call to MotorBabbling() by a probabilistic call to ReuseBabbling() and MotorBabbling(), according to a probability p_{reuse}, producing the Reuse algorithm. The resulting exploration architecture is illustrated Figure 6.2.

Algorithm 4: TRANSFER(ξ_A)

```
Input  : $\xi_A = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{0 \leq i \leq n_A}$, exploration trajectory.
Output: $\mathcal{B}$, a set of motor commands bins.

$\mathcal{B}$ = empty set
Divide $S_A$ into a set of regions $\mathcal{R}$
for $R \in \mathcal{R}$ do
    for $(\mathbf{x}_i, \mathbf{y}_i) \in \xi$, with $y_i \in R$ do
        add $\mathbf{x}_i$ to $b_R$
    add $b_R$ to $\mathcal{B}$
```

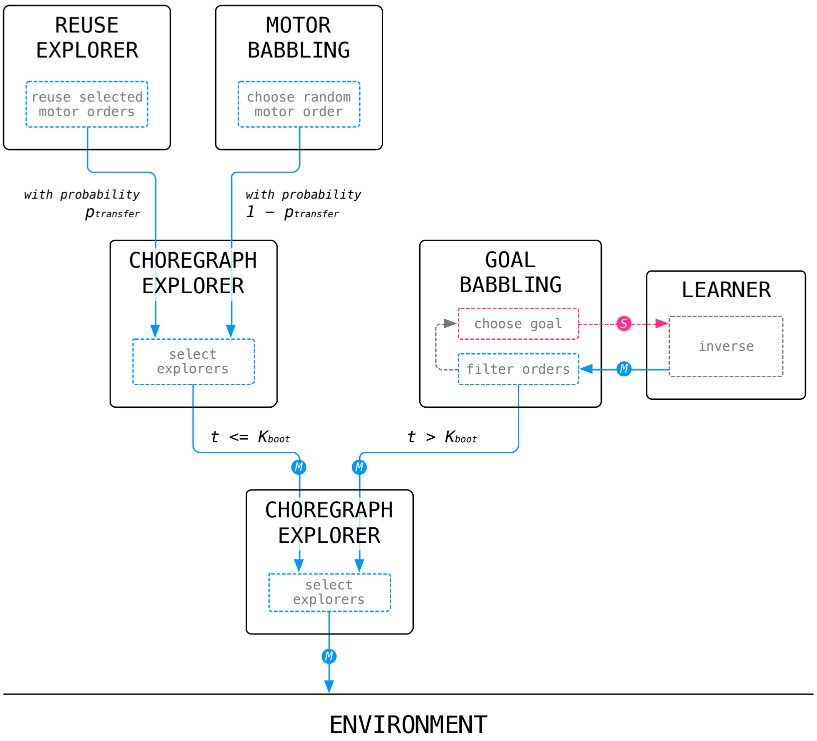


Figure 6.2: The Reuse exploration architecture.

Algorithm 5: REUSE(B, $\mathcal{B}$, K_{boot}, p_{goal}, p_{reuse}, p_{random})

```
Input   :
    • $B = (f_B, n_B)$, target task.
    • $\mathcal{B}$, set of bins of motor commands.
    • $K_{\text{boot}}$, duration of bootstrapping.
    • $p_{\text{reuse}}$, ratio of transfer motor babbling.
Output:
    • $\xi_B = \{\mathbf{x}_i, \mathbf{y}_i\}_{0 \leq i \leq n_B} \in (S_B \times M_B)^{n_B}$, exploration trajectory.

$\xi_B \leftarrow []$
for $t$ from $0$ to $n_B$ do
    if $t \leq K_{boot}$ then
        if RANDOM() $\leq p_{reuse}$ then
            $\mathbf{x}_t$ = REUSEBABBLING($B$, $\mathcal{B}$)
        else
            $\mathbf{x}_t$ = MOTORBABBLING($B$)
    else
        $\mathbf{x}_t$ = GOALBABBLING($B, \xi_B$)
    $\mathbf{y}_t \leftarrow f_B(\mathbf{x}_t)$ // execute the command
    add $(\mathbf{x}_t, \mathbf{y}_t)$ to $\xi_B$

REUSEBABBLING($B$, $\mathcal{B}$)
    if at least one bin of $\mathcal{B}$ is not empty then
        choose a non-empty bin $b_R$ of $\mathcal{B}$ randomly.
        draw $\mathbf{x}_t$ from $b_R$ without replacement
        return $\mathbf{x}_t$
    else
        return MOTORBABBLING($B$)
```

6.3 Diversity Reuse versus Random Reuse

In chapter 5, we presented a example of utilization of the reuse algorithm on two two-dimensional arm environments. The source environment was a 20-joint two-dimensional arm with all segments of the same size, while in the target environment, the arm had segments of decreasing length in the proximo-distal direction. Although we observed that using reuse provided an improvement in exploration on an specific instance, we didn't provide robust quantitative evidence. More over, that selecting the motor commands to be reused with diversity or randomly did not provide a qualitative difference. We address these two points now.

In chapter 5, the explorer on the source and the target task had a bootstrapping period of 50 timesteps. During this period, pure random motor babbling and pure reuse were done in the source task and the target task respectively. The rest of the time, the exploration was done using random goal babbling with a perturbation parameter equal to 0.05.

Exploiting Random Motor Babbling with Diversity

Figure 6.3 shows the coverage performance of the target task with and without reuse, using diversity reuse and random reuse. The performance without reuse is not the performance of the source task (the same-length links task); it is the performance of the target task (the decreasing-length links task) with the 50 timesteps of reuse replaced by 50 timesteps of random motor babbling, as we would proceed if a source task was not available[3]. Because the differences in performance are sometimes small in this section, all experiments have been run 100 times.

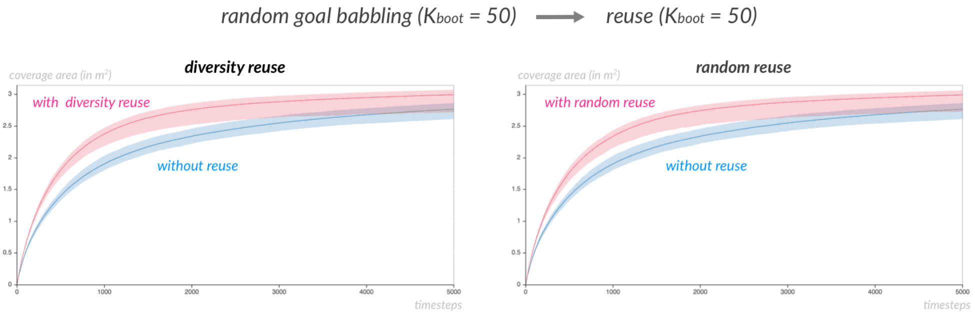


Figure 6.3: The observation made in chapter 6.3 is verified, there is no difference between using diversity reuse or random reuse in this specific example. [source code]

[3] The presentation of those results, with the performance target task with and with reuse in pink and blue respectively will remain the same throughout all chapters of the second part. Should you have a grayscale version of this document, in this chapter the performance with reuse is always superior, sometimes barely, to the one without, at t = 5000.

Figure 6.3 confirms the result of the previous chapter: in this particular instance, no quantitative difference exists between the performance of diversity reuse and random reuse (the curves almost match perfectly, with diversity non-significantly eking out at t = 1000).

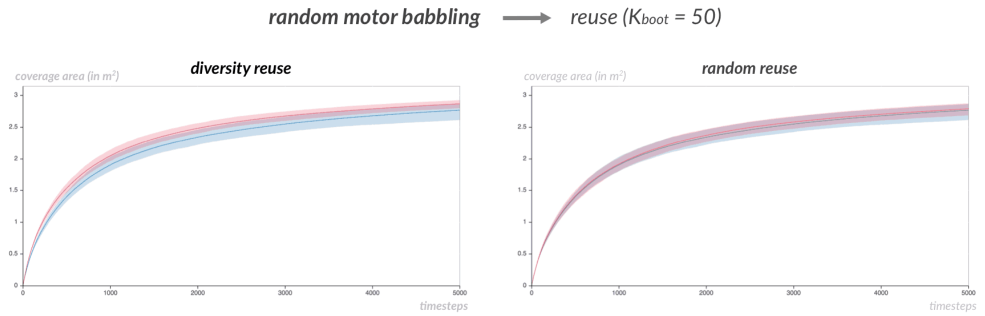


Figure 6.4: Diversity reuse is able to exploit a set of observation generated by random motor babbling, but the improvement in performance is small. [source code]

To show evidence that diversity reuse provides an advantage over random reuse, we consider a scenario where random reuse cannot bring any performance gain: when the source task is pure motor babbling. As shown in Figure 6.4, diversity reuse makes a difference, if a small one. In fact, it illustrates how diversity reuse is able to exploit an exploration that has no particular beneficial structure.

Increasing the Reuse Duration

One explanation of diversity reuse and random reuse not displaying different performance in Figure 6.3 is that over 50 motor commands, a random selection will generate as much diversity as an explicit diversity-driven approach. Over a longer period, random reuse will select similar motor commands from high density areas of the exploration, while diversity reuse will select motor commands uniformly over the sensory space, providing a performance improvement.

To test that hypothesis, we extend the reuse period from 50 timesteps to 500 timesteps. Figure 6.5 shows that this does not provide any significant improvement (the curves almost match perfectly. The standard deviation is slightly better for diversity reuse at t = 1000).

Using a longer reuse duration from a motor babbling source yield interesting result however. In Figure 6.6, diversity reuse is able to yield a significantly better performance during the first 500 steps of the exploration than random reuse.

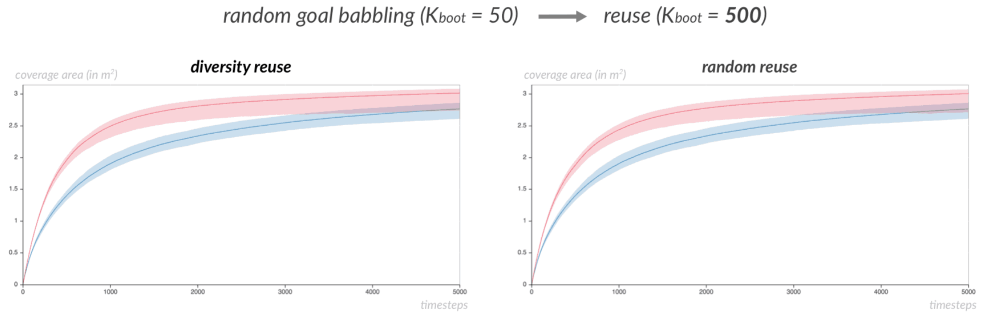


Figure 6.5: Even with an extended reuse period (500 timesteps), no performance difference is exhibited between diversity and random reuse. [source code]

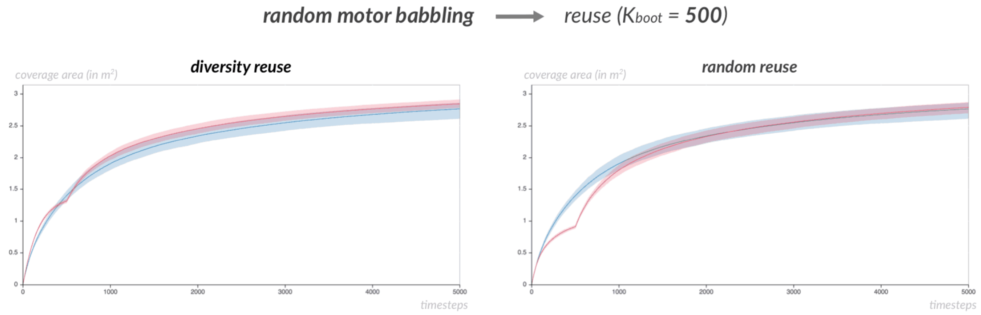


Figure 6.6: Diversity reuse almost matches the performance of the goal babbling strategy without reuse (and 50 timesteps of motor babbling), during the first 500 timesteps. The early performance of motor babbling is much worse, even if the performance of both at t = 5000 is no different from the one of Figure 6.4. [source code]

Opportunistic Diversity Exploitation

A particularly disadvantageous setting for random reuse is when few good observations are mixed with amongst large number of mediocre observations. In that case, diversity will explicitly select the few good observations, while random reuse will miss them, and overwhelmingly select mediocre observations. To that end, we consider a goal babbling source task where motor babbling lasts for 4500 timesteps. Only during the last 500 timesteps does goal babbling is run. Figure 6.7 illustrates the three sources tasks we have considered so far (the performance is shown on the same-length task, as is appropriate).

In that setting, diversity manages to provide a small improvement over random reuse (Figure 6.8).

The real difference happens when the reuse period is extended to 500 timesteps (Figure 6.9). Diversity reuse is able to fully exploit the diversity provided by the source task, while random reuse performs poorly—it will select 50 observations produced during the goal babbling phase of the first tasks out of 500, on average—–, only able to exploit the source task after the 500th timestep, when goal babbling happens. The

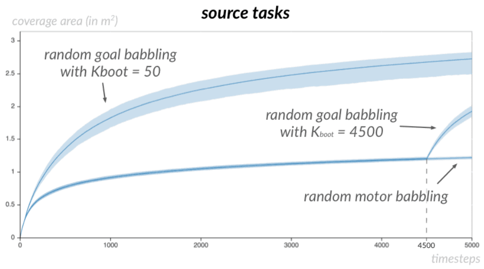


Figure 6.7: Even only over the last 500 timesteps, goal babbling provides a potential exploitable improvement in exploration. [source code]

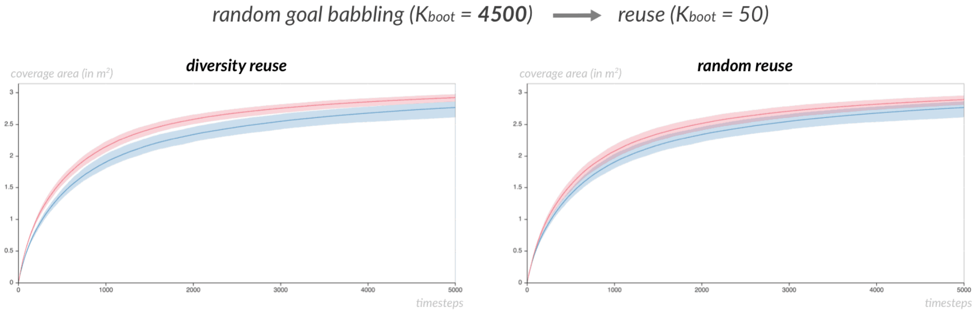


Figure 6.8: Diversity reuse exhibit better performance over a highly biased source task, but not significantly. [source code]

final performance between the two reuse strategy is not different, however.

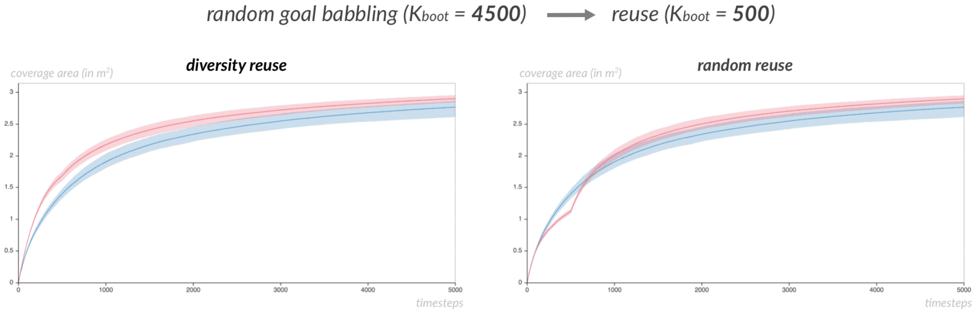


Figure 6.9: Diversity is able to fully exploit the heterogeneously distributed diversity in the observations of the source task, while random reuse performs poorly during all the reuse period. [source code]

Discussion

In these two-dimensional arm experiments, diversity reuse is consistently similar or better than random reuse, but does not exhibit significant quantitative improvement in performance over the long run. Qualitatively however, diversity reuse, at the difference of random reuse, is shown to be able to exploit a distribution of observations randomly distributed over the motor space. Additionally, it brings robustness and significant quantitative improvement during the early phase of the exploration in specific scenarios.

In an ecological context, the early period of learning is the most relevant. Animals, humans, robots do not live at the asymptote. Because time is finite while learnable skills are virtually unlimited, incentives are for engagement with a learning task to be limited. Furthermore, as devoting significant resources to a single task brings diminishing returns (Lopes and Oudeyer 2012), this behaviour can only be justified if the task is of some special importance. In a context where the agent is exploring the environment for new interactions, no interaction is a priori more special than any other. Settling to learn the first new environmental interaction found, that might be less useful and more difficult to learn than others nearby is not a good strategy. Sampling different interactions—for instance estimating their learnability (Baranes and Oudeyer 2013)—, before committing to learning any of them therefore presents a fitness advantage.

Moreover, the early phase of learning is important because learning can be interrupted at any moment. By a predator, by a more pressing physiological need, by a peer, by any sort of environmental perturbation.

Finally, an agent with good early learning performance will be able to amass more knowledge and skills by engaging over short periods of time with a diversity of learning tasks than another agent with bad early learning performance, even if both reach the same medium-term, long-term or asymptotic performance.

Therefore, in robotics, a good early learning performance is better than a good asymptotic performance. A good early learning performance is better than optimality.

Thus, diversity reuse represents a significantly better algorithm for the transfer of exploration in a robotic context where many tasks are available. It increases significantly the knowledge and skills obtained while exploring the environment.

It also favours the interactions that are naturally compatible with the already acquired competences of the agent. As any transfer learning algorithm, it has the tendency to incentivize the agent to learn progressively more complex tasks, the choice of which are dependent on the agent current competence.

Indeed, an agent whose learning is guided by progress in competence for instance (Baranes and Oudeyer 2013) will be motivated to engage with tasks that offer maximum learning progress, i.e. easy tasks. While learning progress motivates the agent to stop learning a task as soon as the (diminishing) returns decrease significantly, it can only direct the agent towards more complex tasks once all the simple tasks are learned. Which means that the agent will increase the complexity of its behaviour only in situations where there is a finite, reasonable amount of easy learning tasks. If the agent uses transfer learning, tasks that were hard at the start of learning can progressively become easier—and therefore more desirable for learning—if they are dependent on simpler, learned tasks. Transfer learning encourages the (self-)scaffolding of increasingly more complex behaviour. We'll see how reuse can foster scaffolded learning in chapter 5.

Another way to look at the previous argument is to consider that an additional *drive towards complexity* or a *boredom drive of simplicity* could be added to the agent to motivate it not to only learn simple tasks. But, alternatively, transfer learning, by modifying the patterns of learning of the agent, can produce similar behaviour. Which is better then? Modify the learning capabilities or modify the motivational drives? This issue warrants further research.

Let's remark here that while we advocate reuse for a task-rich environment, the framework proposed only deals with one source task. In a situation where the agent has multiple acquired tasks available, choosing from which task to reuse motor commands is not trivial.

The most simple way to deal with this issue is to rely on a similarity measure between the available source tasks and the target task. Provided with a slightly novel object, a child will have the tendency to reuse motor commands that provided interesting observations on similar objects. Put differently, the slightly novel objects evoke affordances (Gibson 1977) in the child that were learned by interacting with other objects in the past. These affordances will naturally bias the interactions the child chooses to engage the object with.

In this thesis, we consider the scenario of a totally novel object, that does not bear any visible similarity with past experience. This might be because the object is truly new, or because the agent is unable to accurately recognize the object as similar to other objects he already engaged with. The latter scenario is reasonable in the current context of robotic technology[4].

To choose from which tasks to select motor commands without a similarity measure, a possibility is to estimate similarity empirically from the interaction data with the object. Similar tasks, as we have already stressed, will have a tendency to generate similar level of diversity for the same motor commands. Therefore, a natural way is to proceed in a similar manner as chapter 4. The exploration strategy being selected are the reuse strategies from each source tasks: source tasks are preferentially selected by the diversity they produce during reuse. That is, the creation of diversity acts as an indicator of the compatibility between tasks for transfer. Note that if the sources tasks have already been generated by reusing from each other, they share a number of motor commands, that has to be taken into account when sampling and estimating the diversity contribution of each source task.

We do not test such an algorithm in this thesis, but it is an interesting future venue

[4] With human, such a scenario could be artificially created by asking subjects to recognize the function of novel objects in the dark. Is the haptic exploration random? Rational? In between (Cook et al. 2011)? Yet, before going further with these considerations, they need to be articulated with existing research (in particular on blind individuals), which we did not yet do.

of research.

The best simulator for spacewalking is underwater – it allows full visuals and body movement in 3D. Virtual reality is good, too, and has some advantages, like full Station simulation, not just part. Like all simulators, they have parts that are wrong and misleading: an important thing to remember when preparing for reality.

Chris Hadfield

7 A Real Robot and a Virtual Ball

In this chapter, we describe the second experimental setup used to conduct experiments with the reuse method. We constructed a hardware platform equipped with an articulated arm interacting with a simulated object. We show that reuse is effective when learning to interact with objects with significantly different response behaviour.

7.1 Experimental Setup

We consider a hardware and simulated experimental setup where a 6-joint robotic arm interacts with an object, a cube or a ball, and observes the displacement of the object at the end of the interaction. An interaction task is appropriate to demonstrate the strengths of the reuse methods, because many motor commands do not connect with the object, and thus generates little diversity. As shown in the previous chapter, the diversity reuse method will take advantage of diversity, regardless how little quantity there is.

An object interaction task is also interesting in a developmental context, as it is relevant in early exploration of the world.

Interacting with a real object presents many technical difficulties, and exposes the robot to damage. This was the original motivation for the hybrid approach we chose, where the (real) robot would interact with a simulated object in a physical engine. In the following, we first present the setup technically, and then discuss the choices that were made.

7.1.1 Hardware & Simulation Setup

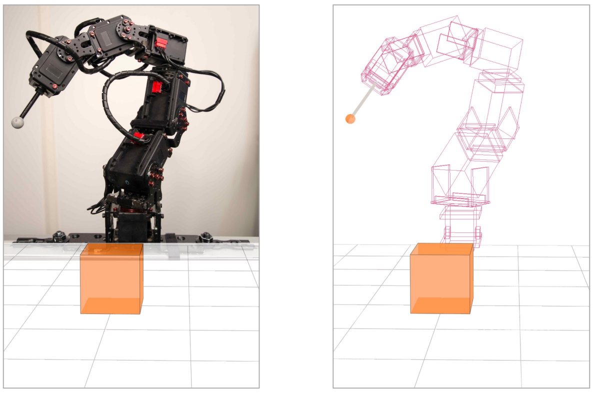


Figure 7.1: The hardware and the simulation setup share the same simulated world. However, when using the real setup, no simulation of the robot is conducted. On the real robot, a reflective marker at the tip allows its position to be tracked by cameras.

The robot is a serial chain of six servomotors. The three proximal motors are Dynamixel RX-64 and the three distal ones are RX-28. Those servomotors are capable of delivering respectively 64 and 28 kg/cm of stall torque, with an angular resolution of 0.29 degrees, measured with a mechanical potentiometer, whose precision is variable (across the angle range *and* between different motors). During the experiments, the servomotors were operated in position using the embedded PIDs, with a control loop for the position running at 100Hz.

Dynamic Movement Primitives

The movements of the robot are generated using dynamic movement primitives (DMP). DMPs are parametrized dynamical systems introduced by Ijspeert, Nakanishi and Schaal (2002). They are computed from sets of differential equations, that provides

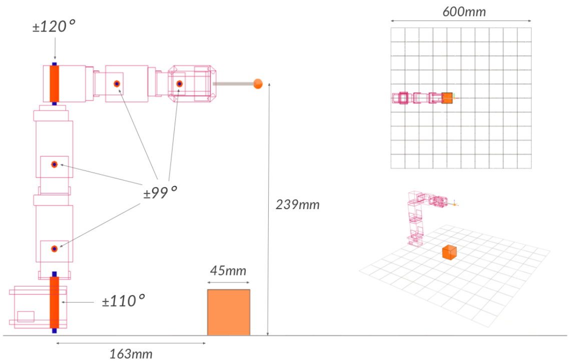


Figure 7.2: The simulated experiment approximates the real robot. Pictured here, the position zero of the robot, which corresponds to the start and target position for each movement.

guarantees of smoothness, convergence, and robustness to perturbation (Konczak 2005). We chose DMPs, and the specific parameterization we explain below, because it allowed to express many different arm trajectories with a compact description (i.e. few motor dimensions). We use the implementation of Stulp (2014), based on Ijspeert, Nakanishi, Hoffmann et al. (2013) with the sigmoid variation of Kulvicius et al. (2012).

DMPs are based on damped spring dynamics, perturbed by a forcing term (equation 7.1). They allow arbitrary smooth movements between start- and end-points. The forcing term is an arbitrary linear function, represented as linear combinations of basis functions. Here Gaussian activation functions $\Phi_i(s_t)$ are used, with centre c_i and width σ_i, weighted by w_i (equations 7.3 and 7.4). s_t is the phase of the forcing term, described by an exponential decay term (equation 7.2). Those equations do not present the more complex case we used, where the sigmoid variation is included, see (Kulvicius et al. 2012) for more details. In the following equations, τ is a temporal scaling factor[1], α and β are constants and g is the target state.

$$\tau \ddot{x}_t = \alpha(\beta(g - x_t) - \dot{x}_t) + f_t \tag{7.1}$$

$$\tau \dot{s_t} = -\alpha s_t \tag{7.2}$$

$$\Phi_i(s_t) = e^{-\frac{(s_t - c_i)^2}{2\sigma_i^2}} \tag{7.3}$$

$$f_t = \frac{\sum_{i=0}^{N} \Phi_i(s_t) w_i}{\sum_{i=0}^{N} \Phi_i(s_t)} s_t \tag{7.4}$$

In this experimental setup, the start- and end-points are set identical ($g = x_0$) and correspond to the motor being in the zero position (Figure 7.2). We use 2 basis functions per motor, with c_0 and c_1 fixed respectively at $1/3\tau$ and $2/3\tau$, with τ = 5s. σ_0 and σ_1 are shared by all motors.

For historical reasons, we don't directly use the weights for parametrizing the motor space. Instead, we use the LWLR function approximator provided with the DMP library (Stulp 2014), and define two linear functions per motor, with slope a_0, a_1 and offsets b_0, b_1 respectively. The function approximator then compute the forcing term to approximate as much as possible these functions at time c_0 and c_1. Although directly manipulating the weights would be more natural, this method provides a rich diversity of trajectories, and, because DMPs were not a focus of our work, we didn't inquire further about making the system perform better or making the representation more compact.

Each motor has independent a_0, a_1, b_0, b_1 parameters, and the motors share σ_0, σ_1, while c_0, c_1 are fixed. With 6 motors, the motion trajectory of the robot is therefore

[1] The τ of the DMPs has no relation with the τ of τ-coverage.

parametrized by a vector of dimension 26. After solving and integrating the dynamical system, we obtain each motor angular position as a function of time.

To avoid the robots removing (rather brutally) their own wires, the range of the first and fourth motor from the base are restricted to $\pm 110^\circ$ and $\pm 120^\circ$ (Figure 7.2). They both generate (potentially unrestricted) rotations around the z axis, when the rest of the robot is in zero position. All other motors were physically restricted by their horn to $\pm 99^\circ$.

The ranges of the DMP parameters are set so that 95% of trajectories of a motor would fall in between the angles the motors were able to produce (using an empirical evaluation), and clipped the rest to legal motor values.

Before executing the motion on the robot, we check for self-collisions, and collisions with the armature of the experiment. If present, the trajectory is truncated and stops just before the collision to avoid damage.

Hybrid Interactions

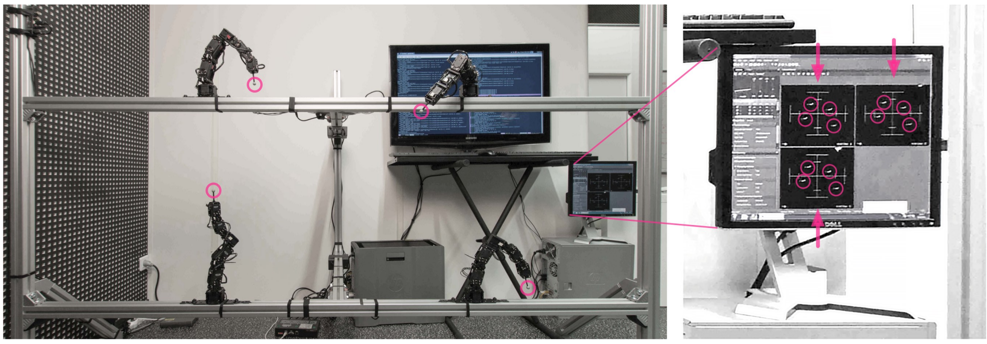


Figure 7.3: The hardware setup consists of four robots, separated so that they cannot interact with each other. The tracking system is positioned in front of the setup, and possess three cameras that capture the position of the four markers. The monitor on the right shows the detection mask of each camera. Most movements of the stems will keep the marker visible, but some will not. However, those movements will overwhelmingly be far away from the virtual objects.

The robot has a reflective marker at the tip, which allows to capture its position at 120Hz during the movement using an OptiTrack Trio camera system, that has sub-millimetre accuracy. A virtual marker then *replays* the trajectory in a simulation where a virtual object has been put. The marker is the only object from the camera that is transported to the simulation, so it is the only part of the robotic arm that can collide with the object.

Let's note that, in order to simplify the setup, the robot executes the movement, and then, after the motion is finished, the trajectory is encoded and transported into the simulation to be replayed. This absence of real-time prevents any immediate feedback to the robot during motion.

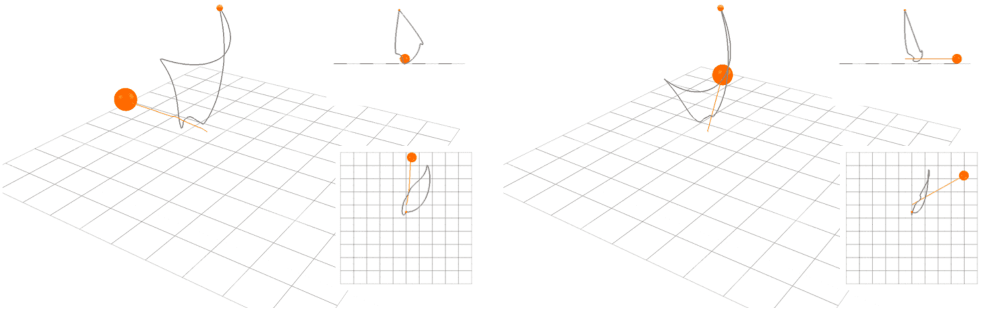


Figure 7.4: The a virtual marker replays the movement captured by the cameras of the real maker at the robot's tip, and interacts with a virtual ball. The two scenes illustrate the perturbation created by the inverse model introduced in part one. The parameters of the right motor trajectory are a random perturbation of the one of the left trajectory, with p = 0.05.

Virtual Environment

The simulated environment features an object placed in a cubic room. While the object cannot fall into the ground, the robot can pass through it, both with the real robot and the simulated robot. While constraining the robot movement to not traverse the ground is possible by truncating the movement before collision, it would remove too much density and diversity of useful movement in the space of parameters.

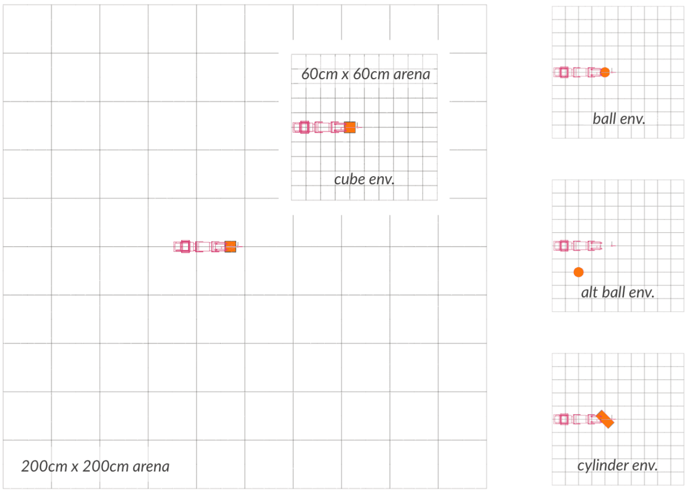


Figure 7.5: The size of the cubic room does not modify the relative position of the robot and the object.

We consider two sizes for the cubic room: 600mm width and 2000mm width. The larger room approximates a unbounded environment, while the interaction between the object and the walls are frequent in the smaller one.

Three different objects are used—one at a time: a ball and a cube of diameter and width 45mm respectively, and a cylinder with diameter 40mm and length 80mm. For the ball, two different positions are considered, as depicted figure Figure 7.5.

The simulation is conducted using the robot simulator V-REP (Virtual Robot Experiment Platform), with the Open Dynamic Engine (ODE) as a physic engine backend. At the end of the simulation, the trajectory of the object is processed by sensory primitives that compute the sensory feedback.

Sensory Primitive

We consider a simple sensory primitive that returns the displacement of the object projected on the ground at the end of the simulation. The displacement is returned as a vector of length 3: the displacement in x, in y, and a discrete dimension of saliency, which has value 0 if no collision happened, and 1 otherwise.

The saliency dimension helps separate observations that create collisions from one that do not. Admittedly, this is not crucial for the perturbation-based inverse model, but when using the LWLR model (Appendix C), this makes learning more robust.

7.1.2 Behaviour of the Setup

Before investigating the behaviour of an agent exploring the environment, we study the general behaviour of the environment itself.

We claimed that the DMPs parameterization creates appropriate movement diversity for the robot. This is illustrated Figure 7.7.

Instances of movement reuse are shown Figure 7.8. Reused movements do not generate necessarily similar effect on the objects, and, in instances, it has a significant impact on the robot's motion. Moreover, not all movements that interact with an object interact with other objects, even when they are the same size and placed at the same location.

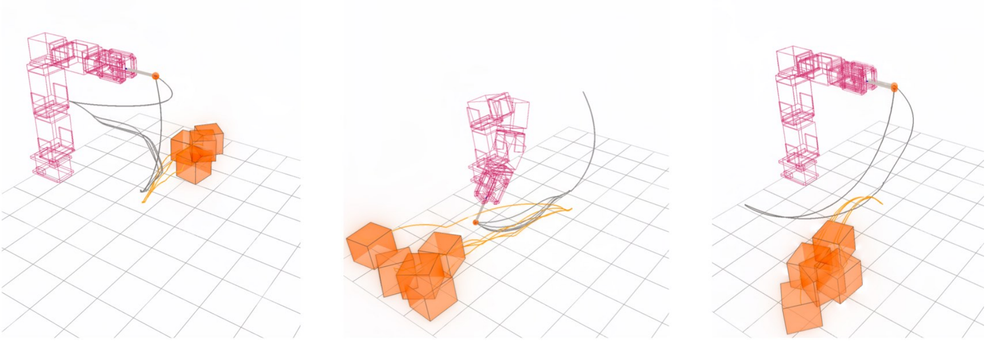


Figure 7.6: The physic engine non-deterministic characteristics generate a lot of variations. In each of these images, multiple executions of the same motor trajectories are overlaid.

Finally, the simulation is not deterministic. Repeated executions of the same movement can generate significantly different effects, as shown Figure 7.6. This is not due to synchronization variability. The motor trajectories are generated to match the simulator step, and the same motor target are fed to the simulation at the same timestep every time. The simulation is also reset to a precisely identical initial situation each time. The source of variability is due to the random seed of ODE not being reset between interactions[2]. As ODE uses the current state of the random generator to decide the order with which to resolve the constraints at each steps, small variations are introduced that are amplified by the chaotic nature of the interaction with the objects. The other physical backend available in V-REP, Bullet, did not have this characteristic and generated consistently identical effects. Because the behaviour of ODE is (slightly) more realistic, we decided to use it rather than Bullet.

We ran experiments on the ball task (because the cube occupies more volume, the ball gives a lower estimate of the collision probability) to decide which number of motor babbling timesteps to use during the experiments. The subsequent goal babbling exploration uses the inverse model introduced in section 3.1.1, with d set to 0.05. The τ of the coverage measure is 22.5mm, the radius of the ball.

The results, Figure 7.9, show a significant diminution of variance until 200 steps of motor babbling. Through independent tests on large quantities (10000+) of motor babbling movements, we estimate the probability to touch the object during a movement at 3.05% for the cube, 1,96% for the ball, and 0.70% for the ball at the alternative location. To ensure that every motor babbling phase had at least one collision, we set the bootstrapping phase to 200 steps. In the results, we will use a source and a reference task using 200 steps. But using a more aggressive reference task with only 50 or 100 steps does improve early learning performance, and this should be taken into account when interpreting the results.

Discussion

Practical Aspects of Random Babbling

A issue not addressed thus far is how practical motor or goal babbling is. Motor babbling, with its blind creativity, can easily damage the robot or endanger users in social experiments.

Avoiding damage in on-board learning robotics is a challenging issue (Levi et al. 2010). Avoiding damage means most of the time placing constraints on the robot actions, such as truncating the trajectories that led up to collisions as we did. These constraints, however, may adversely affect performance, and restrict access to good solutions. A tradeoff must be made between avoiding the robot destructing itself or the environment, and achieving the best possible performance. Wahby et al. (2015)

[2] This is an implementation detail of V-REP, and there was no way to change it the version we used (3.1.2).

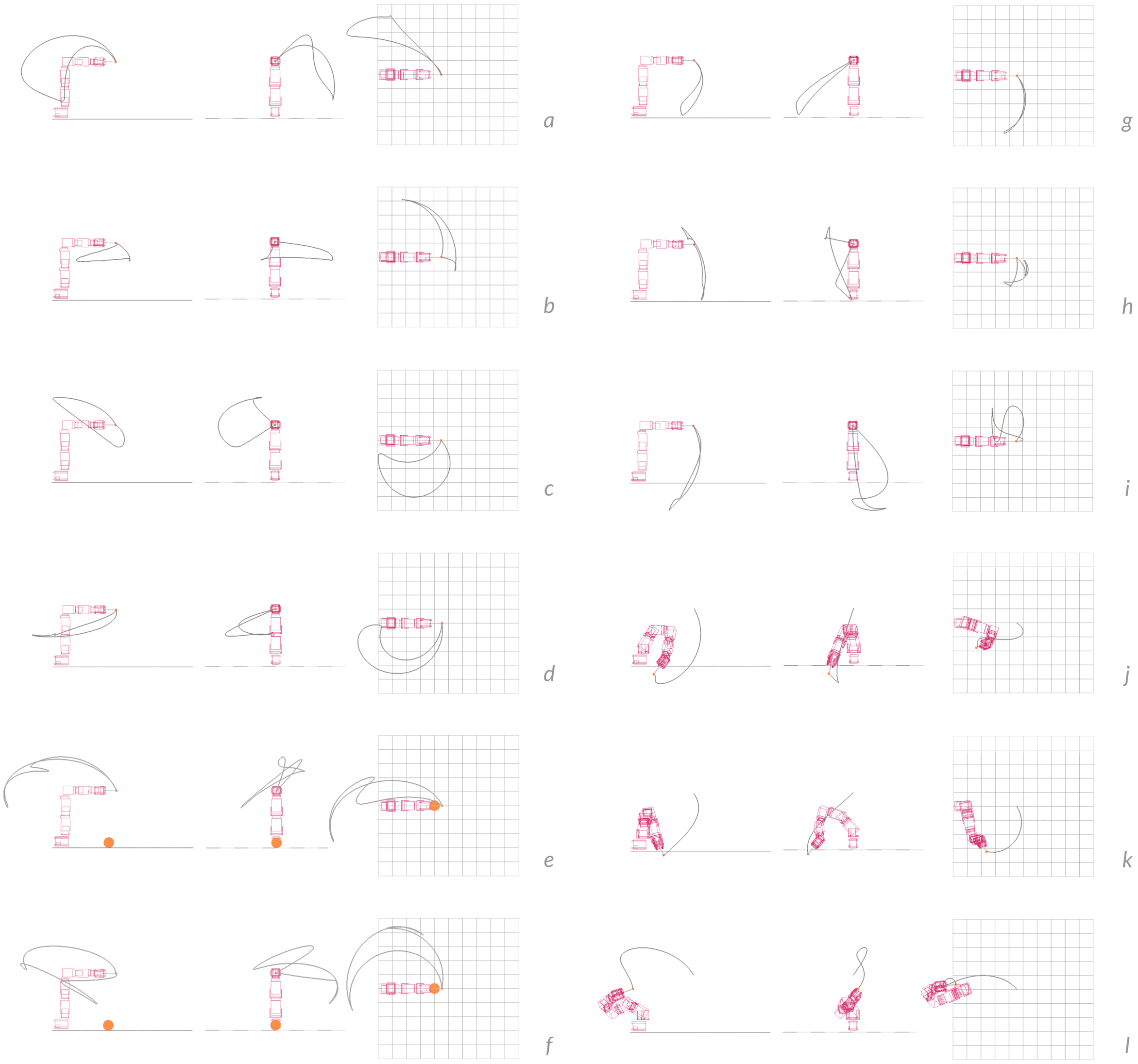


Figure 7.7: The DMP parameterization creates many different trajectories for the tip of the arm. The trajectories of the first column are all different, but they are far from the ground, and will not result in any interaction with the objects. Trajectories [g-i] on the other hand, do approach or even traverse the ground. Let's note the trajectory g demonstrates that the robot can interact with an object even if it is not just below the robot. Finally, trajectories j and k shows how the movement of the robot is stopped before a collision can happen with the environment. And in trajectory l, the system stops a self-collision from happening. While not represented here, the aluminium beam that compose the frame of the hardware setup (see Figure 7.3) can easily damage the robot if bumped into violently. The most practically problematic aspect of it is that the reflective surface of the marker is vulnerable to abrasion. If physical collisions are not prevented, the cameras quickly become unable to track the marker.

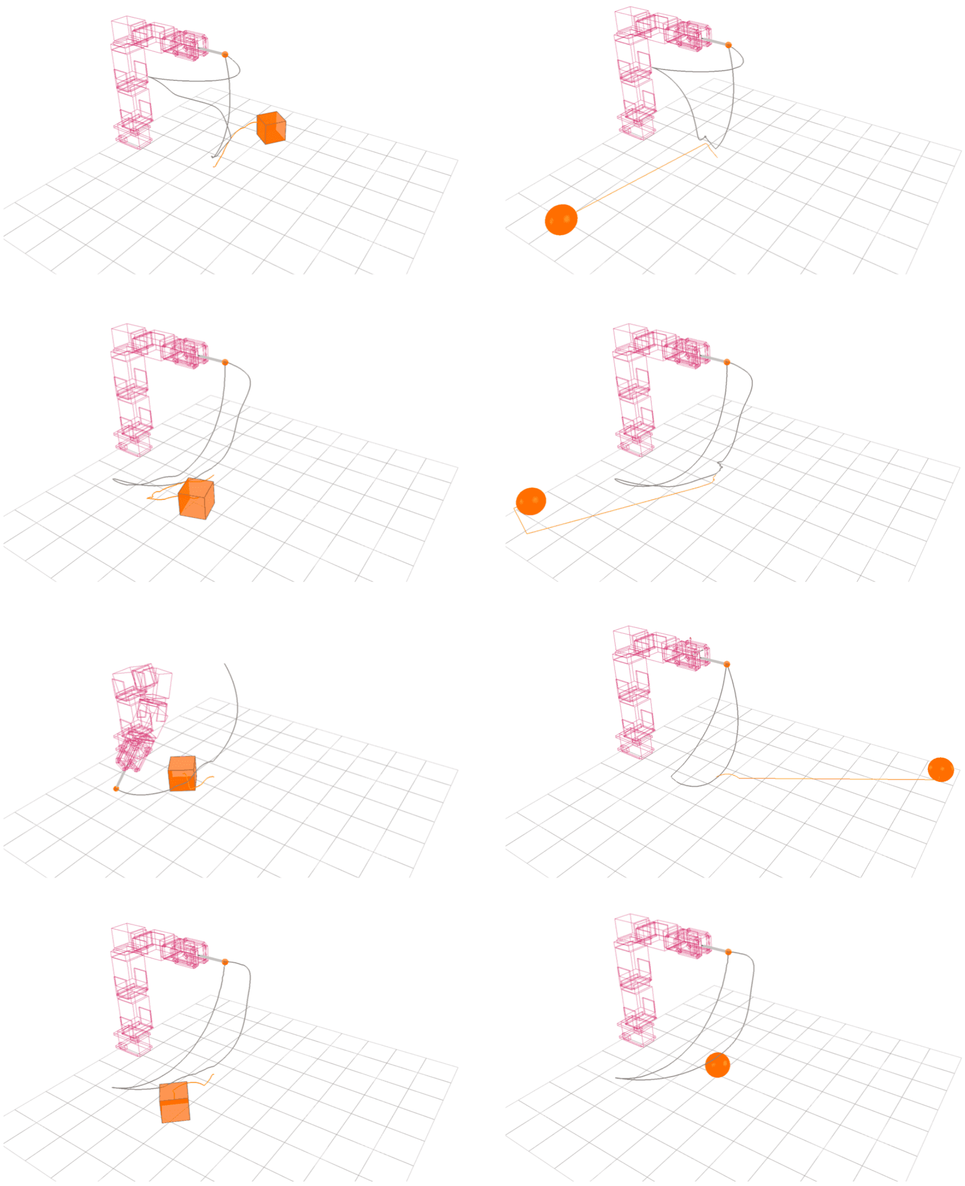


Figure 7.8: Reusing motor commands on different object does not produce similar effects. In these simulated examples, taken from motor commands reused from a cube environment into the ball environment, the two objects respond very differently to the same commands. The first example shows that the displacement can be diametrically opposed. In the second example, the arm interacts two times with the cube (explaining the U-turn in the object trajectory), but does not with the ball, that escapes the reach of the arm quickly after the interaction. The third example shows a motor command that has the arm pushing on the cube from above, creating high reciprocal forces between the two objects. When the tension is liberated the arm overshoots its trajectory, and goes for the beam (invisible here, see Figure 7.3), and therefore the movement get stopped by the anti-collision system. In the case of the ball, none of this happens, as the ball quickly resolves the impulsion from the arm. Finally, the fourth example shows that not all movements that interact with the cube interact with the ball despite their identical size, due to the cube larger space occupancy. In all these examples, the interaction with the object has significant impact on the arm's motion.

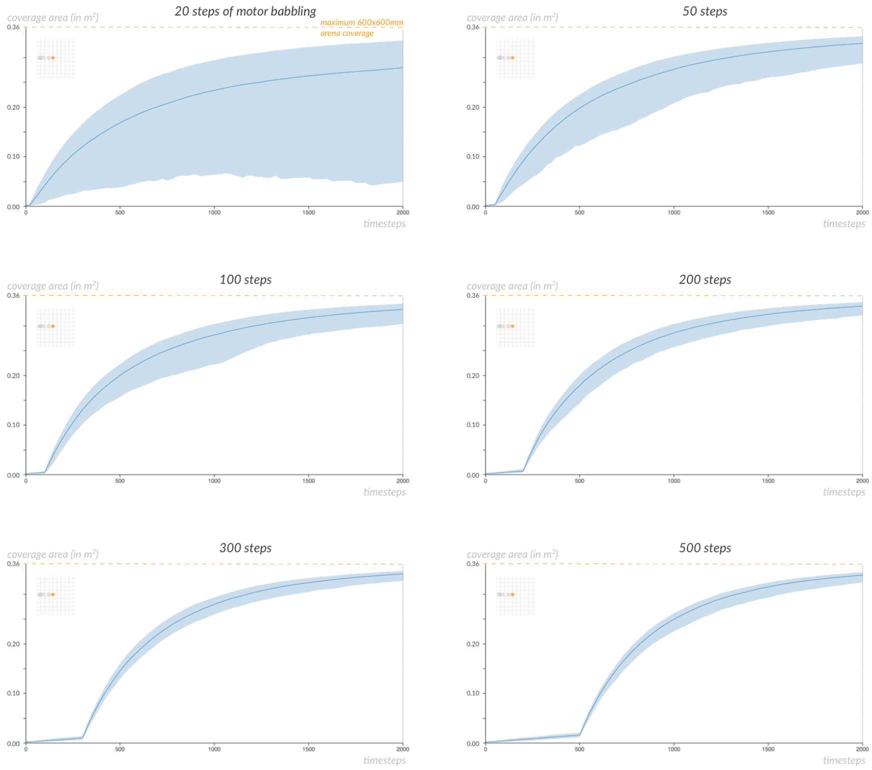


Figure 7.9: The interaction task forces a long bootstrapping phase. Failing that, the overall performance of the exploration varies significantly. Repeated 100 times. [source code]

studied this issue in the context of embodied evolutionary robotics, adding penalties to the fitness function in case of violent interaction with the environment, and effectively measure a trade-off between performance and protection.

Many works have also studied adapting behaviour after damage, in particular in the context of the robustness of gaits when the robot loses or damages one of its limbs (Mahdavi et al. 2003; Bongard and Lipson 2005; Doncieux and Mouret 2010; Cully et al. 2014).

Avoiding damage during random motor babbling on an arbitrary robot platform is difficult. A better approach is to design robots so that they can babble safely.

Of course, the most straightforward solution is to make robot less fragile. While it seems evident that robot deployed in the depths of the ocean must be hardened against a multitude of conditions, because any problem requires to abort the activity and pull the robot to the surface, exploring robots in social environments should be considered as inaccessible to the engineer for repairs as the bottom of the sea. But structural robustness is not enough.

Biology is a great source of inspiration here. As discussed in section 2.3, human fetus start to babble in utero, where the amniotic fluid dampens the motion of the limbs, and the uterine walls provide an elastic source of collision. At birth, newborn are subject to gravity without buoyancy, and their movements more limited than before. Which is just as well, because the muscle have to handle the full inertia of the movements they create: the child movements are reduced at the precise time they become more dangerous. The increase in muscle strength leads to the gradual development of the mobility of the infants, thereby mitigating any risks linked to wandering too far. In animals, some species' have their pups born blind, which also reduces the risk-taking and wandering activity early after birth. The structure of the bone is also conductive of babbling: three-year-old bones absorb three times as much impact energy as ninety-year-old bones (Currey 1979). The small size of children decreases the consequences of a fall.

Moreover, the environment of infants is not arbitrary. Children are kept in safe environments, whose constraints are progressively relaxed as autonomy develops. Danger-prone or injury-prone activities are first experienced safety nets and safety gear: training wheels on bikes, arm buoys for when swimming, and adult supervision. Safety gear in children (and adults) increases risk-taking behaviour (Morrongiello et al. 2007). In other words, children engage in a less restricted repertoire of activities when they feel protected.

A robot that explores the environment is inherently exposed to a non-zero amount of risk. Therefore, at the inverse of industrial car assemblers put in cages, *exploring robots must be designed for risk taking*. Like infants, they must undergo comprehensive developmental constraints, coupled with environments that match their ability for control. They should be initially in padded environments where risks of injury is minimal, their movement range and torque reigned in. And their body must be compliant, and their behaviour reactive to potential damage: when falling, they should react to minimize the fall impact, not try to regain a balance they unequivocally lost Ruiz-del-Solar et al. (2009). Their body must be initially small enough, light enough or compliant enough to withstand fall (Lapeyre, Ly et al. 2011). All those characteristics also reduce the danger to interact—or just to be standing near—the robot (Lapeyre, Rouanet et al. 2013).

Not any robot can babble safely. But developmental robots should be designed so that they can.

The robot arm considered in the experiment displays few qualities conductive of safe motor or goal babbling. To deal with this, we opted for an augmented-reality approach to interaction.

An Hybrid, Augmented-Reality Approach

We chose to use a real robot and a simulated environment for several reasons. Placing an object back into the reach of the robot a few ten of thousands of times after an interaction requires some form of mechanism, or a bigger robot, which makes the experimental setup more complicated. Replacing the object takes time, and slows down the rate of interactions.

Additionally, the robot never experiences physical collisions, which reduces the risk of damage when babbling, given the type of robot we had. Measuring the motion of the object along any conceivable dimension or the force and direction of the collisions does not require equipment and is computationally free.

A virtual environment approach also affords unlimited flexibility in the creating of several different learning tasks, even ones that would be physically possible. And this provides a perfect, transferable and reproducible description of the environment of the robot.

At the same time, using a virtual environment for an interaction task seems to remove the main source of interest of the setup: a realistic, difficult to simulate, interaction with a real object.

These types of contacts are difficult to simulate (Ijspeert 2008)[647]; current physic engines make fundamental simplifying assumptions. They use impulse-based velocity stepping methods for contact dynamics (Mirtich et al. 1995; Stewart et al. 1996; Anitescu and Potra 1997) but solving the methods exactly is NP-hard. Approximations of the simplifications must be made (Anitescu 2005; Kaufman et al. 2008; Drumwright et al. 2010; Todorov 2014), which, as Erez et al. (n.d.) points out, does not make the question of the physical accuracy any simpler. The simulator we used, the Open Dynamic Engine (ODE) is notorious for its weaknesses at simulating interactions.

Our approach, beyond simulation realism, presents another problem: the robot does not receive any kinesthetic feedback, which, as we have seen Figure 7.8, has an important impact on the interaction. While this kinesthetic interaction was simulated between the marker and the object, the marker was following the trajectory of the tip of the robot with a spring constraint, and did not reproduce the precise force generated by the sum of the torque of the motors in a specific posture.

Moreover, while providing great flexibility, simulations always present a danger: as Jakobi et al. (1995) puts it: 'they can lead to both the study of problems that do not exist in the real world, and the ignoring of problems that do.'.

This is facilitated by the shortcuts the simulations afford when designing an experiment: objects tracked to meaningless precision, and can be created or destroyed dynamically, scenes can be perfectly reset to initial conditions. While seemingly innocuous, they actually can hide important issues in a real environment cannot avoid. If the experimenter is oblivious to them, they may affect the whole basis of the experiment or the applicability of the method it develops to real robots.

Specifically for our setup, one of the most problematic behaviour was movements that push the object towards the ground, resulting in the cube sometimes projected with great velocity in a chaotic direction. In the context of our diversity-driven approach, these movements are seen a valuable: they create effects that are often selected as nearest neighbour during goal babbling, and reused during transfer. In reality, those movements might not even be attempted because of the motor damage they represent. If these interactions were absent, the results of our experiments would be much different, creating sharper distinctions. The next step of our research is to reuse motor commands from simulated objects to real objects, and this will provide critical feedback on the validity of our methods, as well as force us to remove those dangerous interactions from the exploration one way or another.

Overall, the hardware platform is somewhat disappointing so far. In our experience, there is no real difference in algorithmic performance between the hardware/simulation hybrid and the fully simulated platform. For these reason, and because much of the experimental data with real robots had to be discarded because of unchecked assumptions, few results on the real setup are presented. At any rate, our setup is unconvincing about the validity of simulated environments.

However, simulated environments—and augmented-reality environments—might turn out to be a useful tool for robotic research. They are a middle ground between a simulated robot and a real environment: they allow robots whose morphology precludes a useful simulation of the robot to be easily subjected to a variety of situations without costs or physical damage. Because they provide full knowledge of the environment, they are a clear experimental asset. Because they provide full control of the environment, they allow to disentangle the reasons for a specific behaviour by systematically controlling different variables.

Of course, they are severely limited, although not impossible, when interaction is required. But for a robot learning to avoid obstacles for instance, this is not an issue.

In a developmental perspective, they allow the environment to be reactive to the development of the robot. First, it allows the environment to actively create specific situations aimed at estimating the degree of development of the robot. Like Bongard and Lipson (2005) co-evolving a behaviours and series of informative tests, the environment can adapt to efficiently estimate the competence of the robot regardless of which development path it chooses. Second, the environment can provide a progressive increase in complexity to scaffold the behaviour of the robot. In Chapter 8, we provide examples of how this can be done.

Simulated environment are not an objective, and they represent the same danger as simulations. Yet, they may be useful during the research process as they represent one more tool to study complex issues.

7.2 Experiments

7.2.1 The Small Arena

We conducted experiments using the reuse exploration for half of the bootstrapping phase ($p_{\text{reuse}} = 0.5$), set at 200 timesteps ($K_{\text{boot}} == 200$). The goal babbling exploration is unchanged, and uses the inverse model introduced in section 3.1.1, with d set to 0.05. Each exploration lasted 1000 timesteps. The τ of the coverage measure is 45 mm.

Figure 7.10 shows the qualitative effect of reuse from the cube task to the ball task. Reuse provides many examples of interactions during the first 200 steps, while motor babbling only provides ten. Still, at the end of the exploration, the reachable space is well covered in both instances.

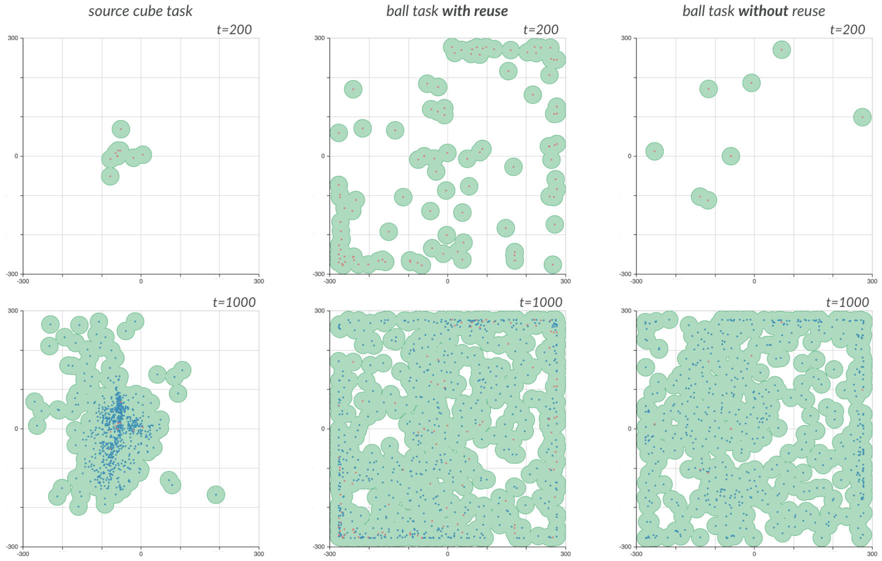


Figure 7.10: Reuse provides early diversity of effects. [source code]

In Figure 7.11, all combinations of the cube and ball task are presented. The effectiveness of reuse is sensitive to the similarity between the tasks: it is better from the same object (ball to ball, cube to cube), than from a different object. Diversity reuse also provides significant differences in early performance over random reuse, further demonstrating the usefulness of diversity as a guiding measure for transfer.

In Figure 7.12, the ball is moved from its central position to create a dissimilar task. The majority of movements that interact with the central ball will not interact with the side ball, and vice-versa. Reuse, in this situation, proves to be robust to dissimilarity, exhibiting no negative transfer.

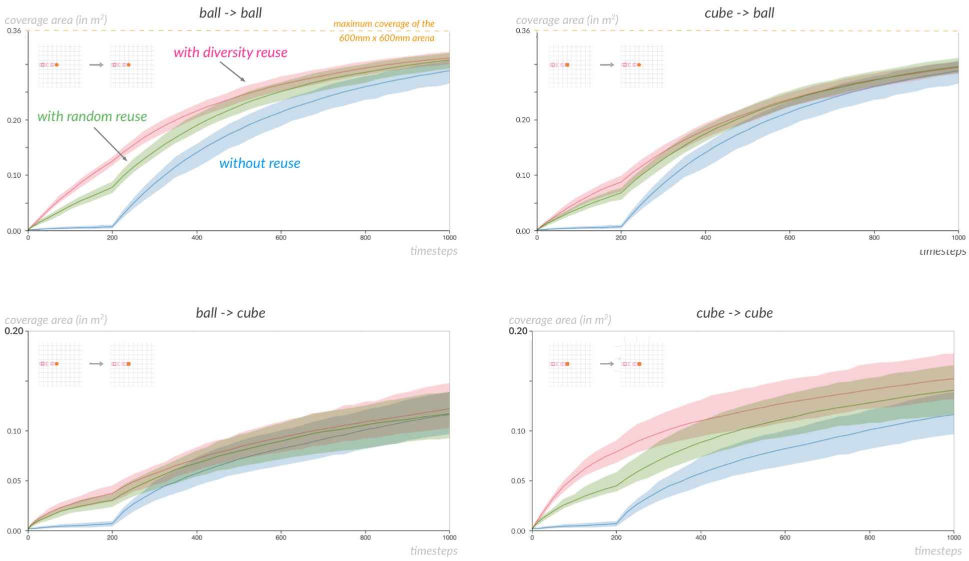


Figure 7.11: Although Reuse is sensitive to the similarity between tasks, it provides significant early exploration improvement between objects responding differently to interactions. Repeated 25 times. [source code]

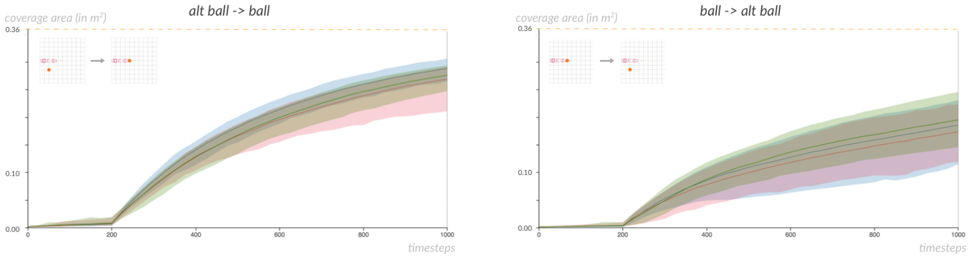


Figure 7.12: Between two dissimilar tasks, Reuse maintains performance similar to an absence of transfer. Repeated 25 times. [source code]

In Figure 7.13, the source task explored the environment using random motor babbling. Naturally, random reuse offers the same the exploration without reuse. Diversity is able to extract useful motor commands from the source task, but their quantity is limited, leading to a visible plateau during the first 200 steps.

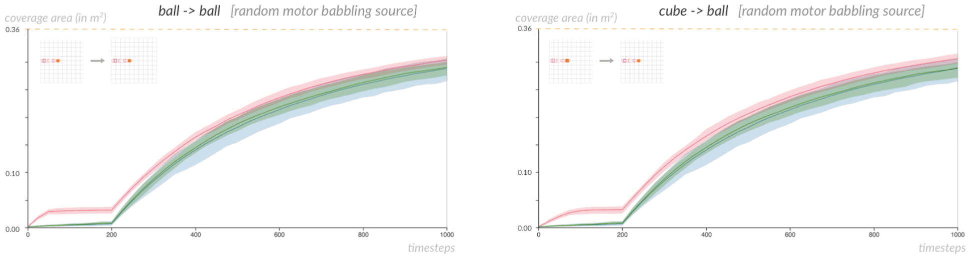


Figure 7.13: Diversity reuse only can exploit random motor babbling data. Repeated 25 times. [source code]

In Figure 7.14, the source task is the cylinder task, but it is used with a different sensory modality. The sensory primitive of the environment capture the rotation of the cylinder along its axis between each timesteps of the simulation, and sum the absolute differences between timesteps. Similarly, the who the cylinder spins is measured by measuring the rotation of the cylinder against the z-axis. The result is a 2D sensory space that expresses different aspects of the interaction that the displacement of the cylinder, in different units. Still, the reuse proves effective, and the difference with the cylinder using the displacement primitive is small. Reuse can be both sensitive and robust to different modalities.

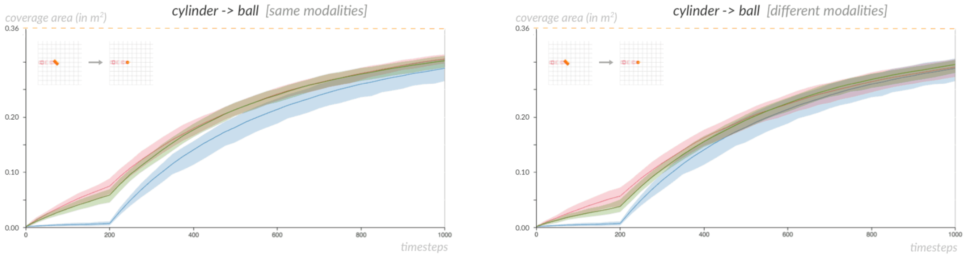


Figure 7.14: Diversity reuse only can exploit random motor babbling data. Repeated 25 times. [source code]

7.2.2 The Big Arena

In this section, we consider the 2000 mm arena, instead of the 600 mm one. With the small arena, the exploration can cover the entire reachable space, as Figure 7.10 illustrates. This is not possible with the larger arena, which is more than 10 times bigger.

Figure 7.15 present results of reuse on the hardware platform between the ball and cube task. For this set of experiments, the LWLR-L-BGFS-B inverse model described in appendix C has been used. The pure goal babbling exploration is also replaced by a mixed exploration of random motor babbling and random goal babbling. 10% of the interactions after the end of the bootstrapping phase are created using random motor babbling. The bootstrapping phase is also extended to 300 timesteps.

The use of those parameters, with a complicated inverse model that does not bring much performance gain, and a long bootstrapping phase, is only justified because they are the one that were used during the only hardware experiments that were deemed correct, and not corrupted by bugs, motor failures, or calibration issues.

Figure 7.16 reproduces the results of Figure 7.15 in simulation. We observe similar patterns as in Figure 7.11: a sensitivity to the task dissimilarity, and diversity reuse consistently providing similar or better exploration than random reuse.

The final performance pattern, however, is different. With a almost unbounded space, the probability for an object final position to be similar to the one produced by

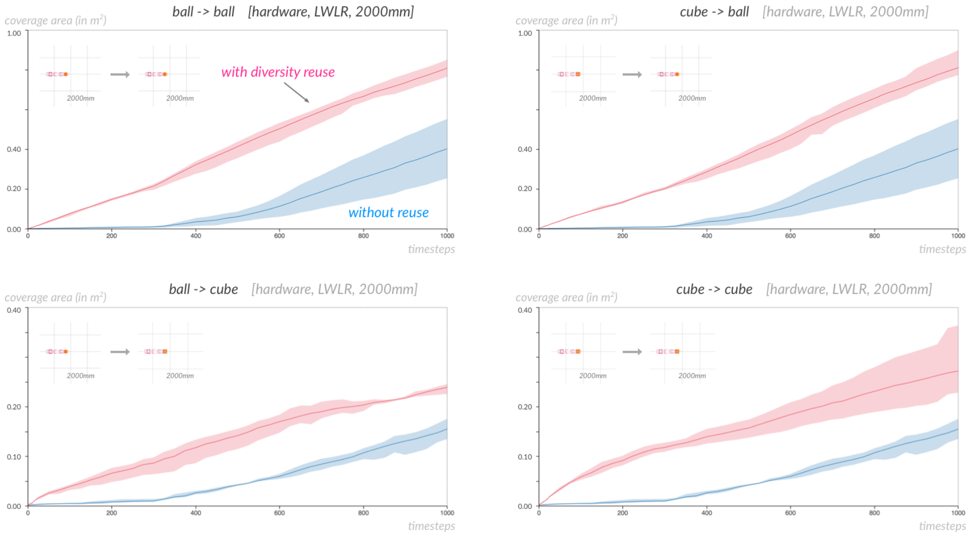


Figure 7.15: The hardware setup provides results very similar to Figure 7.17 and 7.16. However, the low number of repetitions makes those results only prospective. Repeated 4 times. [source code]

a past interaction is lower. Each colliding interaction thus brings additional coverage, as the quasi-parallel coverage curves show. The final performance is then function of the time when goal babbling became efficient. With diversity, useful interactions are provided right from the start, and the head start is kept until the end of the exploration.

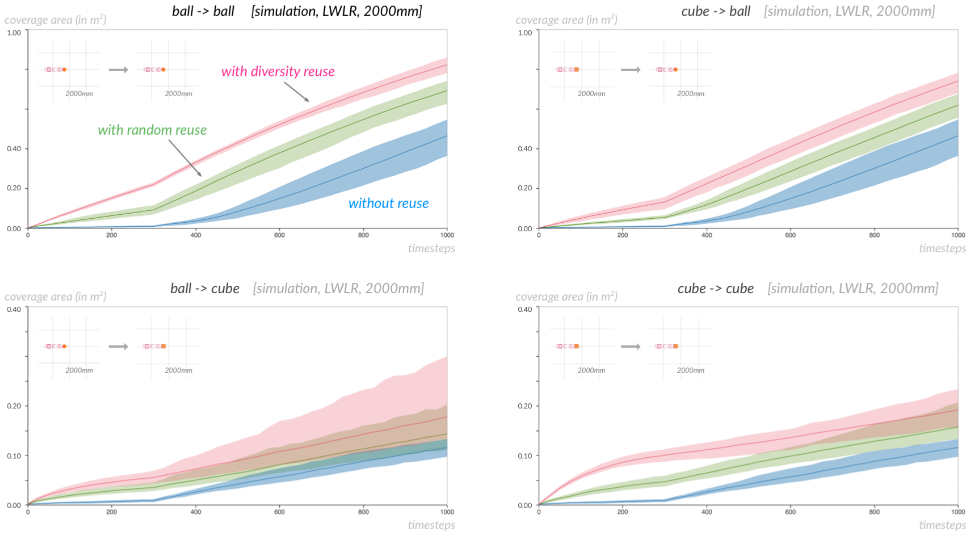


Figure 7.16: Repeated 25 times. [source code]

Figure 7.17 reproduce the result of Figure 7.16 using the exploration strategy of the previous section: a simple perturbation-based inverse model, pure goal babbling after the bootstrapping phase, which is kept to 300 to allow better comparison with LWLR-L-BFGS-B results.

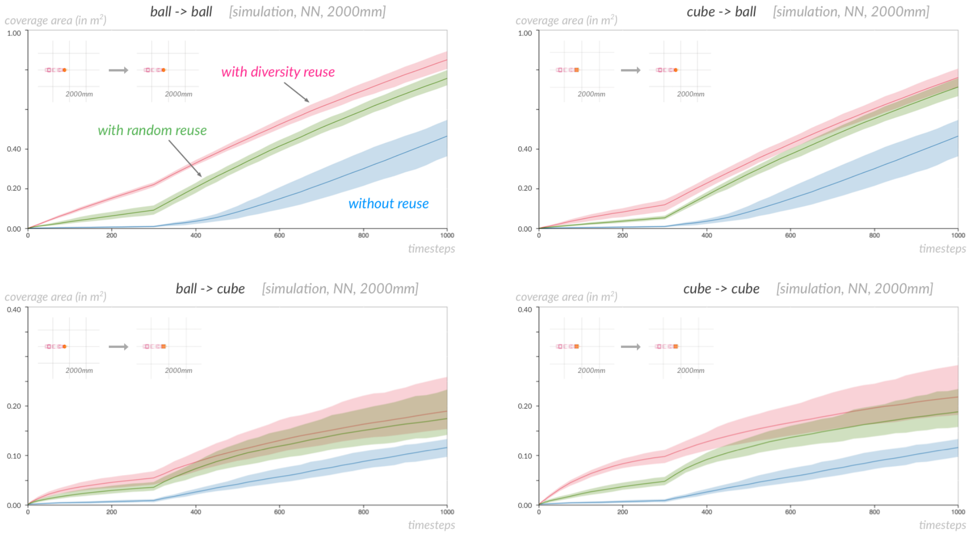


Figure 7.17: Repeated 25 times. [source code]

Discussion

These results are very preliminary. Although they consistently show how reuse can bring good early learning performance, they are fragile.

They seem to rely on the chaotic behaviour of the simulator when the robots pushes the objects towards the ground. However, even the extend of that influence is not properly analysed.

Moreover, they are established only for a simple exploration strategy. The consistency of the rapid increase in coverage after the end of the bootstrapping phase in tasks without reuse indicates that the difference could be reduced by bootstrapping more parsimoniously, as Figure 7.9 suggest.

Successful creative adults seem to combine the wide-ranging exploration and openness we see in children with the focus and discipline we see in adults.

Alison Gopnik

8

Shaping Diversity: Learning Pool

In this brief chapter, we show how reuse can be used to direct exploration by manipulating the environment.

8.1 The Pool Experiment

So far, the impact of reuse has been to improve the performance of early exploration. But after enough time, the exploration without reuse will catch up, and no significant difference will be observed between an exploration that exploited reuse and one that did not.

The experiment in this chapter aims at demonstrating that the reuse method can also make explorable an environment that is not otherwise. To do this, we consider a pool situation, where the robot can interact with a ball, but receive sensory feedback from another ball, out of reach. The only solution for the robot, in order to generate a diversity of effects on the out-of-reach ball is to strike it with the ball it can interact with.

From scratch, it is very difficult to create diversity, as only a very small area of the motor space will, and no guiding signal is provided by the environment. We consider a reuse scenario where the robot first explore how to interact with the ball it can reach, without the other ball present. Then, the out-of-reach ball is introduced, and the sensory feedback of the first ball is removed from the perception of the robot.

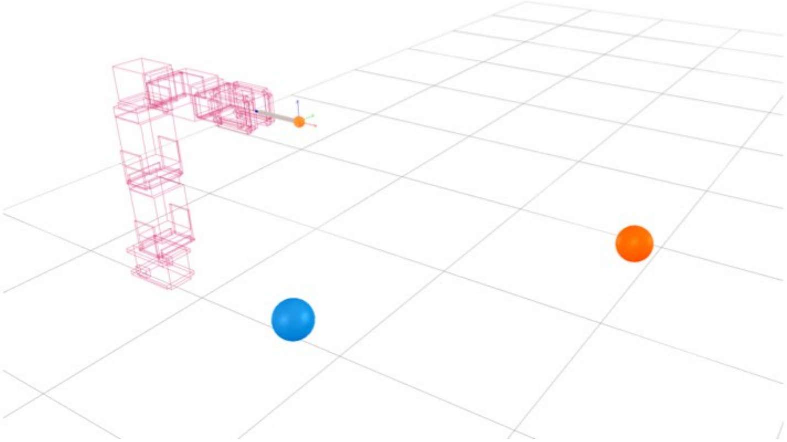


Figure 8.1: The Pool environment. The blue ball is placed at the same location of the ball of the source task, but it is not tracked by the robot. Only the orange ball is, but it is out of reach. The only possibility of interaction is to launch the blue ball at the orange ball.

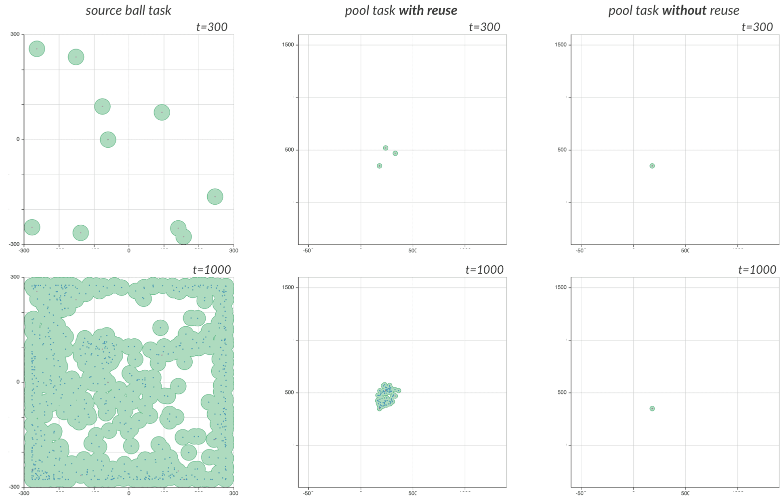


Figure 8.2: From a classic ball task is a 600mm arena, the reuse strategy successfully bootstraps the exploration of the pool environment, which is then able to produce diverse effects on the orange ball during goal babbling. Here the physical properties of the ball and the forces developed by the robot limit the distance the second ball can go. [source code]

Discussion

Staging the exploration this way is similar to reward shaping in reinforcement learning (Dorigo et al. 1994; Mataric 1994) and staging the fitness function in evolutionary robotics (Gomez and Miikkulainen 1997; Urzelai et al. 1998; Kodjabachian et al. 1998).

But there is one important difference here: there is neither a reward nor a fitness

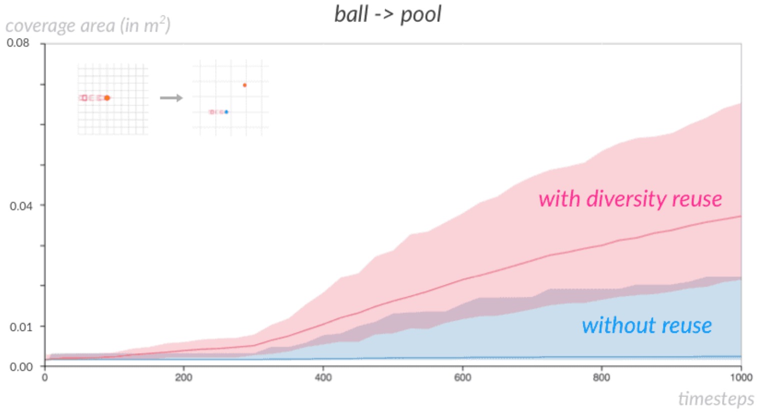


Figure 8.3: Most of the exploration that do not benefit from reuse never makes the ball move in the pool environment. Few do by chance, as the deviation shows. Using reuse however, the exploration of the second ball is consistently done. Naturally, with a task that requires lots of precision, and the stochastic behaviour of the interaction with the object Figure 7.6, lots of variation is observed. Repeated 25 times. [source code]

function. The staging is done through the *sequence of environments*, and object saliency. The diversity fostering exploration strategies (implicit diversity motivation) and the diversity-driven reuse method (explicit diversity motivation) ensure that the robot takes advantage of the relation between the environment.

This opens the door to environment-driven development in robotics. Objectives are abandoned, as proposed Lehman and Stanley (2011a), the growth and behaviour of the robots are dependent on the environments where they are put, and how those environments evolve as they competences increase.

With this experiment, we highlight that we can drive the robot towards the acquisition of complex skills in a closed-skull manner, by only manipulating the environment the robot is exposed to. This is similar to a caregiver manipulating the composition, disposition and saliency of objects a child is playing with.

No plan survives first contact with implementation.

9
Reuse and the Reality Gap

So far we have shown that reuse is effective in situations that involve switching the object (ball/cube experiment in chapter 7), changes in the morphology of the robot (different segment lengths in chapter 5), or increased complexity (scaffolding experiments in chapter 8). The purpose of using reuse in these situations is to leverage past experiences to provide the locations possible good mappings in the sensorimotor space.

In this chapter we apply reuse algorithms to a surrogate context: a simple, computationally efficient simulation is used as source task for a more expensive and more realistic simulation, or for a real robot.

9.1 The Reality Gap

Abstract · Transferring behaviour learned in simulations to real robots is difficult: this is the reality gap. We review the problem and some of its solutions.

We already discussed some of the pitfalls of a full-representation approach to behaviour. Obtaining the representation is difficult or impossible. It needs to simulate the morphology, and hence bears the costs of simulating morphological computational processes that the agent does not otherwise need to have explicit knowledge of to elicit successful behaviour. The simulation itself will be inaccurate, however pain is taken to create it[1]. Moreover, the simulation can be computationally expensive, taking sometimes more time that the execution on the real robot.

Many experiments learning controllers for legged robots have reported remarkable performances for simulated robots. But far fewer have been able to transfer those controllers learned in simulation onto real robots and observe the similar performances (Lipson, Bongard et al. 2006; Palmer et al. 2009). In other words, the transfer from simulation to reality is not efficient: this is the *reality gap* problem (Jakobi et al. 1995; Jakobi 1998).

The *Why* of the Reality Gap

Avoiding the reality gap problem by only learning and exploring on real robots raises practical issues: it is time-expensive, cannot be parallelized (unless many identical-enough robots are available, which leads to other issues and high costs), and can lead to damage or danger—especially if babbling randomly (Wahby et al. 2015). It severely limits the amount of learning the robots can receive, thus undermining their performance—such a problem is particularly acute in evolutionary robotics, where populations of candidates have to be tested over several generations (Floreano and Mondada 1994; Zykov et al. 2004; Regan et al. 2006; Gongora et al. 2009). And when exploring morphology changes, it is usually impractical to work with real robots.

In a scientific context, it may also preclude the opportunity to systematically modify the experiment to assess the robustness of the results; it may be difficult to decide if the works provide far-reaching or anecdotal results linked to the idiosyncrasies of the setup. Furthermore, and this is rarely mentioned or exploited, simulated experiments are highly conductive of the dissemination and the reproduction of research. Thus, simulations make sense for practical and scientific purposes.

One has to acknowledge that, from an embodied perspective, the premise of the *reality gap* problem—relying on a close-to-reality simulation to optimize the behaviour of a real robot—is a methodological error. Still, simulations are required when actual evaluations are prohibitive, consume limited or unique resources, are too dangerous or are simply impossible (for instance, developing a morphology and a gait for a lunar rover).

The *reality gap* problem is not limited to robotics, to machine learning, or to a simulation/reality contrast. It is present every time learning cannot happen in the environment where the exploitation takes place. The aerodynamics of cars and planes are tested in simulation before being tested in wind turbines, plane pilots train in simulators, astronauts train for spacewalks in pools, firefighters create mock emergency situations, surgeons train on cadavers or animals. In all these instances, a balance must be found to make the mock environment as close as possible from the real situation

[1] Jakobi (1997) formulated it nicely: *'any real-life simulation will differ from a perfect copy of the real world on two counts: It will model only a finite set of real-world features and processes, and those features and processes that it does model, it will model inaccurately.'*

to ensure transferability while guaranteeing safety and managing costs and resources.

The *How* of the Reality Gap

In robotics, the *reality gap* is overwhelmingly studied in the context of the optimization of controllers in simulation to be transferred on a real robot, in particular in the context of evolutionary robotics (Nolfi et al. 1994; Koos et al. 2013). This is not always the case; Gongora et al. (2009) report evolving the behaviour of a real helicopter that had to be restrained during learning to avoid damage, perturbing the conditions enough so that the untethered behaviour did not perform as well as the tethered one.

The most straightforward way to deal with the reality gap is to create the most accurate simulation possible. As we have outlined, this is fraught with problems, and can lead to very expensive simulations. Jakobi (1997) proposes to identify the minimal set of features responsible for the behaviour, and to simulate only those. Instead of building one simulation, he proposes to create many, with random variations, to make evolved controllers robust to the specificity of one or the other.

Some approaches improve the simulator during learning based on empirical observations (Zagal et al. 2008; Bongard and Lipson 2005; Bongard, Zykov et al. 2006; Koos et al. 2009). These approaches, when creating a simulation from scratch, has their roots in the domain of *surrogate* models (Sacks et al. 1989; Barton 1998; Jones et al. 1998). A model of the function to optimize is learned empirically, and the optimization takes place on the model rather than in the real environment. Surrogate models only differ from forward models in their intent: surrogate models aim at being useful for optimization while being cheap to evaluate, while forward models typically strive for accurate predictions.

Some methods consider the simulator as fixed, and evaluate the mapping between the simulator and the reality. This allows to estimate the discrepancy between the two, and to only perform simulated optimization in areas when the discrepancy is low (Koos et al. 2013).

In all of those approaches, an underlying assumption is that the simulation can, at least sometimes, be reasonably physically accurate:

> *even if a simulation model is somehow inaccurate, it also contains realistic parts as it is designed to accurately mimic some physical phenomena. Efficient behaviours that mainly rely on these realistic parts of the simulation model should transfer pretty well onto the physical device and then achieve good performances in reality.*
>
> Koos et al. (2013, p. 123)

9.2 Crude Simulations

We take a different perspective. As the complexity of robot's hardware increases, and environments become significantly more complex than a perfectly smooth and flat surface, obtaining physically accurate behaviour in simulation becomes difficult, and will necessitate important modelizing efforts. The risk is to limit the morphology of the robots we create to the methods we have available.

Instead of spending ever increasing efforts to create a realistic simulations, we go in the opposite direction; we search for the most simple, most crude simulation that still affords us an exploratory advantage through reuse. Our objective is not to find an optimal behaviour or even a good behaviour, but to efficiently discover diversity in the environment.

We took the experimental setup of section 7.1 and created a simplified kinematic simulation of it. The arm has been replaced by the forward kinematic computation of the position of the centre of the end-effector according to the forward kinematic model. From this, we compute the trajectory of the end-effector by feeding the kinematic model with the joint trajectories produced by the motor primitives. The object is approximated to its axis-aligned bounding box. If the trajectory of the end-effector enters the bounding box, the velocity of the end-effector is averaged from its last 10 positions, and the displacement of the object is a vector of the same direction as the velocity of the end-effector. The norm of the displacement is proportional to the end-effector velocity, and inversely proportional to the mass set for the object. There is no floor to interact with, the displacement of the object is done in three dimensions, and then projected to the ground plane.

This model is highly unrealistic in many ways. There is no way to have objects with different geometry. No contact is simulated except the one between the object and the end-effector—and it does not even take into account where the trajectory of the end effector hit the object with respect to its centre of mass.

The kinematic simulation is run for 1000 timesteps using a normal exploration strategy ($K_{\text{boot}} = 300$, $d = 0.05$). The exploration is then transferred to the full simulation scenario with a ball placed at the same place as the object in the kinematic model. The exploration on the full simulation is parametrized normally ($K_{\text{boot}} = 300$, $d = 0.05$, $p_{\text{transfer}} = 50\%$). The results are available Figure 9.1.

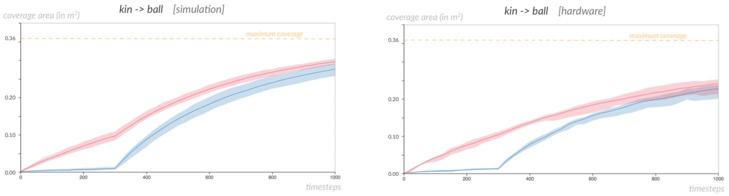


Figure 9.1: Even with a crude model, the reuse transfer is effective. Averaged over 25 repetitions for the simulation, and 4 repetitions for the hardware. [source code]

Even with a crude simulation devoid of most physical modelizations, the reuse strategy is able to take significant advantage of the exploration.

A Cruder Simulation

We simplify the previous simulation. Instead of computing the displacement of the object, the sensory response is only conditioned to the trajectory of the end-effector entering the bounding box. If that happens, a *random* value between 0 and 1 is returned. If not a random value between -1 and 0 is returned. The sensory signal has only one dimension.

Learning with such a poor sensory feedback is more difficult. The simulation has essentially become an indicator for a possible collision. Yet, reuse still provides an improvement (Figure 9.2). As should be expected, the improvement is less than when the simulation is more informative.

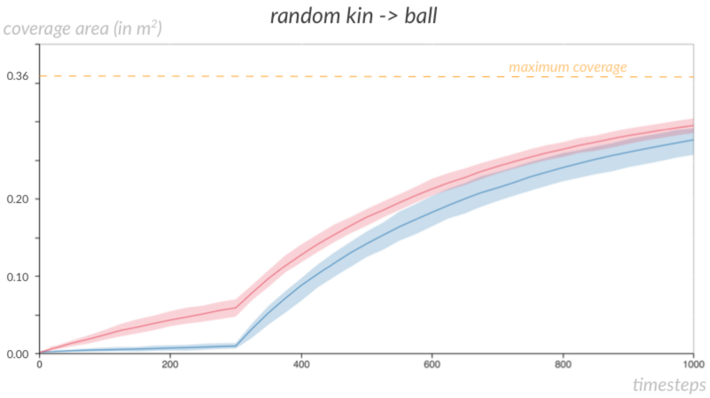


Figure 9.2: Even with a cruder model, the reuse transfer is still effective. Averaged over 25 repetitions. [source code]

Crossing The Smaller Reality Gap

A fortiori, reusing motor commands from the full simulated setup on the hardware setup should be easy. But we verify nonetheless. The results are available Figure 9.3.

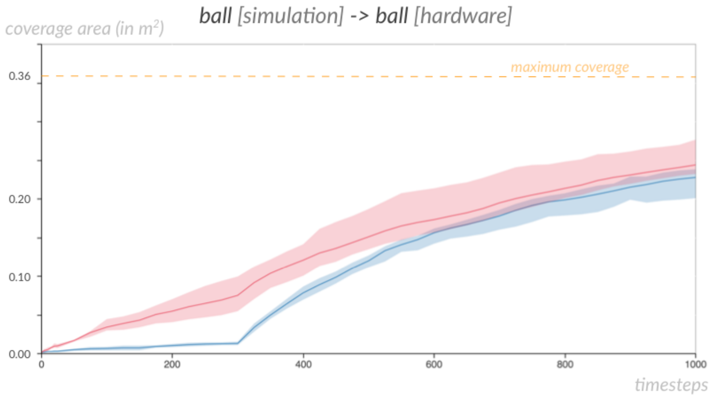


Figure 9.3: The hardware setup is able to take advantage of the motor commands reused from simulated exploration. Of course, since the simulated environment is shared, the feat is hardly impressive. Averaged over 4 repetitions. [source code]

Discussion

A weakness of the work presented here is that even a simple forward kinematic model usually display good performance on a rigid body robotic arm. Although we removed many aspects of the physical simulation, we retained the essential part. The discrepancy then, between a collision detected in simulation and one produced in reality is low. This easily explains the results obtained. And while we claimed to not assume that the simulation needs to be physically accurate, it actually is, but qualitatively.

The way the object displacement is computed in the first crude simulation can also be criticized. Although it seems that, by not taking into account any geometry of the object, or not considering the floor we have lost much information, the direction of the displacement is directly correlated to the direction of the end-effector when a collision happens. This sensory feedback is probably richer in information that the final position of the object in the physical simulation. It is also a signal that is easier to learn. The first crude simulation could be considered as scaffolding, that offers knowledge of a pivotal aspect of the interaction—the direction and velocity of the colliding tip of the arm just before the collision—that was hidden before.

Of course, these criticisms can also be considered positively: yes, the crude models are qualitatively accurate with regards to the presence of a collision, and reuse is able to take advantage of a merely qualitative, rather than numerical, accuracy. Yes, this is scaffolding, and reuse is taking advantage of it without the experimenter noticing it: reuse do not need to be explained how the two environments relate to each other.

Compared to previous works, the context in which we consider the reality gap problem is different. We are not trying to transfer controllers while conserving performances; we are looking for an exploratory advantage. While we presented our work in the context of the reality gap problem, it is not comparable with the other methods we discussed section 9.1: it addresses a different problem.

Yet, reuse could be of use for learning controllers. One could derive the first population of an evolutionary algorithm by reusing the genetic code of a set of candidate solutions that generated a diversity of behaviour during a cheap, pared-down simulation. This could help mitigate the early convergence and bootstrap problems. In a single agent optimization scenario, using reuse from a simulation would not provide the best controller. But using reuse increases the probability the robot is given access to controllers early in the exploration that are close to good solutions, compared to a random motor babbling exploration. The simulation that provides those initial solutions does not need to model all aspects of the real robot. Actually, it can be arbitrarily selective about which features of reality it decides to model. The transfer should stay robust, as long as a diversity of candidate solutions is transferred.

In a self-sufficient perspective, the crude simulations can be considered as cognitive models. The simplicity and relaxed qualitative nature of the correspondence to reality that they must provide makes their acquisition by a self-sufficient robot more reasonable than a full-featured realistic simulation. In that context, the results suggest another way to engage with the reality gap problem. Instead of reproducing reality, cognitive simulations can do away with much of the realism, without losing their power to direct and inform behaviour. They pose as a reasonable artifice of cognition that allows agents to think about the world without having to predict or simulate it accurately. Cheap cognitive simulations can create diversity, and give robots—dare I say—*ideas*.

Exploring Ahead

Roboticists are demiurges.

They create bodies, the minds inside them, and, more often than not, the worlds around them.

That makes roboticists their own worst foes.

The risk is that roboticists, creating both problems and solutions, may tailor problems to solutions, and not the other way around. This may lead to inventing and investigating artificial problems that contribute little to advancing scientific inquiry, whilst systematically avoiding hard problems, by modifying them into easier ones each time seemingly insurmountable obstacles are encountered.

But there is another risk, more pernicious, and, more fundamental. It is to design robots from a human perspective, choosing features and characteristics that make sense for the external observer, but none for the robot itself and its egocentric experience of the world. In other words, the risk is that the features that make robots easy for humans to design, control and understand make it hard for the robot itself to interact with the world, and end up fundamentally limiting its capabilities.

An illustration of this is found in how robots are created: the hardware is usually created before the software. This allows to decouple the two activities, and hence, the somewhat different skillsets. And it allows to sweep aside the myriad of interactions that would need to be considered if the body and mind were designed together. The software here truly plays the part of the ghost in the machine, investing it, animating it after it was created. This way of proceeding is obviously suited to the robotic workcells of assembly lines. But this same design pattern, the one essentially used for smartphone development, is repeated for state-of-the-art robotic research platforms such as the iCub, the PR2 or the Baxter, where researchers are tasked to find out how to program finished and hardly reconfigurable hardware products.[2]

This paradigm works fine for a number of scientific endeavours but is problematic others, such as legged locomotion. Designing legs divorced from the gait algorithms that are used to actuate them has produced many robots that not only requires precise, low-latency, computationally expensive algorithms, but that are also brittle to unexpected, even if small, perturbations of their environment.

[2] As a result, many works in robotic algorithms list as a positive feature the ability to adapt to arbitrary hardware. Although seemingly desirable, the broader implications of such a goal make it a potentially dangerous one.

Evolutionary robotics has attacked this problem directly, by proposing algorithms inspired from natural selection to design robots directly from an evaluation of their behaviour, removing the human designer from the process, and allowing the morphology and the controllers to tailor themselves to one another.

Yet, even in evolutionary robotics, one key human element remains in the design process: the fitness function. It encodes the ultimate goal of the evolutionary process, and is decided beforehand by the designer, oftentimes in an extremely specific manner: the distance covered by the robot in a given amount of time in locomotion experiments, for instance. The most immediate consequence is to create one-trick robotic ponies. This is made worse by the tendency of evolutionary algorithms to routinely outsmart the designer by producing robots whose behaviour is unsuitable in ways that are not encoded in the fitness function. Examples include evolutionary processes exploiting bugs of the physic engine in simulated experiments, or producing real robots that cover the most ground by irreversibly damaging themselves. Those are serious concerns compounded by that setting a goal is not always the most effective way to reach it, as the work of Stanley and Lehman (2015) shows.

However, the most fundamental problem is elsewhere: the designer gets to choose the goals that will single-mindedly direct the activity of the robot. Problem is, it is not clear how qualified or well-positioned a human is for choosing the goals of a robot, an entity with vastly different embodiment and cognitive processes.

This is where some strands of developmental robotics try to distinguish themeselves from the rest of the robotics. They study robots that must create their own goals, using their own *motivational systems*.

Developmental robotics originated around the realisation that creating robot adults with fully-formed knowledge and skills out of the assembly line was too difficult. Programming commonsense, for instance, proved remarkably difficult. Observing that humans naturally acquire it during their childhood, it was proposed to create child robots, that would be equipped with learning abilities that would allow them to gather knowledge and skills that made sense for them, for their particular embodiment and environment.

Motivational systems, in turn, are to designer-set goals what learning abilities are to preformed knowledge. They are goal factories, the same way learning abilities are knowledge and skills factories. They allow robots to choose goals that make sense for their particular embodiment, environment, and current experience. Motivational systems also naturally compliment learning abilities, because there are too many things to learn in any sort of modestly complex environment; they allow to choose what activities to engage in, and therefore, what to learn and not to learn.

All this brings us to the subject of this thesis: exploration. Robots that choose their own goals, that acquire skills on their own need to explore the environment for two reasons. The first so that they can acquire experience, which can be in turn used to modify their behaviour (i.e. learning). The second, to discover new kind goals they can pursue.

In this thesis, we focused on exploration in sensorimotor spaces, that is spaces where the mapping between a motor action and its corresponding sensory feedback can be expressed. Also, we considered only exploration that is conducted by the robot itself, without any social guidance or any externally provided knowledge. Hence the title: 'Self-Exploration of Sensorimotor Spaces in Robots'.

The thesis had three goals. First, establish exploration as a scientific problem. Second, do a study of some simple exploration algorithms, and what impact different factors had on them, in order to provide a bedrock on which to think and build more elaborate exploration strategies. And third, start exploring some ways the exploration capabilities of an agent could improve over time, as experience accumulates. This thesis fulfills those goals, if only specifically.

To establish robotic exploration as a scientific problem, we start, in chapter one, at the very beginning: we define what a robot is, we explain the effect embodiment has on the robot experience of the world, and why all problems cannot be solved by an ambitious-enough simulation of the real world in the robot's head. We conclude that to be effective in the unstructured part of the real world, robots need to pass through an extended development phase in order to build up skills, knowledge and the commonsense needed to deal with future unexpected situations. During this development phase, exploration skills are crucial.

We then formalize the exploration problem: exploring is creating access to different aspects of the environment. Exploration is not solely spatial: you can explore the responses an object gives to external input, such as the different sounds it can make. This definition allows us to draw an important distinction between exploring and learning. Learning is modifying your behaviour as a result of experience. As such, learning is independent from exploration; you can learn without exploring: this is what a weather prediction system does. And you can explore without learning: this is what some robot vacuum cleaner do; they manage to cover and clean a room without ever learning its shape. Of course, most of the time, we want to combine learning and exploring.

Now, to make a scientific problem out of exploration, we need to be able to evaluate it using a quantitative measure. If exploration is creating access to different aspects of the environment, then one way to evaluate it is measuring the *diversity* of the sensory feedback the robot is able to produce. Diversity is a great measure for a number of reasons: it is a concept that can adapt to almost any setting. It is intrinsic, i.e. it can be measured by the robot itself, and without disturbing its behaviour in any way—contrary to, for instance, a measure of the robot's prediction abilities. This, as a side-effect, allows to envision sharing common experimental setups with other domains where peering into the explorer's thought process is hard, such as cognitive science experiments on children.

This leads us to an important and inescapable point: the related work. The notion of exploration and diversity has seen scant explicit attention in robotics, outside of spatial exploration[3]. But many neighbouring domains of developmental robotics feature informative works. In developmental robotics itself, the study of intrinsic motivation is relevant; diversity can be used and is used in some algorithms of this thesis as an explicit intrinsic motivation. Moreover, diversity in computer science has seen a steady rise in interest since 2000 from many different areas such as ensemble classifiers, swarm optimization and recommender systems. And in cognitive science, exploratory behaviour has been the subject of important works, even if almost all the quantitative data comes from spatial exploration experiments.

For our experiments, we introduced a diversity measure called τ-coverage. It measures the volume of the union of balls of radius τ centered around the observed sensory feedback points in the sensory space. If the sensory feedbacks are diverse, they are far from each other, and overlap between the balls is low: the volume of the union is high. If the sensory feedbacks are similar, the overlap is high and the volume lower, for the same amount of feedback points.

The second goal was to study exploration algorithms. The idea there was to create one of the simplest algorithm possible, and study it under different conditions. The simplicity was warranted by two factors: first, it allowed to understand the results in their every detail without suspending intuition. The behaviour of linear weighted regression or more complex learning algorithms in high-dimensional spaces can be complex, which is why we opted for a simpler perturbation-based nearest-neighbor learning method. And second, it was hoped that being simple, the lessons learned and the intuition gained could be carried over a wider range of situations than a more complex, more specific algorithm.

One of the first contributions of the study was to clarify that exploring the motor space was inefficient because of the *combined* contributions the high-dimensionality *and* the heterogeneous distribution of the redundancy of the sensorimotor space (i.e., how many different motor commands produce the same sensory feedback). High-dimensionality alone does not make exploring the motor space ineffective.

Next, we systematically analysed the contribution of each aspect of the algorithm. The impact of the distribution of goals was studied, outlining the potential directed methods represent (in most of the thesis, goals are chosen at random by the algorithms). The effects of a bad inverse model were shown, and an algorithm for boundless goals space was introduced.

The next experiments focused on showing how even rudimentary implementations of motor synergies, developmental constraints and external demonstration could positively affect the exploration. One takeaway is that improving embodiment potentially offers cheaper and larger gains that improving the learning performance.

So far, all the algorithmic variations studied did not make use of any explicit in-

[3] Spatial exploration is a highly specific case of exploration, and is mainly distinguished from general sensorimotor exploration in that movement in the sensory space is already explicitly mastered.

trinsic motivation measure. Diversity was only used as an evaluation tool. In chapter four, we introduce an algorithm that uses diversity to direct which exploration strategy to use among a set, and doing so can adapt to different situations as well as any fixed mixture of the strategies.

The third goal was to investigate ways for the exploration capabilities of an agent to improve over time, as experience accumulates.

To understand the underlying challenge here, one must consider that a successful exploration of a given environment should give access to different features of said environment, i.e., from the point of view of the robot, produce a diversity of sensory feedbacks. To produce a diversity of sensory feedbacks efficiently, one would need knowledge of the dynamics of the environment, in order to avoid its inherent redundancy, i.e., to avoid executing actions that produce the same effects. Pushing and pulling on a closed door illustrates this point: two different actions that produce the same effect—hence producing no diversity in sensory feedback—and afford new knowledge about the environment. Should the state of the door had been known beforehand, the robot could have engaged with other actions, more likely to produce diversity. Therefore, producing diversity faces a chicken-and-egg problem: the knowledge needed to perform an efficient exploration is the knowledge the exploration is supposed to produce in the first place. This means that the exploration process can feed itself, but can also remain stalled if incapable to produce informative interactions, leading to long early periods of poor exploration in challenging environments.

This what drove us to find a solution to improve early exploration. To this end we introduced the *reuse* method, that leverages experience acquired in a past environment, to explore a new one. The core idea is to select motor commands that produced a diversity of sensory feedbacks in the past environment, and to reexecute them in the new one. This method has the benefit of being conceptually simple, and to be agnostic about the sensory modalities of the past and new environments, which can be arbitrarily different. The exploration strategy or learning algorithm used in the past environment need not to be the same in the current one either: the method can leverage data that has been arbitrarily collected. The only constraint is that motor commands executed in the past environment can be reexecuted in the new one.

The rationale behind the *reuse* method can be understood by considering how redundancy makes to different motor commands produce the same effect in the environment: either by body redundancy, or environmental redundancy. The body redundancy makes different motor commands produce the same movements: the robot applies the same forces on the environment. Environmental redundancy leads differents forces to the same effects, as the closed-door example illustrates. Typically, different effects both avoid body and environmental redundancy. When changing from one environment to another, environmental redundancy is not conserved, but the body redundancy is in most cases. Moreover, if the environments are similar, some of the environmental redundancy generally overlap. Therefore, by reusing a set of motor commands that generated a diversity of effects, the *reuse* method capitalizes on knowledge gained about body redundancy, and opportunistically on the environmental one.

We conducted experiments to demonstrate the viability of the approach on a real robot manipulating different objects in augmented reality. The results showed conclusively that the *reuse* method is effective when reusing experience gained from interacting with one object (a ball), to explore another object with a significantly different behaviour (a cube). The method is also robust to dissimilar environments, where diversity from one environment does not transfer well to the other. Moreover, we established that choosing which motor commands to reexecute according to the diversity of the sensory feedbacks they produced is better than random.

In the previous experiments, *reuse* proved to improve early exploration. But after enough time, whether using *reuse* or not, the exploration process produced similar results. To show that *reuse* could do more than only improve early exploration, we designed an experiment to show that it could make explorable an environment that would not be otherwise. An interesting part of the experiment was that the exploration was shaped not by changing a reward function, but only by manipulating the environment and the saliency of the objects in it, much like a caretaker would do with a child.

Finally, we got interested by the applicability of the *reuse* method to situations where the exploration of the past environment happened entirely in simulation, while the exploration of the new one would happen in the real world, on a real robot. Transferring results from simulation to reality has proven difficult in robotics, a problem known as the *reality gap*. The results obtained, although warranting more work, are excellent. They raise the prospect of using cheap and crude simulations of reality as efficient cognitive artifacts for self-sufficient robots to explore the real world better.

This is where the thesis end. From there, what is the way forward? There are three research directions that stand out: diversity in robotics, interdisciplinary work with cognitive sciences, and evolutionary developmental robotics.

First, diversity in robotics. In 1255, in his Commentary on Sentences, Thomas Aquinas argued that while an angel is better than a stone, it does not follow that two angels is better than one angel and one stone[4]. A modernised version of Aquinas' argument is proposed by Nehring et al. (2002): 'A human being is more valuable than a chimpanzee. It does not follow, however, that 6,000,130,000 humans and no chimpanzee are more valuable than 6,000,000,000 humans and 130,000 chimpanzees.' Diversity has value. Such an observation can be made in domains as different as biodiversity, art, hiring practices, investment portfolios, search engines results, ensemble classifiers, and even, scientific progress. In Lehman, Clune et al. (2014), Pierre-Yves Oudeyer remarked that 'because we don't deeply understand intelligence

[4] For latin readers out there: 'quod quamvis Angelus absolute sit melior quam lapis, tamen utraque natura est melior quam altera tantum' (Lib. 1 d. 44 q. 1 a. 2 ad 6)

or know how to produce general AI, rather than cutting off any avenues of exploration, to truly make progress we should embrace AI's "anarchy of methods".' In other words, when fumbling in the dark, diversity is a powerful tool.

It is tempting there to apply this lesson to developmental robotics, and that is actually what this thesis tries to do: developmental robots, plunged in the complexity of the real world, and with no choice but to make sense of it with their learning and exploration capabilities, must fumble in the dark for a time. The lack of literature on diversity in developmental robotics pales in comparison to the potential benefits it could bring.

There are, however, many ways to abuse the lesson. First, diversity for diversity's sake is hardly justified, however intrinsically valuable it may be. In particular, a motivational system only driven by diversity seems like a poor idea. Some have argued that since simplicity is in finite supply, diversity-driven development will naturally lead to discovering ever more complex phenomena. The scarceness of simplicity, however, has never been justified outside of toy examples, and simple things to discover and learn in the real-world seem to be plentiful-enough to fill multiple lifespans. All this conspire to suggest that robotic motivational systems should embrace a *diversity* of motivations, with diversity being one of them. Competing and complementary motivations should lead to behaviour alternating between broad exploration, where new features of the world are discovered, and more focused study, where specific skills are mastered.

Second is the issue of how to use the experience collected through diversity-driven exploration. In this thesis, we have shown, through the *reuse* method, that this experience is precious to conduct more exploration of other environments. But exploration is hardly the only behaviour of a developmental robot. The question of how to marshall and apply experiences gained by diversity for precise problem-solving, and whether it is competitive with more directed approaches, remains open.

Finally, many specific issues about diversity are not yet satisfactorily answered. Diversity-driven exploration differs from novelty-driven exploration in that novelty-driven approaches cannot explicitly control the amount of diversity they produce. Maintaining a certain level of behavioural diversity, especially when changing environmental conditions decrease the options available to the robot, can only be obtained from the global perspective diversity affords, not using the local one offered by novelty-based approaches. Still, diversity is more computationally expensive: when is it necessary versus simpler novelty-based motivations? What are good diversity measures? Does diversity makes any sense in high-dimensional sensory space, or should it always be supported by low-dimensional abstract representations?

Answering those questions is not easy; a possible source of intuition is to turn to cognitive science. How do children use diversity during development? This is the second research direction that seems promising.

It is remarkable that amongst all the literature on play, exploration, and problem-solving in children and animals, quantitative measurements of the diversity of the interactions they engage in and of the solutions they try out, is almost completely absent. Studies usually stop at vague qualitative descriptions. Quantitative studies on behavioural diversity in exploration could shed useful light on how to design robotic motivational systems. Moreover, this line of research, by its compatible methodology promise to be able to conduct similar experiments on humans and robots, potentially leading to fruitful exchange and emulation between the two domains.

The third avenue of research is evolutionary developmental robotics, affectionally called 'evo-devo-robo'. Evolutionary robotics mimics the natural selection process, while developmental robotics mimics the morphological and cognitive development of biological systems. Most of their respective work, however similar, has remained separated so far. Given the interest of AI for producing human-level intelligence, this separation is puzzling; after all, the only known examples of entities possessing human-level intelligence were created by a combination the two processes.

Combining evolutionary and developmental robotics raises a tremendous issue: time. The typical timescales of development and evolution—lifespans and eons, respectively—already have their respective discipline struggling. Combining both seems therefore completely intractable, whether simulations are involved or not. The way to look as this is to consider that the scale of the problem is so important that it won't change as technological progress piles-up in the next, let's say, 50 years: waiting does not help.

Another objection is to argue that evolutionary and developmental robotics are still young disciplines, and not yet ripe for being combined. Although mostly speculative, this argument could end up being true. But the difficulties encountered along the way could shed precious light on shortcomings in the two disciplines that are hard to detect otherwise.

Adding a long development phase to evolutionary robotics could give rise to new, more complex dynamics in the evolutionary process, and a better selection process. Some ground has already been covered by Bongard (2011), who showed that morphological development could act as a sieve filtering brittle behaviour in legged locomotion. Conversely, developmental robotics could benefit from an overarching evolutionary process, which could reduce the arbitrary decisions current researchers have to make about the representation and learning abilities they give to robots.

Evolutionary developmental robotics certainly represents a tremendous challenge, but the results that are there to reap are equally so. This is a domain that we simply cannot afford not to investigate.

If only because it promises to chip away at the demiurgic nature of roboticists. Roboticists are demiurges; evo-devo-robo is part of the solution.

A
Volume of Union of Geometrical Objects

Computing the union of many disks is expensive. Here we propose a more computationally efficient way to compute the diversity of an effect, based on grid partitioning.

A.1 Volume of the Union of Hyperballs

Computing the volume of the union of an arbitrary set of hyperballs is not a straightforward problem.

A.1.1 The Klee Measure Problem

The Klee's measure problem (KMP) is an open problem of computational geometry, introduced in 1977 by Victor Klee (Klee 1977). It is stated as follow:

> Given a set B of n axis-parallel boxes (hyperrectangles) in $\mathbb{R}^d$, compute the volume of the union of B. (Chan 2013)

Under the Euclidean norm, i.e. the L^2 norm[1], the KMP is the equivalent of the problem the volume of the union of hyperballs but for hyperrectangles. Under the L^∞

[1] For any real number $p \geq 1$, the L^p norm of a given vector $\mathbf{x} \in \mathbb{R}^d$ is $\|\mathbf{x}\|_p = (|x_0|^p + |x_1|^p + \ldots + |x_{d-1}|^p)^{\frac{1}{p}}$. The L^2 norm is thus the familiar Euclidean norm.

norm, which is defined as $\|\mathbf{x}\|_\infty = \max\{|x_0|, |x_1|, ..., |x_{d-1}|\}$, the two problems are identical.

The KMP has been a continuous subject of study in the computational geometry community. For $d = 1$, the original $\mathcal{O}(n \log n)$ algorithm by Victor Klee was proven optimal in 1978 (Fredman et al. 1978). Bentley (Bentlley 1977) proposed the problem for $d = 2$ and provided a $\mathcal{O}(n \log n)$ algorithm as well—*a fortiori* optimal. In 2013, an algorithm for $d \geq 3$, was provided by Chan (2013) with a time complexity of $\mathcal{O}(n^{\frac{d}{2}})$, improving on previous works. As the only known lower bound for any dimension is $\Omega(n \log n)$, the existence, for $d \geq 3$, of faster algorithms than Chan (2013) remains an open problem.

Chan (2013) proposed a slightly faster algorithm in $\mathcal{O}(n^{\frac{d+1}{3}} \log^{\mathcal{O}(1)} n))$ for the special case of unit hypercubes, which applies to our case. This imply for $d \leq 4$, the algorithm is subquadratic.

Yet, the KMP, behind an apparent simplicity, proves costly, and makes the use of an exact coverage performance based on the volume of the union of polygons impractical for $d \geq 4$, over long periods.

A.1.2 Union of Hyperballs

For the problem of computing the volume of the union of hyperballs, exact methods exist using Voronoi Power Diagrams (Cazals et al. 2011), that partition the space into as many areas as there are balls; in each area, the center of only one ball is present, and the contribution of this ball to the overall volume can be computed independently of the others (Kim et al. 2012). There also are approximate methods based on Monte-Carlo sampling (Till et al. 2009).

B
Grid Diversity

Computing the union of many disks is expensive—see appendix A, and is impractical in dimensions higher than four. To that end, we introduce a more computationally efficient method for estimating strategy diversity, that may be used in higher dimensions, based on *grid partitioning*.

B.1 Grid Diversity

Definition 7. *Given a set of effects $E \subset \mathbb{R}^d$, a grid overlaying $\mathbb{R}^d$, and a new effect $\mathbf{y}$, let's define k as the number of observed effects already contained in the cell of the grid $\mathbf{y}$ belongs to (excluding $\mathbf{y}$). The diversity of the new effect $\mathbf{y}$ is then defined as:*

$$\text{diversity}(\mathbf{y}, E) = 2^{-k}$$

This definition of diversity values effects that create new cells, or that belong to cells with a low number of already observed effects. The total diversity value that each cell represents is $\sum_0^\infty 2^{-k} = 2.0$.

Finding the coordinates of the cell where a effect belongs takes $\mathcal{O}(d)$ steps, and then finding k is in $\mathcal{O}(\log n)$—we only store in memory the size of non-empty cells, using a hashmap, ensuring a $\mathcal{O}(nd)$-space complexity, we obtain a $\mathcal{O}(nd \log n)$-time complexity for the grid diversity algorithm.

We use this definition of effect diversity to define *strategy diversity* in the same way as in section 4.2, and reruns the same experiments. The width of a cell is set to five millimetres.

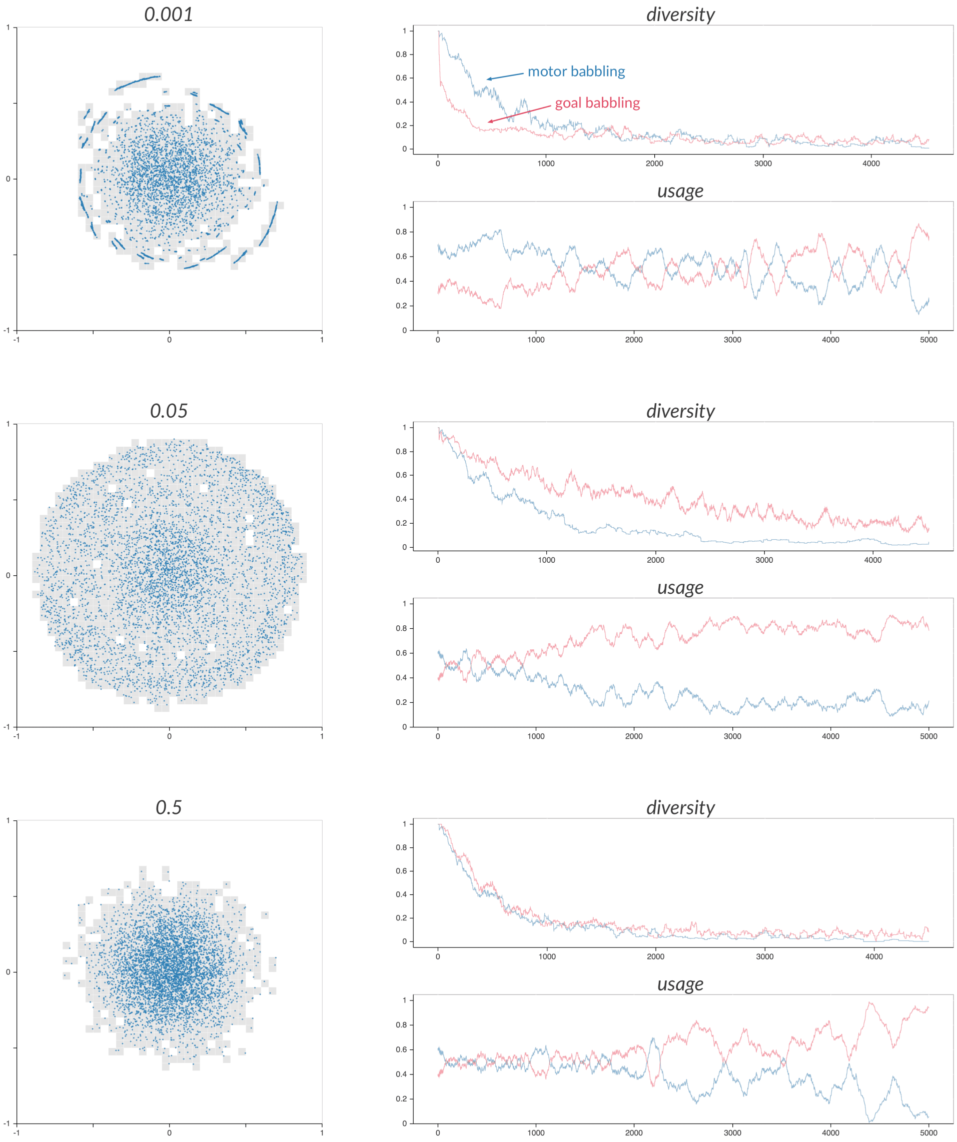


Figure B.1: [source code]

C
A More Sophisticated Inverse Model

In section 3.1.1, we defined a simple inverse model for the two-dimensional arm. In some experiments on the interaction setup of the second part, we use a more sophisticated inverse model, based on a optimization routine, L-BFGS-B (Byrd et al. 1995; Zhu et al. 1997), and a predictor, Locally Weighted Linear Regression (LWLR) (Cleveland et al. 1988; Atkeson et al. 1997a,b). The algorithmic change in the exploration strategy is a simple replacement of the Inverse() routine.

C.1 Forward Model

To approximate the function f from a set of observations, we employ Locally Weighted Linear Regression (LWLR) (Cleveland et al. 1988; Atkeson et al. 1997a,b), a incremental machine learning algorithm. Although LWLR is considerably more sophisticated than the inverse model used in the first part, it is still a simple method compared to the state-of-the-art. Here, again, the absolute performance is of little concern, as we are interested in comparing different exploration strategies. Still, LWLR is reasonably robust (Munzer et al. 2014) for the learning tasks we are considering. The main differences between LWLR and our perturbation-based inverse model are that LWLR is able to extrapolate—how far the goal is from the data is taken into account—, and LWLR uses, and needs, multiple observations to predict the outcome of an hypothetical input.

Given a set of observations $E = \{(\mathbf{x}_t, \mathbf{y}_t)\}_{0 \leq t < N}$ where for each i, $f(\mathbf{x}_t) = \mathbf{y}_t$, and a query vector $\mathbf{x}_q$, for which we wish to predict the effect, we compute, for each point $\mathbf{x}_t$, the euclidean distance to $\mathbf{x}_q$ and derive a gaussian weight w_t :

$$w_t = e^{\frac{-\|\mathbf{x}_t - \mathbf{x}_q\|^2}{\sigma^2}}$$

We consider the matrices X with $X_{i,k} = (\mathbf{x}_i)_k$, Y with $Y_{i,k} = (\mathbf{y}_i)_k$, and $W = \text{diag}(w_0, w_1, ..., w_n)$, and compute :

$$\boldsymbol{\beta} = (X^T W X)^{+}(X^T W Y)$$

where $X^T W X$ is a positive definite symmetric matrix, and $(X^T W X)^{+}$ is its Moore-Penrose inverse (Penrose et al. 1955).

Then:

$$\mathbf{y}_e = \boldsymbol{\beta}^T \mathbf{x}_q$$

$\mathbf{y}_e$ is the LWLR estimate of $\mathbf{x}_q$, given the observed data E. We define the function PredictLWLR$(\mathbf{x}_q, E)$ that compute $\mathbf{y}_e$ for any $\mathbf{x}_q \in M$ given E.

In our implementation, σ, which control the locality of the regression, is dynamically computed. With m as the dimension of the motor space, we define a constant $k = 2m + 1$, and compute σ as the average distance of the k closest points of the query vector $\mathbf{x}_q$. All other points of E besides the k closest neighbours are given a weight of zero.

C.2 Inverse Model

Given a query point $\mathbf{y}_q \in S$, we want to produce a motor command $\mathbf{x}_e \in M$ so that $\|f(\mathbf{x}_e) - \mathbf{y}_q\|$ is minimal.

Since M is a hyperrectangle of $\mathbb{R}^m$, we use L-BFGS-B (Limited-memory Broyden–Fletcher–Goldfarb–Shanno Bound-constrained (Byrd et al. 1995; Zhu et al. 1997); we used version 3.0 (Morales et al. 2011)), a quasi-Newton method for bound-constrained optimization, to minimize the error. L-BFGS-B use an approximation of the Hessian matrix to direct the optimization (because the Hessian cannot be directly computed, it is approximated using finite differences). We approximate $\|f(\mathbf{x}_e) - \mathbf{y}_q\|$ with $\|\mathbf{y}_q - \text{PredictLWLR}(\mathbf{x}, E)\|$ and L-BFGS-B, in turns, approximates :

$$\text{argmin}_{\mathbf{x} \in M}(\|\mathbf{y}_q - \text{PredictLWLR}(\mathbf{x}, E)\|)$$

The optimization process is initialized with the motor command corresponding to the closest neighbour of $\mathbf{y}_q$ in the set of observations.

Algorithm 6: InverseLBFGSB-LWLR($\mathbf{y}_g, E$)

```
Data:
    • $E = \{(\mathbf{x}_t, \mathbf{y}_t)\}_{0 \leq t < N} \in (M \times S)^N$, past observations.
    • $\mathbf{y}_g \in S$, a goal.
Result:
    • $\mathbf{x}_e \in M$ a motor command.

$\mathbf{x}_e = \textsc{Minimize}\text{LBFGSB}_{\mathbf{x} \in M}(\|\mathbf{y}_q - \textsc{Predict}\text{LWLR}(\mathbf{x}, E)\|)$
```

Bibliography

Adler, B. Thomas, Alfaro, Luca de, Mola-Velasco, Santiago M., Rosso, Paolo and West, Andrew G. (2011). Wikipedia Vandalism Detection: Combining Natural Language, Metadata, and Reputation Features. In: *Computational Linguistics and Intelligent Text Processing*. Springer Science + Business Media, 277–288. doi:10.1007/978-3-642-19437-5_23 [page 47].

Adolph, K. E., Karasik, L. and Tamis-LeMonda, C. S. (2010). Motor Skills. In: *Handbook of Cultural Developmental Science*. Ed. by Marc Bornstien. New York: Taylor & Francis. Chap. 5, 241–302. isbn: 978-1-84872-871-4 [pages 100, 101].

Agrawal, Rakesh, Gollapudi, Sreenivas, Halverson, Alan and Ieong, Samuel (2009). Diversifying search results. In: *Proceedings of the Second ACM International Conference on Web Search and Data Mining - WSDM '09*. Association for Computing Machinery (ACM). doi:10.1145/1498759.1498766 [page 118].

Alexandridis, Georgios, Siolas, Georgios and Stafylopatis, Andreas (2015). Accuracy Versus Novelty and Diversity in Recommender Systems: A Nonuniform Random Walk Approach. In: *Recommendation and Search in Social Networks*. Springer Science + Business Media, 41–57. doi:10.1007/978-3-319-14379-8_3 [page 119].

Algara-Siller, G., Lehtinen, O., Wang, F. C., Nair, R. R., Kaiser, U., Wu, H. A., Geim, A. K. and Grigorieva, I. V. (2015). Square ice in graphene nanocapillaries. In: *Nature* 519.7544, 443–445. doi:10.1038/nature14295 [page 97].

Allee, W. C. (1931). Animal Aggregations: A Study in General Sociology. In: *Journal of Educational Sociology* 5.2, 130. doi:10.2307/2961735 [page 109].

Anderson, S. O., Wisse, M., Atkeson, C. G., Hodgins, J. K., Zeglin, G. J. and Moyer, B. (2005). Powered bipeds based on passive dynamic principles. In: *5th IEEE-RAS International Conference on Humanoid Robots, 2005*. IEEE. doi:10.1109/ichr.2005.1573554 [page 66].

Anitescu, M. and Potra, F. A. (1997). Formulating Dynamic Multi-Rigid-Body Contact Problems with Friction as Solvable Linear Complementarity Problems. In: *Nonlinear Dynamics* 14.3, 231–247. doi:10.1023/a:1008292328909 [page 212].

Anitescu, Mihai (2005). Optimization-based simulation of nonsmooth rigid multibody dynamics. In: *Mathematical Programming* 105.1, 113–143. doi:10.1007/s10107-005-0590-7 [page 212].

Argall, Brenna D., Chernova, Sonia, Veloso, Manuela and Browning, Brett (2009). A Survey of Robot Learning From Demonstration. In: *Robotics and Autonomous Systems* 57.5, 469–483. doi:10.1016/j.robot.2008.10.024 [page 151].

Arkin, Ronald (1998). *Behavior-based robotics*. Cambridge, Mass: MIT Press. isbn: 9780262011655 [page 62].

Artac, M., Jogan, M. and Leonardis, A. (2002). Incremental PCA for on-line visual learning and recognition. In: *Object recognition supported by user interaction for service robots*. Institute of Electrical & Electronics Engineers (IEEE). doi:10.1109/icpr.2002.1048133 [page 115].

Arthur, Wallace (1990). *The green machine : ecology and the balance of nature*. Oxford, UK Cambridge, Mass., USA: B. Blackwell. isbn: 9780631178538 [page 97].

Asada, M., Hosoda, K., Kuniyoshi, Y., Ishiguro, H., Inui, T., Yoshikawa, Y., Ogino, M. and Yoshida, C. (2009). Cognitive Developmental Robotics: A Survey. In: *IEEE Trans. Auton. Mental Dev.* 1.1, 12–34. doi:10.1109/tamd.2009.2021702 [pages 53, 66].

Asadi, Mehran and Huber, Manfred (2007). Effective Control Knowledge Transfer through Learning Skill and Representation Hierarchies. In: *IJCAI*. Vol. 7, 2054–2059 [page 188].

Ashby, W. Ross (1940). Adaptiveness and Equilibrium. In: *The British Journal of Psychiatry* 86.362, 478–483. doi:10.1192/bjp.86.362.478 [page 99].

Ashby, W. Ross (1947). Principles of the Self-Organizing Dynamic System. In: *The Journal of General Psychology* 37.2, 125–128. doi:10.1080/00221309.1947.9918144 [page 99].

Ashby, W. Ross (1960). *Design for a Brain*. Springer Netherlands. doi:10.1007/978-94-015-1320-3 [page 96].

Ashby, W. Ross (1962). Principles of the Self-Organizing System. In: *Transactions of the University of Illinois Symposium*. Ed. by H. Von Foerster and Jr. (eds.) G. W. Zopf. Pergamon Press, 255–278 [page 96].

Atkeson, Christopher G., Moore, Andrew W. and Schaal, Stefan (1997a). Locally Weighted Learning. In: *Lazy Learning*. Springer Netherlands, 11–73. doi:10.1007/978-94-017-2053-3_2 [pages 124, 138, 243].

Atkeson, Christopher G., Moore, Andrew W. and Schaal, Stefan (1997b). Locally Weighted Learning for Control. In: *Lazy Learning*. Springer Netherlands, 75–113. doi:10.1007/978-94-017-2053-3_3 [pages 124, 138, 243].

Auer, Peter, Cesa-Bianchi, Nicolò, Freund, Yoav and Schapire, Robert E. (2002). The Nonstochastic Multiarmed Bandit Problem. In: *SIAM J. Comput.* 32.1, 48–77. doi:10.1137/s0097539701398375 [pages 163, 164, 169].

Auerbach, Joshua E. and Bongard, Josh C. (2010). Evolving CPPNs to grow three-dimensional physical structures. In: *Proceedings of the 12th annual conference on Genetic and evolutionary computation - GECCO '10*. Association for Computing Machinery (ACM). doi:10.1145/1830483.1830597 [page 69].

Ay, N., Bertschinger, N., Der, R., GĂźttler, F. and Olbrich, E. (2008). Predictive information and explorative behavior of autonomous robots. In: *The European Physical Journal B* 63.3, 329–339. doi:10.1140/epjb/e2008-00175-0 [page 111].

Balch, Tucker (1997). Learning roles: Behavioral diversity in robot teams. In: *College of Computing Technical Report GIT-CC-97-12, Georgia Institute of Technology, Atlanta, Georgia* 73 [page 118].

Balch, Tucker and Parker, Lynne E., eds. (2002). *Robot Teams: From Diversity to Polymorphism*. Natick, MA, USA: A. K. Peters, Ltd. isbn: 1568811551 [page 118].

Baldassarre, Gianluca (2011). What are intrinsic motivations? A biological perspective. In: *2011 IEEE International Conference on Development and Learning (ICDL)*. Institute of Electrical & Electronics Engineers (IEEE). doi:10.1109/devlrn.2011.6037367 [pages 111, 113].

Baldassarre, Gianluca and Mirolli, Marco, eds. (2013). *Intrinsically Motivated Learning in Natural and Artificial Systems*. Springer Science + Business Media. doi:10.1007/978-3-642-32375-1 [page 111].

Ballard, Dana H. (1991). Animate vision. In: *Artificial Intelligence* 48.1, 57–86. doi:10.1016/0004-3702(91)90080-4 [page 65].

Ballard, Dana H., Hayhoe, Mary M., Pook, Polly K. and Rao, Rajesh P. N. (1997). Deictic codes for the embodiment of cognition. In: *Behavioral and Brain Sciences* 20.04. doi:10.1017/s0140525x97001611 [page 62].

Ballaz, Santiago J. (2009). Differential novelty detection in rats selectively bred for novelty-seeking behavior. In: *Neuroscience Letters* 461.1, 45–48. doi:10.1016/j.neulet.2009.05.066 [page 109].

Baram, Yoram, El-Yaniv, Ran and Luz, Kobi (2004). Online Choice of Active Learning Algorithms. In: *J. Mach. Learn. Res.* 5, 255–291. issn: 1532-4435 [pages 163, 164].

Baranes, Adrien F., Oudeyer, Pierre-Yves and Gottlieb, Jacqueline (2014). The effects of task difficulty, novelty and the size of the search space on intrinsically motivated exploration. In: *Frontiers in Neuroscience* 8. doi:10.3389/fnins.2014.00317 [pages 79, 106].

Baranes, Adrien and Oudeyer, Pierre-Yves (2009). R-IAC: Robust Intrinsically Motivated Exploration and Active Learning. In: *IEEE Trans. Auton. Mental Dev.* 1.3, 155–169. doi:10.1109/tamd.2009.2037513 [pages 111, 134].

Baranes, Adrien and Oudeyer, Pierre-Yves (2010). Intrinsically motivated goal exploration for active motor learning in robots: A case study. In: *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE. doi:10.1109/iros.2010.5651385 [pages 103, 111, 134, 138, 163].

Baranes, Adrien and Oudeyer, Pierre-Yves (2013). Active learning of inverse models with intrinsically motivated goal exploration in robots. In: *Robotics and Autonomous Systems* 61.1, 49–73. issn: 0921-8890. doi:10.1016/j.robot.2012.05.008 [pages 131, 154, 198].

Barbot, Baptiste and Lubart, Todd (2012). Creative thinking in music: Its nature and assessment through musical exploratory behaviors. In: *Psychology of Aesthetics, Creativity, and the Arts* 6.3, 231 [page 120].

Barto, Andrew G. (2012). Intrinsic Motivation and Reinforcement Learning. In: *Intrinsically Motivated Learning in Natural and Artificial Systems*. Springer Science + Business Media, 17–47. doi:10.1007/978-3-642-32375-1_2 [page 111].

Barto, Andrew G, Singh, Satinder and Chentanez, Nuttapong (2004). Intrinsically motivated learning of hierarchical collections of skills. In: *Proc. of ICDL 2004*, 112–119 [page 110].

Barto, Andrew, Mirolli, Marco and Baldassarre, Gianluca (2013). Novelty or Surprise? In: *Frontiers in Psychology* 4. doi:10.3389/fpsyg.2013.00907 [page 114].

Barton, R. R. (1998). Simulation metamodels. In: *1998 Winter Simulation Conference. Proceedings (Cat. No.98CH36274)*. Institute of Electrical & Electronics Engineers (IEEE). doi:10.1109/wsc.1998.744912 [page 225].

Basawapatna, Ashok R., Repenning, Alexander, Koh, Kyu Han and Nickerson, Hilarie (2013). The zones of proximal flow. In: *Proceedings of the ninth annual international ACM conference on International computing education research - ICER '13*. Association for Computing Machinery (ACM). doi:10.1145/2493394.2493404 [page 106].

Beer, Randall D. (2003). The Dynamics of Active Categorical Perception in an Evolved Model Agent. In: *Adaptive Behavior* 11.4, 209–243. doi:10.1177/1059712303114001 [page 62].

Beer, Randall D, Chiel, Hillel J, Quinn, Roger D and Ritzmann, Roy E (1998). Biorobotic approaches to the study of motor systems. In: *Current Opinion in Neurobiology* 8.6, 777–782. doi:10.1016/s0959-4388(98)80121-9 [page 53].

Belsky, Jay and Most, Robert K. (1981). From exploration to play: A cross-sectional study of infant free play behavior. In: *Developmental Psychology* 17.5, 630–639. doi:10.1037/0012-1649.17.5.630 [page 120].

Bénard, Henri (1901). Les tourbillons cellulaires dans une nappe liquide. - Méthodes optiques d'observation et d'enregistrement. In: *Journal de Physique Théorique et Appliquée* 10.1, 254–266. doi:10.1051/jphystap:0190100100025400 [page 97].

Bennett, James and Lanning, Stan (2007). The Netflix prize. In: *Proceedings of KDD cup and workshop*. Vol. 2007, 35 [page 118].

Bentlley, J. L. (1977). Algorithms for Klee's rectangle problem. In: *Unpublished manuscript* [page 240].

Berglund, Michael and Wieser, Michael E. (2011). Isotopic compositions of the elements 2009 (IUPAC Technical Report). In: *Pure and Applied Chemistry* 83.2. doi:10.1351/pac-rep-10-06-02 [page 98].

Berlyne, D. E. (1950). Novelty and Curiosity as Determinants of Exploratory Behavior. In: *British Journal of Psychology. General Section* 41.1-2, 68–80. doi:10.1111/j.2044-8295.1950.tb00262.x [page 104].

Berlyne, D. E. (1960). *Conflict, arousal, and curiosity.* McGraw-Hill Book Company. doi:10.1037/11164-000 [page 104].

Berlyne, D. E. (1966). Curiosity and Exploration. In: *Science* 153.3731, 25–33. doi:10.1126/science.153.3731.25 [page 104].

Bernshteĭn, Nikolaĭ Aleksandrovich (1967). *The Co-ordination and Regulation of Movements.* (Collection of papers from 1934 to 1962 translated from Russian and German). New York: Pergamon Press [page 145].

Bernstein, Daniel S. (1999). *Reusing old policies to accelerate learning on new MDPs.* Tech. rep. UM-CS-1999-026. Department of Computer Science, University of Massachusetts at Amherst [page 188].

Berthouze, Luc and Lungarella, Max (2004). Motor Skill Acquisition Under Environmental Perturbations: On the Necessity of Alternate Freezing and Freeing of Degrees of Freedom. In: *Adaptive Behavior* 12.1, 47–64. doi:10.1177/105971230401200104 [page 147].

Bertschinger, Nils, Olbrich, Eckehard, Ay, Nihat and Jost, JĂźrgen (2008). Autonomy: An information theoretic perspective. In: *Biosystems* 91.2, 331–345. doi:10.1016/j.biosystems.2007.05.018 [page 55].

Bhattacharyya, A. (1943). On A Measure of Divergence Between Two Statistical Populations Defined by their Probability Distributions. In: *Bulletin of Cal. Math. Soc.* 35.1 [page 185].

Billard, Aude, Calinon, Sylvain, Dillmann, RĂźdiger and Schaal, Stefan (2008). Robot Programming by Demonstration. In: *Springer Handbook of Robotics*. Springer Berlin Heidelberg, 1371–1394. doi:10.1007/978-3-540-30301-5_60 [page 151].

Billing, David (2007). Teaching for transfer of core/key skills in higher education: Cognitive skills. In: *High Educ* 53.4, 483–516. doi:10.1007/s10734-005-5628-5 [page 185].

Blackwell, T. M. (2005). Particle swarms and population diversity. In: *Soft Comput* 9.11, 793–802. doi:10.1007/s00500-004-0420-5 [page 118].

Blitzer, John, McDonald, Ryan and Pereira, Fernando (2006). Domain adaptation with structural correspondence learning. In: *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing - EMNLP '06*. Association for Computational Linguistics (ACL). doi:10.3115/1610075.1610094 [page 188].

Blumberg, Mark S., Marques, Hugo Gravato and Iida, Fumiya (2013). Twitching in Sensorimotor Development from Sleeping Rats to Robots. In: *Current Biology* 23.12, R532–R537. doi:10.1016/j.cub.2013.04.075 [pages 52, 102].

Bolado-Gomez, Rufino and Gurney, Kevin (2013). A biologically plausible embodied model of action discovery. In: *Frontiers in Neurorobotics* 7.4. issn: 1662-5218. doi:10.3389/fnbot.2013.00004 [page 114].

Bonawitz, Elizabeth Baraff, Schijndel, Tessa J. P. van, Friel, Daniel and Schulz, Laura (2012). Children balance theories and evidence in exploration, explanation, and learning. In: *Cognitive Psychology* 64.4, 215–234. doi:10.1016/j.cogpsych.2011.12.002 [page 96].

Bongard, Josh (2011). Morphological change in machines accelerates the evolution of robust behavior. In: *Proceedings of the National Academy of Sciences* 108.4, 1234–1239. doi:10.1073/pnas.1015390108 [pages 28, 69, 148, 238].

Bongard, Josh and Hornby, Gregory S. (2010). Guarding against premature convergence while accelerating evolutionary search. In: *Proceedings of the 12th annual conference on Genetic and evolutionary computation - GECCO '10*. Association for Computing Machinery (ACM). doi:10.1145/1830483.1830504 [page 116].

Bongard, Josh and Lipson, Hod (2005). Nonlinear System Identification Using Coevolution of Models and Tests. In: *IEEE Transactions on Evolutionary Computation* 9.4, 361–384. doi:10.1109/tevc.2005.850293 [pages 95, 210, 213, 225].

Bongard, Josh and Lipson, Hod (2014). Evolved Machines Shed Light on Robustness and Resilience. In: *Proc. IEEE* 102.5, 899–914. doi:10.1109/jproc.2014.2312844 [page 53].

Bongard, Josh, Zykov, Victor and Lipson, Hod (2006). Resilient Machines Through Continuous Self-Modeling. In: *Science* 314.5802, 1118–1121. doi:10.1126/science.1133687 [page 225].

Bonilla, Edwin, Chai, Kian Ming and Williams, Christopher (2008). Multi-task Gaussian process prediction. In: [page 188].

Bornstein, Marc H. (2014). Human Infancy...and the Rest of the Lifespan. In: *Annu. Rev. Psychol.* 65.1, 121–158. doi:10.1146/annurev-psych-120710-100359 [page 75].

Brafman, Ronen I. and Tennenholtz, Moshe (2003). R-MAX — a General Polynomial Time Algorithm for Near-optimal Reinforcement Learning. In: *J. Mach. Learn. Res.* 3, 213–231. issn: 1532-4435. doi:10.1162/153244303765208377 [page 115].

Braitenberg, Valentino (1986). *Vehicles : experiments in synthetic psychology*. Cambridge, Mass: MIT Press. isbn: 9780262521123 [pages 72, 74].

Braun, C. Christoph, Heinz, Udo, Schweizer, Renate, Wiech, Katja, Birbaumer, Niels and Topka, Helge (2001). Dynamic organization of the somatosensory cortex induced by motor activity. In: *Brain* 124.11, 2259–2267. issn: 0006-8950. doi:10.1093/brain/124.11.2259 [page 102].

Breiman, Leo (1996). In: *Machine Learning* 24.1, 41–47. doi:10.1023/a:1018094028462 [page 94].

Brooks, Rodney A. (1990). Elephants don't play chess. In: *Robotics and Autonomous Systems* 6.1-2, 3–15. doi:10.1016/s0921-8890(05)80025-9 [page 62].

Brooks, Rodney A. (1991a). Intelligence Without Reason. In: *Proceedings of the 12th International Joint Conference on Artificial Intelligence - Volume 1*. IJCAI'91. Sydney, New South Wales, Australia: Morgan Kaufmann Publishers Inc., 569–595. isbn: 1-55860-160-0 [page 62].

Brooks, Rodney A. (1991b). Intelligence Without Representation. In: *Artificial Intelligence* 47.1-3, 139–159. doi:10.1016/0004-3702(91)90053-m [pages 62, 152].

Brooks, Rodney A. (1991c). New Approaches to Robotics. In: *Science* 253.5025, 1227–1232. doi:10.1126/science.253.5025.1227 [page 62].

Brooks, Rodney A. (1999). *Cambrian Intelligence: The Early History of the New AI*. Cambridge, Mass: MIT Press. isbn: 9780262024686 [page 63].

Brown, E., Rodenberg, N., Amend, J., Mozeika, A., Steltz, E., Zakin, M. R., Lipson, H. and Jaeger, H. M. (2010). Universal robotic gripper based on the jamming of granular material. In: *Proceedings of the National Academy of Sciences* 107.44, 18809–18814. doi:10.1073/pnas.1003250107 [page 64].

Brown, Gavin, Wyatt, Jeremy, Harris, Rachel and Yao, Xin (2005). Diversity creation methods: a survey and categorisation. In: *Information Fusion* 6.1, 5–20. doi:10.1016/j.inffus.2004.04.004 [page 118].

Brown, Gillian R. and Nemes, Christopher (2008). The exploratory behaviour of rats in the hole-board apparatus: Is head-dipping a valid measure of neophilia? In: *Behavioural Processes* 78.3, 442–448. doi:10.1016/j.beproc.2008.02.019 [page 108].

Byers-Heinlein, Krista and Fennell, Christopher T. (2013). Perceptual narrowing in the context of increased variation: Insights from bilingual infants. In: *Dev Psychobiol* 56.2, 274–291. doi:10.1002/dev.21167 [page 71].

Byrd, Richard H., Lu, Peihuang, Nocedal, Jorge and Zhu, Ciyou (1995). A Limited Memory Algorithm for Bound Constrained Optimization. In: *SIAM J. Sci. Comput.* 16.5, 1190–1208. doi:10.1137/0916069 [pages 243, 244].

Byrge, Lisa, Sporns, Olaf and Smith, Linda B. (2014). Developmental process emerges from extended brain–body–behavior networks. In: *Trends in Cognitive Sciences* 18.8, 395–403. doi:10.1016/j.tics.2014.04.010 [page 67].

Calinon, Sylvain (2009). *Robot Programming by Demonstration : a Probabilistic Approach*. Lausanne, Switzerland Boca Raton, FL: EPFL Press Distributed by CRC Press. isbn: 9781439808672 [page 151].

Camazine, Scott (2003). *Self-organization in biological systems*. Princeton, N.J. Oxford: Princeton University Press. isbn: 9780691116242 [page 98].

Cannon, W.B. (1932). *The wisdom of the body*. W.W. Norton & Company, inc. [pages 99, 103].

Carter, Alecia J., Feeney, William E., Marshall, Harry H., Cowlishaw, Guy and Heinsohn, Robert (2012). Animal personality: what are behavioural ecologists measuring? In: *Biol Rev* 88.2, 465–475. doi:10.1111/brv.12007 [page 108].

Cazals, Frederic, Kanhere, Harshad and Loriot, Sébastien (2011). Computing the volume of a union of balls. In: *ACM Trans. Math. Softw.* 38.1, 1–20. doi:10.1145/2049662.2049665 [page 240].

Cen, Renyue (2014). Temporal Self-Organization in Galaxy Formation. In: *ApJ* 785.2, L21. doi:10.1088/2041-8205/785/2/l21 [pages 97, 98].

Ceriani, Lidia and Verme, Paolo (2012). The origins of the Gini index: extracts from Variabilità e Mutabilità (1912) by Corrado Gini. In: *J Econ Inequal* 10.3, 421–443. doi:10.1007/s10888-011-9188-x [page 88].

Cesa-Bianchi, Nicolo and Lugosi, Gabor (2006). *Prediction, Learning, and Games*. Cambridge University Press. doi:10.1017/cbo9780511546921 [page 163].

Chai, Kian M, Williams, Christopher, Klanke, Stefan and Vijayakumar, Sethu (2009). Multi-task gaussian process learning of robot inverse dynamics. In: *Advances in Neural Information Processing Systems*, 265–272 [page 188].

Chaloner, Kathryn and Verdinelli, Isabella (1995). Bayesian Experimental Design: A Review. In: *Statist. Sci.* 10.3, 273–304. doi:10.1214/ss/1177009939 [page 94].

Chan, T.M. (2013). Klee's Measure Problem Made Easy. In: *Foundations of Computer Science (FOCS), 2013 IEEE 54th Annual Symposium on*, 410–419. doi:10.1109/FOCS.2013.51 [pages 239, 240].

Chandola, Varun, Banerjee, Arindam and Kumar, Vipin (2007). Outlier detection: A survey. In: *ACM Computing Surveys* [page 115].

Chandola, Varun, Banerjee, Arindam and Kumar, Vipin (2009). Anomaly detection. In: *CSUR* 41.3, 1–58. doi:10.1145/1541880.1541882 [pages 49, 115].

Chapple, David G., Simmonds, Sarah M. and Wong, Bob B. M. (2011). Know when to run, know when to hide: can behavioral differences explain the divergent invasion success of two sympatric lizards? In: *Ecology and Evolution* 1.3, 278–289. doi:10.1002/ece3.22 [page 109].

Chapple, David G., Simmonds, Sarah M. and Wong, Bob B. M. (2012). Can behavioral and personality traits influence the success of unintentional species introductions? In: *Trends in Ecology & Evolution* 27.1, 57–64. doi:10.1016/j.tree.2011.09.010 [page 109].

Cheney, Nick, MacCurdy, Robert, Clune, Jeff and Lipson, Hod (2013). Unshackling Evolution: Evolving Soft Robots with Multiple Materials and a Powerful Generative Encoding. In: *Proceeding of the fifteenth annual conference on Genetic and evolutionary computation conference - GECCO '13*. Association for Computing Machinery (ACM). doi:10.1145/2463372.2463404 [page 69].

Cheng, Shi, Shi, Yuhui and Qin, Quande (2013). A Study of Normalized Population Diversity in Particle Swarm Optimization. In: *International Journal of Swarm Intelligence Research* 4.1, 1–34. doi:10.4018/jsir.2013010101 [page 118].

Churchland, Patricia S, Ramachandran, VS and Sejnowski, Terrence J (1994). A Critique of Pure Vision1. In: *Large-scale neuronal theories of the brain*, 23 [page 65].

Clark, Andy (1997). *Being there putting brain, body, and world together again*. Cambridge, Mass: MIT Press. isbn: 9780262531566 [page 62].

Clark, Fay E. and Smith, Lauren J. (2013). Effect of a Cognitive Challenge Device Containing Food and Non-Food Rewards on Chimpanzee Well-Being. In: *American Journal of Primatology* 75.8, 807–816. doi:10.1002/ajp.22141 [page 108].

Clement, Benjamin, Roy, Didier, Oudeyer, Pierre-Yves and Lopes, Manuel (2015). Multi-Armed Bandits for Intelligent Tutoring Systems. In: *(submitted)* [page 164].

Clements, Douglas H. and Gullo, Dominic F. (1984). Effects of computer programming on young children's cognition. In: *Journal of Educational Psychology* 76.6, 1051–1058. doi:10.1037/0022-0663.76.6.1051 [page 185].

Cleveland, William S. and Devlin, Susan J. (1988). Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting. In: *Journal of the American Statistical Association* 83.403, 596–610. doi:10.1080/01621459.1988.10478639 [pages 124, 243].

Cliff, D., Husbands, P. and Harvey, I. (1993). Explorations in Evolutionary Robotics. In: *Adaptive Behavior* 2.1, 73–110. doi:10.1177/105971239300200104 [page 69].

Cohen, Jerome S. and Stettner, Laurence J. (1968). Effect of deprivation level on exploratory behavior in the albino rat. In: *Journal of Comparative and Physiological Psychology* 66.2, 514–517. doi:10.1037/h0026350 [page 105].

Cohn, David A., Atlas, Les and Ladner, Richard (1994). Improving generalization with active learning. In: *Machine Learning* 15.2, 201–221. doi:10.1007/bf00993277 [pages 94, 95].

Cohn, David A., Ghahramani, Zoubin and Jordan, Michael I. (1996). Active Learning with Statistical Models. In: *Journal of Aritficial Intelligence Research* 4, 129–145 [pages 94, 95].

Conkur, E. Sahin and Buckingham, Rob (1997). Clarifying the definition of redundancy as used in robotics. In: *Robotica* 15.5, 583–586. doi:10.1017/s0263574797000672 [page 37].

Connolly, J-F, Granger, Eric and Sabourin, Robert (2012). On the correlation between genotype and classifier diversity. In: *Pattern Recognition (ICPR), 2012 21st International Conference on*. IEEE, 1068–1071 [page 118].

Cook, Claire, Goodman, Noah D. and Schulz, Laura E. (2011). Where science starts: Spontaneous experiments in preschoolers' exploratory play. In: *Cognition* 120.3, 341–349. doi:10.1016/j.cognition.2011.03.003 [pages 72, 95, 96, 199].

Cote, J., Fogarty, S., Weinersmith, K., Brodin, T. and Sih, A. (2010). Personality traits and dispersal tendency in the invasive mosquitofish (Gambusia affinis). In: *Proceedings of the Royal Society B: Biological Sciences* 277.1687, 1571–1579. doi:10.1098/rspb.2009.2128 [page 109].

Cousins, Steven H. (1991). Species diversity measurement: Choosing the right index. In: *Trends in Ecology & Evolution* 6.6, 190–192. doi:10.1016/0169-5347(91)90212-g [page 87].

Cruse, Holk (1990). What mechanisms coordinate leg movement in walking arthropods? In: *Trends in Neurosciences* 13.1, 15–21. doi:10.1016/0166-2236(90)90057-h [page 61].

Csikszentmihalyi, M., Abuhamdeh, S. and Nakamura, J. (2005). Flow. In: *Handbook of Competence and Motivation*. Ed. by Andrew J. Elliot et Carol S. Dweck. New York, NY, US: The Guilford Press. Chap. XVI, 598–608 [page 105].

Csikszentmihalyi, Mihaly (1990). *Flow : the psychology of optimal experience*. New York: Harper & Row. isbn: 0060920432 [page 105].

Cully, Antoine, Clune, Jeff and Mouret, Jean-Baptiste (2014). Robots that can adapt like natural animals. In: *arXiv preprint arXiv:1407.3501* [page 210].

Currey, J. D. (1979). Changes in the impact energy absorption of bone with age. In: *Journal of Biomechanics* 12.6, 459–469. doi:10.1016/0021-9290(79)90031-9 [page 211].

Daerden, Frank and Lefeber, Dirk (2002). Pneumatic Artificial Muscles: Actuators for Robotics and Automation. In: *European Journal of Mechanical and Environmental Engineering* 47.1, 11–21 [page 49].

Dai, Wenyuan, Xue, Gui-Rong, Yang, Qiang and Yu, Yong (2007). Co-clustering based classification for out-of-domain documents. In: *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '07*. Association for Computing Machinery (ACM). doi:10.1145/1281192.1281218 [page 188].

Dai, Wenyuan, Yang, Qiang, Xue, Gui-Rong and Yu, Yong (2007). Boosting for transfer learning. In: *Proceedings of the 24th international conference on Machine learning - ICML '07*. Association for Computing Machinery (ACM). doi:10.1145/1273496.1273521 [page 187].

Darchen, Roger (1952). Sur l'activité exploratrice de Blattella germanica. In: *Zeitschrift fĂźr Tierpsychologie* 9.3, 362–372. doi:10.1111/j.1439-0310.1952.tb01655.x [page 108].

Dasgupta, Sanjoy (2011). Two faces of active learning. In: *Theoretical Computer Science* 412.19, 1767–1781. doi:10.1016/j.tcs.2010.12.054 [page 93].

Daumé III, Hal (2007). Frustratingly Easy Domain Adaptation. In: *CoRR* abs/0907.1815 [page 188].

Davies, T. Jonathan and Cadotte, Marc W. (2011). Quantifying Biodiversity: Does It Matter What We Measure? In: *Biodiversity Hotspots*. Springer Science + Business Media, 43–60. doi:10.1007/978-3-642-20992-5_3 [page 87].

Davis, Abe, Rubinstein, Michael, Wadhwa, Neal, Mysore, Gautham J., Durand, Frédo and Freeman, William T. (2014). The visual microphone. In: *ACM Trans. Graph.* 33.4, 1–10. doi:10.1145/2601097.2601119 [page 57].

De Moivre, Abraham (1733). Approximatio ad summam terminorum binomii a+b)n in seriem expansi. In: *(self-published pamphlet)*. English translation: A. De Moivre, *The Doctrine of Chances*, 2nd ed. (London, England: H. Woodfall, 1738), pp. 235-243. [page 98].

deCharms, R. (1968). *Personal causation*. New York: Academic Press [pages 105, 107].

Deci, Edward L. (1975). *Intrinsic Motivation*. Springer US. doi:10.1007/978-1-4613-4446-9 [page 105].

Deci, Edward L. and Ryan, Richard M. (1985). *Intrinsic Motivation and Self-Determination in Human Behavior*. Springer US. doi:10.1007/978-1-4899-2271-7 [page 105].

Delarboulas, Pierre, Schoenauer, Marc and Sebag, Michèle (2010). Open-Ended Evolutionary Robotics: An Information Theoretic Approach. In: *Parallel Problem Solving from Nature, PPSN XI*. Springer Science + Business Media, 334–343. doi:10.1007/978-3-642-15844-5_34 [pages 69, 76, 87, 117, 121].

Dember, William N (1965). The new look in motivation. In: *American Scientist* 53, 409–427. issn: 00030996 [pages 104, 105].

Dember, William N. and Earl, Robert W. (1957). Analysis of exploratory, manipulatory, and curiosity behaviors. In: *Psychological Review* 64.2, 91–96. doi:10.1037/h0046861 [page 105].

Dempster, M.B.L (1998). A Self-Organising Systems Perspective on Planning for Sustainability. MA thesis. University of Waterloo, School of Urban and Regional Planning [page 96].

Deneubourg, J. L. and Goss, S. (1989). Collective patterns and decision-making. In: *Ethology Ecology & Evolution* 1.4, 295–311. doi:10.1080/08927014.1989.9525500 [page 97].

Der, Ralf (2014). On the Role of Embodiment for Self-Organizing Robots: Behavior As Broken Symmetry. In: *Emergence, Complexity and Computation*. Springer Science + Business Media, 193–221. doi:10.1007/978-3-642-53734-9_7 [page 97].

Der, Ralf and Martius, Georg (2012). *The Playful Machine*. Springer Berlin Heidelberg. doi:10.1007/978-3-642-20253-7 [pages 97, 99].

Der, Ralf and Martius, Georg (2013). Behavior as broken symmetry in embodied self-organizing robots. In: *Advances in Artificial Life, ECAL 2013*. MIT Press. doi:10.7551/978-0-262-31709-2-ch086 [page 97].

Dickman, J. David, Beyer, Matt and Hess, Bernhard J. M. (2000). Three-dimensional organization of vestibular related eye movements to rotational motion in pigeons. In: *Vision Research* 40.20, 2831–2844. doi:10.1016/s0042-6989(00)00128-0 [page 65].

Dingemanse, N (2002). Repeatability and heritability of exploratory behaviour in great tits from the wild. In: *Animal Behaviour* 64.6, 929–938. doi:10.1006/anbe.2002.2006 [page 109].

Dominguez, Melissa and Jacobs, Robert A. (2003). Developmental Constraints Aid the Acquisition of Binocular Disparity Sensitivities. In: *Neural Computation* 15.1, 161–182. doi:10.1162/089976603321043748 [page 148].

Doncieux, Stephane, Bredeche, Nicolas, Mouret, Jean-Baptiste and Eiben, Agoston E. (Gusz) (2015). Evolutionary Robotics: What, Why, and Where to. In: *Frontiers in Robotics and AI* 2. doi:10.3389/frobt.2015.00004 [pages 69, 116].

Doncieux, Stephane and Mouret, Jean-Baptiste (2010). Behavioral diversity measures for Evolutionary Robotics. In: *IEEE Congress on Evolutionary Computation*. Institute of Electrical & Electronics Engineers (IEEE). doi:10.1109/cec.2010.5586100 [pages 86, 87, 117, 210].

Doncieux, Stephane and Mouret, Jean-Baptiste (2014). Beyond black-box optimization: a review of selective pressures for evolutionary robotics. In: *Evolutionary Intelligence* 7.2, 71–93. doi:10.1007/s12065-014-0110-x [page 117].

Doncieux, Stéphane, Mouret, Jean-Baptiste, Bredeche, Nicolas and Padois, Vincent (2011). Evolutionary Robotics: Exploring New Horizons. In: *Studies in Computational Intelligence*. Springer Science + Business Media, 3–25. doi:10.1007/978-3-642-18272-3_1 [page 53].

Dorigo, Marco and Colombetti, Marco (1994). Robot shaping: developing autonomous agents through learning. In: *Artificial Intelligence* 71.2, 321–370. doi:10.1016/0004-3702(94)90047-7 [pages 116, 220].

Drumwright, Evan and Shell, Dylan A. (2010). Modeling Contact Friction and Joint Friction in Dynamic Robotic Simulation Using the Principle of Maximum Dissipation. In: *Springer Tracts in Advanced Robotics*. Springer Science + Business Media, 249–266. doi:10.1007/978-3-642-17452-0_15 [page 212].

Duda, Richard, Hart, Peter E. and Stork, David G. (2001). *Pattern Classification*. New York: Wiley. isbn: 9780471056690 [page 87].

Duncan, I. J. (1998). Behavior and behavioral needs. In: *Poultry Science* 77.12, 1766–1772. doi:10.1093/ps/77.12.1766 [page 107].

Düzel, Emrah, Bunzeck, Nico, Guitart-Masip, Marc and DĂźzel, Sandra (2010). NOvelty-related Motivation of Anticipation and exploration by Dopamine (NOMAD): Implications for healthy aging. In: *Neuroscience & Biobehavioral Reviews* 34.5, 660–669. doi:10.1016/j.neubiorev.2009.08.006 [page 120].

Edwards, Brian J., Rottman, Benjamin M., Shankar, Maya, Betzler, Riana, Chituc, Vladimir, Rodriguez, Ricardo, Silva, Liara, Wibecan, Leah, Widness, Jane and Santos, Laurie R. (2014). Do Capuchin Monkeys (Cebus apella) Diagnose Causal Relations in the Absence of a Direct Reward? In: *PLoS ONE* 9.2. Ed. by Emma Flynn, e88595. doi:10.1371/journal.pone.0088595 [page 108].

Eggenberger Hotz, Peter (2003). Genome-physics interaction as a new concept to reduce the number of genetic parameters in artificial evolution. In: *The 2003 Congress on Evolutionary Computation, 2003. CEC '03*. Institute of Electrical & Electronics Engineers (IEEE). doi:10.1109/cec.2003.1299574 [page 98].

Elman, Jeffrey L. (1993). Learning and development in neural networks: the importance of starting small. In: *Cognition* 48.1, 71–99. doi:10.1016/0010-0277(93)90058-4 [page 147].

Erez, Tom, Tassa, Yuval and Todorov, Emanuel. Simulation tools for model-based robotics: Comparison of Bullet, Havok, MuJoCo, ODE and PhysX. In: [page 212].

Espenschied, K. S., Chiel, H. J., Quinn, R. D. and Beer, R. D. (1993). Leg Coordination Mechanisms in the Stick Insect Applied to Hexapod Robot Locomotion. In: *Adaptive Behavior* 1.4, 455–468. doi:10.1177/105971239300100404 [page 61].

Fan, Wei, Davidson, I., Zadrozny, B. and Yu, Philip S. (2005). An Improved Categorization of Classifier's Sensitivity on Sample Selection Bias. In: *Fifth IEEE International Conference on Data Mining (ICDM'05)*. Institute of Electrical & Electronics Engineers (IEEE). doi:10.1109/icdm.2005.24 [page 187].

Farroni, T., Csibra, G., Simion, F. and Johnson, M. H. (2002). Eye contact detection in humans from birth. In: *Proceedings of the National Academy of Sciences* 99.14, 9602–9605. doi:10.1073/pnas.152159999 [page 102].

Fedorov, V. V. (1972). *Theory of optimal experiments*. New York: Academic Press. isbn: 9780122507502 [page 94].

Feige, Uriel (1998). A Threshold of Ln N for Approximating Set Cover. In: *J. ACM* 45.4, 634–652. issn: 0004-5411. doi:10.1145/285055.285059 [page 164].

Fernández, Fernando and Veloso, Manuela (2006a). Policy Reuse for Transfer Learning Across Tasks with Different State and Action Space. In: *ICML workshop on Structural Knowledge Transfer for Machine Learning* [page 189].

Fernández, Fernando and Veloso, Manuela (2006b). Probabilistic Policy Reuse in a Reinforcement Learning Agent. In: *Proc. of AAMAS 2006*. AAMAS '06. Hakodate, Japan: ACM, 720–727. isbn: 1-59593-303-4. doi:10.1145/1160633.1160762 [pages 186, 189].

Festinger, Leon (1957). *A Theory of Cognitive Dissonance*. Stanford, Calif: Stanford University Press. isbn: 9780804709118 [page 104].

Festinger, Leon and Carlsmith, James M. (1959). Cognitive consequences of forced compliance. In: *The Journal of Abnormal and Social Psychology* 58.2, 203–210. doi:10.1037/h0041593 [page 105].

Floreano, Dario and Keller, Laurent (2010). Evolution of Adaptive Behaviour in Robots by Means of Darwinian Selection. In: *PLoS Biol* 8.1, e1000292. doi:10.1371/journal.pbio.1000292 [page 69].

Floreano, Dario and Mondada, Francesco (1994). Automatic Creation of an Autonomous Agent: Genetic Evolution of a Neural-network Driven Robot. In: *Proceedings of the Third International Conference on Simulation of Adaptive Behavior : From Animals to Animats 3: From Animals to Animats 3*. SAB94. Brighton, United Kingdom: MIT Press, 421–430. isbn: 0-262-53122-4 [page 224].

Fowler, H. (1965). *Curiosity and Exploratory Behavior*. The Critical issues in psychology series. Macmillan [page 104].

Fredman, Michael L. and Weide, Bruce (1978). On the complexity of computing the measure of $\bigcup[a_i, b_i]$. In: *Commun. ACM* 21.7, 540–544. doi:10.1145/359545.359553 [page 240].

French, Robert M., Mermillod, Martial and Chauvin, Alan (2002). The importance of starting blurry: simulating improved basic-level category learning in infants due to weak visual acuity. In: *In Proceedings of the 24th Annual Conference of the Cognitive Science Society*, 322–327 [page 148].

Freund, Yoav, Seung, H. Sebastian, Shamir, Eli and Tishby, Naftali (1997). In: *Machine Learning* 28.2/3, 133–168. doi:10.1023/a:1007330508534 [pages 94, 95].

Friedrich, Tobias, Oliveto, Pietro S., Sudholt, Dirk and Witt, Carsten (2008). Theoretical analysis of diversity mechanisms for global exploration. In: *Proceedings of the 10th annual conference on Genetic and evolutionary computation - GECCO '08*. Association for Computing Machinery (ACM). doi:10.1145/1389095.1389276 [page 117].

Friston, Karl J., Daunizeau, Jean, Kilner, James and Kiebel, Stefan J. (2010). Action and behavior: a free-energy formulation. In: *Biol Cybern* 102.3, 227–260. doi:10.1007/s00422-010-0364-z [page 111].

Fritzke, Bernd (1995). A Growing Neural Gas Network Learns Topologies. In: *Advances in Neural Information Processing Systems 7*. Ed. by G. Tesauro, D. S. Touretzky and T. K. Leen. MIT Press, 625–632. isbn: 978-0262201049 [page 111].

Gallese, Vittorio, Fadiga, Luciano, Fogassi, Leonardo and Rizzolatti, Giacomo (1996). Action recognition in the premotor cortex. In: *Brain* 119.2, 593–609. doi:10.1093/brain/119.2.593 [page 63].

Gao, Jing, Fan, Wei, Jiang, Jing and Han, Jiawei (2008). Knowledge transfer via multiple model local structure mapping. In: *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 08*. Association for Computing Machinery (ACM). doi:10.1145/1401890.1401928 [page 188].

Garivier, Aurélien and Moulines, Eric (2008). On Upper-Confidence Bound Policies for Non-Stationary Bandit Problems. 24 pages [page 163].

Gerken, LouAnn, Balcomb, Frances K. and Minton, Juliet L. (2011). Infants avoid 'labouring in vain' by attending more to learnable than unlearnable linguistic patterns. In: *Developmental Science* 14.5, 972–979. doi:10.1111/j.1467-7687.2011.01046.x [page 106].

Gerken, Louann, Wilson, Rachel and Lewis, William (2005). Infants can use distributional cues to form syntactic categories. In: *J. Child Lang.* 32.2, 249–268. doi:10.1017/s0305000904006786 [page 106].

Gershman, Samuel J. and Blei, David M. (2012). A tutorial on Bayesian nonparametric models. In: *Journal of Mathematical Psychology* 56.1, 1–12. doi:10.1016/j.jmp.2011.08.004 [page 87].

Gibson, Eleanor J. (1988). Exploratory Behavior in the Development of Perceiving, Acting, and the Acquiring of Knowledge. In: *Annu. Rev. Psychol.* 39.1, 1–42. doi:10.1146/annurev.ps.39.020188.000245 [pages 71, 72].

Gibson, Eleanor J., Riccio, Gary, Schmuckler, Mark A., Stoffregen, Thomas A., Rosenberg, David and Taormina, Joanne (1987). Detection of the traversability of surfaces by crawling and walking infants. In: *Journal of Experimental Psychology: Human Perception and Performance* 13.4, 533–544. doi:10.1037/0096-1523.13.4.533 [page 72].

Gibson, James Jerome (1977). The Theory of Affordances. In: *Perceiving, acting, and knowing : toward an ecological psychology*. Ed. by Robert Shaw and J. Bransford. Hillsdale, N.J. New York: Lawrence Erlbaum Associates Distributed by the Halsted Press Division, Wiley. isbn: 0470990147 [pages 63, 72, 199].

Gini, Corrado (1912). *Variabilità e Mutuabilità. Contributo allo Studio delle Distribuzioni e delle Relazioni Statistiche.* Bologna: C. Cuppini [page 88].

Goda, Yukiko and Davis, Graeme W (2003). Mechanisms of Synapse Assembly and Disassembly. In: *Neuron* 40.2, 243–264. doi:10.1016/s0896-6273(03)00608-1 [page 100].

Goldberg, D. E. (1987). Simple genetic algorithms and the minimal, deceptive problem. In: *Genetic algorithms and simulated annealing*. Morgan Kaufmann, pp. 74–88 [pages 116, 117].

Gomez, Faustino J. (2009). Sustaining Diversity Using Behavioral Information Distance. In: *Proceedings of the 11th Annual Conference on Genetic and Evolutionary Computation*. GECCO '09. Montreal, Québec, Canada: ACM, 113–120. isbn: 978-1-60558-325-9. doi:10.1145/1569901.1569918 [page 117].

Gomez, Faustino and Miikkulainen, R. (1997). Incremental Evolution of Complex General Behavior. In: *Adaptive Behavior* 5.3-4, 317–342. doi:10.1177/105971239700500305 [pages 116, 220].

Gongora, Mario A., Passow, Benjamin N. and Hopgood, Adrian A. (2009). Robustness analysis of evolutionary controller tuning using real systems. In: *2009 IEEE Congress on Evolutionary Computation*. Institute of Electrical & Electronics Engineers (IEEE). doi:10.1109/cec.2009.4983001 [pages 224, 225].

Gopnik, A. (2012). Scientific Thinking in Young Children: Theoretical Advances, Empirical Research, and Policy Implications. In: *Science* 337.6102, 1623–1627. doi:10.1126/science.1223416 [pages 72, 95].

Gopnik, Alison (1997). *Words, thoughts, and theories*. Cambridge, Mass: MIT Press. isbn: 9780262571265 [pages 72, 95].

Gopnik, Alison, Sobel, David M., Schulz, Laura E. and Glymour, Clark (2001). Causal learning mechanisms in very young children: Two-, three-, and four-year-olds infer causal relations from patterns of variation and covariation. In: *Developmental Psychology* 37.5, 620–629. doi:10.1037/0012-1649.37.5.620 [page 95].

Gosling, Samuel D. and John, Oliver P. (1999). Personality Dimensions in Nonhuman Animals: A Cross-Species Review. In: *Current Directions in Psychological Science* 8.3, 69–75. doi:10.1111/1467-8721.00017 [page 109].

Gottlieb, Jacqueline, Oudeyer, Pierre-Yves, Lopes, Manuel and Baranes, Adrien (2013). Information-seeking, curiosity, and attention: computational and neural mechanisms. In: *Trends in Cognitive Sciences* 17.11, 585–593. doi:10.1016/j.tics.2013.09.001 [pages 71, 73, 75, 110, 120].

Granmo, M., Petersson, P. and Schouenborg, J. (2008). Action-Based Body Maps in the Spinal Cord Emerge from a Transitory Floating Organization. In: *Journal of Neuroscience* 28.21, 5494–5503. doi:10.1523/jneurosci.0651-08.2008 [page 102].

Gupta, A. K., Smith, K. G. and Shalley, C. E. (2006). The Interplay Between Exploration and Exploitation. In: *Academy of Management Journal* 49.4, 693–706. doi:10.5465/amj.2006.22083026 [page 120].

Gweon, Hyowon and Schulz, L. (2008). Stretching to learn: Ambiguous evidence and variability in preschoolers' exploratory play. In: [pages 72, 95, 96].

Hadjitodorov, Stefan T., Kuncheva, Ludmila I. and Todorova, Ludmila P. (2006). Moderate diversity for better cluster ensembles. In: *Information Fusion* 7.3, 264–275. doi:10.1016/j.inffus.2005.01.008 [page 118].

Haith, Marshall (1980). *Rules that babies look by : the organization of newborn visual activity*. Hillsdale, N.J: L. Erlbaum Associates. isbn: 9780898590333 [page 102].

Harlow, Harry F., Harlow, Margaret Kuenne and Meyer, Donald R. (1950). Learning Motivated by a Manipulation Drive. In: *Journal of Experimental Psychology* 40.2, 228–234. doi:10.1037/h0056906 [page 103].

Harnad, Stevan (1990). The symbol grounding problem. In: *Physica D: Nonlinear Phenomena* 42.1-3, 335–346. doi:10.1016/0167-2789(90)90087-6 [page 68].

Harvey, Inman, Husbands, Philip and Cliff, Dave (1992). *Issues in evolutionary robotics*. School of Cognitive and Computing Sciences, University of Sussex [page 70].

Hasenjäger, M. and Ritter, H. (2002). Active Learning in Neural Networks. In: *Studies in Fuzziness and Soft Computing*. Springer Science + Business Media, 137–169. doi:10.1007/978-3-7908-1803-1_5 [page 93].

Hausdorff, Felix (1914). *Grundzüge der Mengenlehre*. Leipzig: republished in 1949 by Chelsea Pub. Co. isbn: 978-0-8284-0061-9 [page 81].

Hebb, Donald Olding (1949). *The Organization of Behavior: A Neuropsychological Theory*. Wiley [page 46].

Heisenberg, W. (1927). Über den anschaulichen Inhalt der quantentheoretischen Kinematik und Mechanik. In: *Z. Physik* 43.3-4, 172–198. doi:10.1007/bf01397280 [page 57].

Helbing, Dirk, Buzna, Lubos, Johansson, Anders and Werner, Torsten (2005). Self-Organized Pedestrian Crowd Dynamics: Experiments, Simulations, and Design Solutions. In: *Transportation Science* 39.1, 1–24. doi:10.1287/trsc.1040.0108 [page 97].

Hendriks-Jansen, Horst (1996). *Catching ourselves in the act situated activity, interactive emergence, evolution, and human thought*. Cambridge, Mass: MIT Press. isbn: 9780262082464 [page 62].

Hervouet, Fabien and Bourreau, Eric (2012). Improvement proposals to intrinsically motivational robotics. In: *Proc. ICDL-EpiRob 2012*. IEEE. doi:10.1109/devlrn.2012.6400806 [page 111].

Hervouet, Fabien and Bourreau, Eric (2013). FIMO: Framework for Intrinsic Motivation. In: *Advances in Artificial Life, ECAL 2013*. MIT Press. doi:10.7551/978-0-262-31709-2-ch148 [pages 89, 103, 131].

Hester, Todd, Lopes, Manuel and Stone, Peter (2013). Learning Exploration Strategies in Model-based Reinforcement Learning. In: *Proceedings of the 2013 International Conference on Autonomous Agents* *and Multi-agent Systems*. AAMAS '13. St. Paul, MN, USA: International Foundation for Autonomous Agents and Multiagent Systems, 1069–1076. isbn: 978-1-4503-1993-5 [page 164].

Hilgard, Ernest R. and McGraw, Myrtle B. (1945). The Neuromuscular Maturation of the Human Infant. In: *The American Journal of Psychology* 58.2, 296. doi:10.2307/1417865 [pages 74, 101].

Hirai, K., Hirose, M., Haikawa, Y. and Takenaka, T. (1998). The development of Honda humanoid robot. In: *Proceedings. 1998 IEEE International Conference on Robotics and Automation (Cat. No.98CH36146)*. Institute of Electrical & Electronics Engineers (IEEE). doi:10.1109/robot.1998.677288 [page 49].

Hodge, Victoria and Austin, Jim (2004). A Survey of Outlier Detection Methodologies. In: *Artificial Intelligence Review* 22.2, 85–126. doi:10.1023/b:aire.0000045502.10941.a9 [page 115].

Hofsten, Claes von (1982). Eye-hand coordination in the newborn. In: *Developmental Psychology* 18.3, 450–461. doi:10.1037/0012-1649.18.3.450 [page 102].

Hofsten, Claes von (2004). An action perspective on motor development. In: *Trends in Cognitive Sciences* 8.6, 266–272. doi:10.1016/j.tics.2004.04.002 [pages 71, 100, 102, 103].

Holland, John H. (1975). *Adaptation in natural and artificial systems. an introductory analysis with applications to biology, control and artificial intelligence*. University of Michigan Press [page 69].

Holland, John H. (1992). *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*. Cambridge, Mass: MIT Press. isbn: 0262581116 [page 117].

Holst, Erich von (1939). Die relative Koordination als Phänomen und als Methode zentralnervöser Funktionsanalyse. In: *Ergebnisse der Physiologie* 42, 228–306 [page 145].

Holway, David A and Suarez, Andrew V (1999). Animal behavior: an essential component of invasion biology. In: *Trends in Ecology & Evolution* 14.8, 328–330. doi:10.1016/s0169-5347(99)01636-5 [page 109].

Huang, Jiayuan, Gretton, Arthur, Borgwardt, Karsten M, Schölkopf, Bernhard and Smola, Alex J (2006). Correcting sample selection bias by unlabeled data. In: *Advances in neural information processing systems*, 601–608 [page 187].

Huang, Xiao and Weng, Juyang (2002). *Novelty and Reinforcement Learning in the Value System of Developmental Robots*. Ed. by Christopher G. Prince, Yiannis Demiris, Yuval Marom, Hideki Kozima and Christian Balkenius [page 114].

Huang, Xiao and Weng, Juyang (2004). Motivational System for Human-Robot Interaction. In: *Computer Vision in Human-Computer Interaction*. Springer Science + Business Media, 17–27. doi:10.1007/978-3-540-24837-8_3 [page 114].

Hughes, Robert N. (1965). Food deprivation and locomotor exploration in the white rat. In: *Animal Behaviour* 13.1, 30–32. doi:10.1016/0003-3472(65)90068-0 [page 107].

Hughes, Robert N (1997). Intrinsic exploration in animals: motives and measurement. In: *Behavioural Processes* 41.3, 213–226. doi:10.1016/s0376-6357(97)00055-7 [pages 108, 121].

Hull, C.L. (1943). *Principles of Behavior: An Introduction to Behavior Theory*. The Century psychology series. D. Appleton-Century Company, Incorporated [page 103].

Humphrey, Tryphena (1944). Primitive neurons in the embryonic human central nervous system. In: *Journal of Comparative Neurology* 81.1, 1–45 [page 100].

Hunt, J. M. (1965). Intrinsic motivation and its role in psychological development. In: vol. 13, 189–282 [page 105].

Hurlbert, Stuart H. (1971). The Nonconcept of Species Diversity: A Critique and Alternative Parameters. In: *Ecology* 52.4, 577. doi:10.2307/1934145 [page 87].

Hutchins, Edwin (1995). *Cognition in the Wild*. Cambridge, Mass: MIT Press. isbn: 9780262581462 [page 62].

Ijspeert, A. J., Nakanishi, J. and Schaal, S. (2002). Movement imitation with nonlinear dynamical systems in humanoid robots. In: *Proceedings 2002 IEEE International Conference on Robotics and Automation*

(Cat. No.02CH37292). Institute of Electrical & Electronics Engineers (IEEE). doi:10.1109/robot.2002.1014739 [page 202].

Ijspeert, Auke Jan (2008). Central pattern generators for locomotion control in animals and robots: A review. In: *Neural Networks* 21.4, 642–653. doi:10.1016/j.neunet.2008.03.014 [pages 52, 53, 212].

Ijspeert, Auke Jan, Crespi, Alessandro and Cabelguen, Jean-Marie (2005). Simulation and Robotics Studies of Salamander Locomotion: Applying Neurobiological Principles to the Control of Locomotion in Robots. In: *Neuroinformatics* 3.3, 171–196. doi:10.1385/ni:3:3:171 [pages 52, 53].

Ijspeert, Auke Jan, Nakanishi, Jun, Hoffmann, Heiko, Pastor, Peter and Schaal, Stefan (2013). Dynamical Movement Primitives: Learning Attractor Models for Motor Behaviors. In: *Neural Computation* 25.2, 328–373. doi:10.1162/neco_a_00393 [page 203].

Ito, Masao (1972). Neural design of the cerebellar motor control system. In: *Brain Research* 40.1, 81–84. doi:10.1016/0006-8993(72)90110-2 [page 54].

Jakobi, N. (1997). Evolutionary Robotics and the Radical Envelope-of-Noise Hypothesis. In: *Adaptive Behavior* 6.2, 325–368. doi:10.1177/105971239700600205 [pages 224, 225].

Jakobi, Nick (1998). Running across the reality gap: Octopod locomotion evolved in a minimal simulation. In: *Evolutionary Robotics*. Springer Science + Business Media, 39–58. doi:10.1007/3-540-64957-3_63 [page 224].

Jakobi, Nick, Husbands, Phil and Harvey, Inman (1995). Noise and the reality gap: The use of simulation in evolutionary robotics. In: *Advances in Artificial Life*. Springer Science + Business Media, 704–720. doi:10.1007/3-540-59496-5_337 [pages 212, 224].

Jamone, Lorenzo, Natale, Lorenzo, Hashimoto, Kenji, Sandini, Giulio and Takanishi, Atsuo (2011). Learning task space control through goal directed exploration. In: *2011 IEEE International Conference on Robotics and Biomimetics*. Institute of Electrical & Electronics Engineers (IEEE). doi:10.1109/robio.2011.6181368 [pages 103, 131].

Jauffret, Adrien, Cuperlier, Nicolas, Tarroux, Philippe and Gaussier, Philippe (2013). From self-assessment to frustration, a small step toward autonomy in robotic navigation. In: *Front. Neurorobot.* 7. doi:10.3389/fnbot.2013.00016 [page 164].

Jepma, Marieke, Verdonschot, Rinus G., Steenbergen, Henk van, Rombouts, Serge A. R. B. and Nieuwenhuis, Sander (2012). Neural mechanisms underlying the induction and relief of perceptual curiosity. In: *Frontiers in Behavioral Neuroscience* 6. doi:10.3389/fnbeh.2012.00005 [page 120].

Jiang, Jing and Zhai, Chengxiang (2007). Instance weighting for domain adaptation in NLP. In: *In ACL 2007*, 264–271 [page 187].

Jin, Yaochu and Meng, Yan (2011). Morphogenetic Robotics: An Emerging New Field in Developmental Robotics. In: *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on* 41.2, 145–160. doi:10.1109/TSMCC.2010.2057424 [page 69].

Johnson, Scott H (2000). Thinking ahead: the case for motor imagery in prospective judgements of prehension. In: *Cognition* 74.1, 33–70. doi:10.1016/s0010-0277(99)00063-3 [page 54].

Jones, Donald R., Schonlau, Matthias and Welch, William J. (1998). Efficient Global Optimization of Expensive Black-Box Functions. In: *Journal of Global Optimization* 13.4, 455–492. doi:10.1023/a:1008306431147 [pages 94, 225].

Jouzel, Jean and Merlivat, Liliane (1984). Deuterium and oxygen 18 in precipitation: Modeling of the isotopic effects during snow formation. In: *J. Geophys. Res.* 89.D7, 11749. doi:10.1029/jd089id07p11749 [page 98].

Kagan, Jerome (1972). Motives and development. In: *Journal of Personality and Social Psychology* 22.1, 51–66. doi:10.1037/h0032356 [pages 104, 105].

Kahrs, Björn Alexander (2012). Rhythmical stereotypies in infancy. PhD thesis. Tulane University School of Science and Engineering. isbn: 9781267656087 [pages 74, 101].

Kaneko, Kenji, Kanehiro, Fumio, Morisawa, Mitsuharu, Miura, Kanako, Nakaoka, Shin'ichiro and Kajita, Shuuji (2009). Cybernetic human HRP-4C. In: *2009 9th IEEE-RAS International Conference on Humanoid Robots*. Institute of Electrical & Electronics Engineers (IEEE). doi:10.1109/ichr.2009.5379537 [page 49].

Kang, Min Jeong, Hsu, Ming, Krajbich, Ian M., Loewenstein, George, McClure, Samuel M., Wang, Joseph Tao-yi and Camerer, Colin F. (2009). The Wick in the Candle of Learning: Epistemic Curiosity Activates Reward Circuitry and Enhances Memory. In: *Psychological Science* 20.8, 963–973. doi:10.1111/j.1467-9280.2009.02402.x [pages 108, 120].

Kaufman, Danny M., Sueda, Shinjiro, James, Doug L. and Pai, Dinesh K. (2008). Staggered projections for frictional contact in multibody systems. In: *ACM Trans. Graph.* 27.5, 1. doi:10.1145/1409060.1409117 [page 212].

Kawato, Mitsuo (1999). Internal models for motor control and trajectory planning. In: *Current Opinion in Neurobiology* 9.6, 718–727. doi:10.1016/s0959-4388(99)00028-8 [page 54].

Keil, Frank C. (2003). Folkscience: coarse interpretations of a complex reality. In: *Trends in Cognitive Sciences* 7.8, 368–373. doi:10.1016/s1364-6613(03)00158-x [page 71].

Kelso, J. A. Scott (1995). *Dynamic Patterns: The Self-Organization of Brain and Behavior (Complex Adaptive Systems)*. The MIT Press. isbn: 9780262611312 [page 99].

Kennedy, J. and Eberhart, R. (1995). Particle swarm optimization. In: *Proceedings of ICNN'95 - International Conference on Neural Networks*. Institute of Electrical & Electronics Engineers (IEEE). doi:10.1109/icnn.1995.488968 [page 118].

Kidd, Celeste, Piantadosi, Steven T. and Aslin, Richard N. (2012). The Goldilocks Effect: Human Infants Allocate Attention to Visual Sequences That Are Neither Too Simple Nor Too Complex. In: *PLoS ONE* 7.5. Ed. by Antoni Rodriguez-Fornells, e36399. doi:10.1371/journal.pone.0036399 [pages 71, 106].

Kidd, Celeste, Piantadosi, Steven T. and Aslin, Richard N. (2014). The Goldilocks Effect in Infant Auditory Attention. In: *Child Dev*, n/a–n/a. doi:10.1111/cdev.12263 [pages 71, 106].

Kim, Deok-Soo, Ryu, Joonghyun, Shin, Hayong and Cho, Youngsong (2012). Beta-decomposition for the volume and area of the union of three-dimensional balls and their offsets. In: *J. Comput. Chem.* 33.13, 1252–1273. doi:10.1002/jcc.22956 [page 240].

Klee, Victor (1977). Can the Measure of $\bigcup_1^n [a_i, b_i]$ be Computed in Less Than $\mathcal{O}(n\log n)$ Steps? In: *The American Mathematical Monthly* 84.4, 284. doi:10.2307/2318871 [page 239].

Kleibeuker, Sietske W., Dreu, Carsten K. W. De and Crone, Eveline A. (2012). The development of creative cognition across adolescence: distinct trajectories for insight and divergent thinking. In: *Developmental Science* 16.1, 2–12. doi:10.1111/j.1467-7687.2012.01176.x [page 120].

Klyubin, Alexander S., Polani, Daniel and Nehaniv, Chrystopher L. (2005a). All Else Being Equal Be Empowered. In: *Advances in Artificial Life*. Springer Science + Business Media, 744–753. doi:10.1007/11553090_75 [page 111].

Klyubin, Alexander S., Polani, Daniel and Nehaniv, Chrystopher L. (2005b). Empowerment: A Universal Agent-Centric Measure of Control. In: *2005 IEEE Congress on Evolutionary Computation*. IEEE. doi:10.1109/cec.2005.1554676 [page 111].

Klyubin, Alexander S., Polani, Daniel and Nehaniv, Chrystopher L. (2008). Keep Your Options Open: An Information-Based Driving Principle for Sensorimotor Systems. In: *PLoS ONE* 3.12. Ed. by Olaf Sporns, e4018. doi:10.1371/journal.pone.0004018 [page 111].

Kochenderfer, Mykel J. and Gupta, Rakesh (2003). Common Sense Data Acquisition for Indoor Mobile Robots. In: *In Nineteenth National Conference on Artificial Intelligence (AAAI-04*. AAAI Press / The MIT Press, 605–610 [page 67].

Kodjabachian, J. and Meyer, J. -A. (1998). Evolution and development of neural controllers for locomotion, gradient-following, and obstacle-avoidance in artificial insects. In: *IEEE Trans. Neural Netw.* 9.5, 796–812. doi:10.1109/72.712153 [pages 116, 220].

Konczak, Jürgen (2005). On the notion of motor primitives in humans and robots. In: *Proc. of Epirob 2005* 123. Ed. by Luc Berthouze, Frédéric Kaplan, Hideki Kozima, Hiroyuki Yano, Jürgen Konczak, Giorgio Metta, Jacqueline Nadel, Giulio Sandini, Georgi Stojanov and Christian Balkenius, 47–53 [page 203].

Konidaris, George and Barto, Andrew (2008). Sensorimotor abstraction selection for efficient, autonomous robot skill acquisition. In: *2008 7th IEEE International Conference on Development and Learning*. Institute of Electrical & Electronics Engineers (IEEE). doi:10.1109/devlrn.2008.4640821 [page 163].

Kononenko, Igor (2001). Machine learning for medical diagnosis: history, state of the art and perspective. In: *Artificial Intelligence in Medicine* 23.1, 89–109. doi:10.1016/s0933-3657(01)00077-x [page 47].

Koos, Sylvain, Mouret, J-B and Doncieux, S. (2013). The Transferability Approach: Crossing the Reality Gap in Evolutionary Robotics. In: *IEEE Transactions on Evolutionary Computation* 17.1, 122–145. doi:10.1109/tevc.2012.2185849 [page 225].

Koos, Sylvain, Mouret, Jean-Baptiste and Doncieux, Stephane (2009). Automatic system identification based on coevolution of models and tests. In: *2009 IEEE Congress on Evolutionary Computation*. Institute of Electrical & Electronics Engineers (IEEE). doi:10.1109/cec.2009.4982995 [page 225].

Krause, Andreas and Golovin, Daniel (2014). Submodular Function Maximization. In: *Practical Approaches to Hard Problems*. Ed. by Lucas Bordeaux, Youssef Hamadi, Pushmeet Kohli and Robert Mateescu. Cambridge University Press, 71–104. doi:10.1017/cbo9781139177801.004 [page 164].

Krause, Andreas and Guestrin, Carlos (2005). Near-optimal Value of Information in Graphical Models. In: *Conference on Uncertainty in Artificial Intelligence (UAI)* [page 164].

Krawczyk, Bartosz and Wozniak, Michal (2013). Accuracy and diversity in classifier selection for one-class classification ensembles. In: *2013 IEEE Symposium on Computational Intelligence and Ensemble Learning (CIEL)*. Institute of Electrical & Electronics Engineers (IEEE). doi:10.1109/ciel.2013.6613139 [page 118].

Krawczyk, Bartosz and Woźniak, Michał (2014). Diversity measures for one-class classifier ensembles. In: *Neurocomputing* 126, 36–44. doi:10.1016/j.neucom.2013.01.053 [page 118].

Krcah, Peter (2010). Solving deceptive tasks in robot body-brain co-evolution by searching for behavioral novelty. In: *2010 10th International Conference on Intelligent Systems Design and Applications*. Institute of Electrical & Electronics Engineers (IEEE). doi:10.1109/isda.2010.5687250 [page 117].

Krig, Daniel G. (1951). A Statistical Approach to Some Basic Mine Valuation Problems on the Witwatersrand. In: *Journal of Chemical, Metallurgical, and Mining Society of South Africa* 52 (6), 119–139 [page 95].

Krink, T., Vesterstrom, J. S. and Riget, J. (2002). Particle swarm optimisation with spatial particle extension. In: *Proceedings of the 2002 Congress on Evolutionary Computation. CEC'02 (Cat. No.02TH8600)*. Institute of Electrical & Electronics Engineers (IEEE). doi:10.1109/cec.2002.1004460 [page 118].

Kuhl, Patricia K (2000). Language, mind, and brain: Experience alters perception. In: *The new cognitive neurosciences* 2, 99–115 [page 68].

Kulvicius, T., Ning, K., Tamosiunaite, M. and Worgötter, F. (2012). Joining Movement Sequences: Modified Dynamic Movement Primitives for Robotics Applications Exemplified on Handwriting. In: *Robotics, IEEE Transactions on* 28.1, 145–157. issn: 1552-3098, doi:10.1109/TRO.2011.2163863 [page 203].

Kumaran, Dharshan and Maguire, Eleanor A. (2007). Which computational mechanisms operate in the hippocampus during novelty detection? In: *Hippocampus* 17.9, 735–748. doi:10.1002/hipo.20326 [page 114].

Kuncheva, L. I. (2001). Ten measures of diversity in classifier ensembles: limits for two classifiers. In: *DERA/IEE Workshop Intelligent Sensor Processing*. Institution of Engineering and Technology (IET). doi:10.1049/ic:20010105 [page 118].

Kuncheva, Ludmila I. and Whitaker, Christopher J. (2003). In: *Machine Learning* 51.2, 181–207. doi:10.1023/a:1022859003006 [page 118].

Lakoff, George (1999). *Philosophy in the flesh : the embodied mind and its challenge to Western thought*. New York: Basic Books. isbn: 9780465056743 [page 62].

Lande, Russell (1996). Statistics and Partitioning of Species Diversity, and Similarity among Multiple Communities. In: *Oikos* 76.1, 5. doi:10.2307/3545743 [page 87].

Lapeyre, Matthieu, Ly, Olivier and Oudeyer, Pierre-Yves (2011). Maturational constraints for motor learning in high-dimensions: The case of biped walking. In: *2011 11th IEEE-RAS International Conference on Humanoid Robots*. IEEE. doi:10.1109/humanoids.2011.6100909 [page 211].

Lapeyre, Matthieu, Rouanet, Pierre and Oudeyer, Pierre-Yves (2013). Poppy: a New Bio-Inspired Humanoid Robot Platform for Biped Locomotion and Physical Human-Robot Interaction. In: *Proceedings of the 6th International Symposium on Adaptive Motion in Animals and Machines (AMAM)* [page 211].

Larranaga, P. (2006). Machine learning in bioinformatics. In: *Briefings in Bioinformatics* 7.1, 86–112. doi:10.1093/bib/bbk007 [page 47].

Laucht, Manfred, Becker, Katja and Schmidt, Martin H. (2006). Visual exploratory behaviour in infancy and novelty seeking in adolescence: two developmentally specific phenotypes of DRD4? In: *Journal of Child Psychology and Psychiatry* 47.11, 1143–1151. doi:10.1111/j.1469-7610.2006.01627.x [page 120].

Lawrence, Neil D. and Platt, John C. (2004). Learning to learn with the informative vector machine. In: *Twenty-first international conference on Machine learning - ICML '04*. Association for Computing Machinery (ACM). doi:10.1145/1015330.1015382 [page 188].

Lazaric, Alessandro (2012). Transfer in Reinforcement Learning: A Framework and a Survey. In: *Reinforcement Learning*. Springer Science + Business Media, 143–173. doi:10.1007/978-3-642-27645-3_5 [page 189].

Lee, M. H. and Meng, Q. (2005). Staged development of Robot Motor Coordination. In: *2005 IEEE International Conference on Systems, Man and Cybernetics*. IEEE. doi:10.1109/icsmc.2005.1571593 [pages 114, 148].

Lee, M. H., Meng, Q. and Chao, F. (2007a). Developmental learning for autonomous robots. In: *Robotics and Autonomous Systems* 55.9, 750–759. doi:10.1016/j.robot.2007.05.002 [page 148].

Lee, M. H., Meng, Q. and Chao, F. (2007b). Staged Competence Learning in Developmental Robotics. In: *Adaptive Behavior* 15.3, 241–255. doi:10.1177/1059712307082085 [page 148].

Lee, Mark H. (2011). Intrinsic Activitity: from motor babbling to play. In: *2011 IEEE International Conference on Development and Learning (ICDL)*. Institute of Electrical & Electronics Engineers (IEEE). doi:10.1109/devlrn.2011.6037375 [pages 102, 103].

Lee, Rachel, Walker, Ryan, Meeden, Lisa and Marshall, James (2009). Category-based Intrinsic Motivation. In: *Proc. Epirob 2009*. Vol. 146, 81–88 [pages 111, 134].

Lee, Tae-Hee and Crompton, John (1992). Measuring novelty seeking in tourism. In: *Annals of Tourism Research* 19.4, 732–751. doi:10.1016/0160-7383(92)90064-v [page 120].

Lehman, Joel, Clune, Jeff and Risi, Sebastian (2014). An Anarchy of Methods: Current Trends in How Intelligence Is Abstracted in AI. In: *IEEE Intelligent Systems* 29.6, 56–62. doi:10.1109/mis.2014.92 [pages 26, 236].

Lehman, Joel and Stanley, Kenneth O. (2008). Exploiting Open-Endedness to Solve Problems Through the Search for Novelty. In: *Proc. of the Eleventh Intl. Conf. on Artificial Life (ALIFE XI)*. Cambridge, MA: MIT Press [pages 85, 117, 121].

Lehman, Joel and Stanley, Kenneth O. (2011a). Abandoning Objectives: Evolution Through the Search for Novelty Alone. In: *Evolutionary Computation* 19.2, 189–223. doi:10.1162/evco_a_00025 [pages 117, 221].

Lehman, Joel and Stanley, Kenneth O. (2011b). Evolving a diversity of virtual creatures through novelty search and local competition. In: *Proceedings of the 13th annual conference on Genetic and evolutionary computation - GECCO '11*. Association for Computing Machinery (ACM). doi:10.1145/2001576.2001606 [page 69].

Lehrer, Richard, Guckenberg, Thomas and Sancilio, Leonard (1988). Influences of LOGO on children's intellectual development. In: *Teaching and learning computer programming: Multiple research perspectives.* Ed. by Richard E. (Ed) Mayer. Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc, pp. 75–110 [page 185].

Lemaignan, Séverin, Ros, R, MoĔ⍰senlechner, Lorenz, Alami, R and Beetz, M (2010). ORO, a knowledge management platform for cognitive architectures in robotics. In: *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*. Institute of Electrical & Electronics Engineers (IEEE). doi:10.1109/iros.2010.5649547 [pages 59, 68].

Lenarcic, J. (1999). On the quantification of robot redundancy. In: *Proc. ICRA 1999*. IEEE. doi:10.1109/robot.1999.774079 [page 37].

Levi, Paul and Kernbach, Serge (2010). *Symbiotic Multi-Robot Organisms*. Springer Berlin Heidelberg. doi:10.1007/978-3-642-11692-6 [page 207].

Lewis, David D. and Catlett, Jason (1994). Heterogeneous Uncertainty Sampling for Supervised Learning. In: *Machine Learning Proceedings 1994*. Elsevier, 148–156. doi:10.1016/b978-1-55860-335-6.50026-x [page 94].

Liang, Z. S., Nguyen, T., Mattila, H. R., Rodriguez-Zas, S. L., Seeley, T. D. and Robinson, G. E. (2012). Molecular Determinants of Scouting Behavior in Honey Bees. In: *Science* 335.6073, 1225–1228. doi:10.1126/science.1213962 [page 108].

Liao, Xuejun, Xue, Ya and Carin, Lawrence (2005). Logistic Regression with an Auxiliary Data Source. In: *Proc. ICML '05*. ICML '05. Bonn, Germany: ACM, 505–512. isbn: 1-59593-180-5. doi:10.1145/1102351.1102415 [page 187].

Liebl, A. L. and Martin, L. B. (2012). Exploratory behaviour and stressor hyper-responsiveness facilitate range expansion of an introduced songbird. In: *Proceedings of the Royal Society B: Biological Sciences* 279.1746, 4375–4381. doi:10.1098/rspb.2012.1606 [page 109].

Lindauer, Martin (1952). Ein Beitrag zur Frage der Arbeitsteilung im Bienenstaat. In: *Zeitschrift für Vergleichende Physiologie* 34.4, 299–345. doi:10.1007/bf00298048 [page 108].

Lipson, Hod (2005). Evolutionary Robotics and Open-Ended Design Automation. In: *Biomimetics*. CRC Press, 129–155. doi:10.1201/9781420037715.ch4 [pages 53, 69].

Lipson, Hod, Bongard, Josh, Zykov, Victor and Malone, Evan (2006). Evolutionary Robotics for Legged Machines: From Simulation to Physical Reality. In: *Proc. of the 9th Int. Conference on Intelligent Autonomous Systems.* 11–18 [page 224].

Lipson, Hod and Pollack, Jordan B. (2000). Automatic design and manufacture of robotic lifeforms. In: *Nature* 406.6799, 974–978. doi:10.1038/35023115 [page 69].

Lisman, John E. and Grace, Anthony A. (2005). The Hippocampal-VTA Loop: Controlling the Entry of Information into Long-Term Memory. In: *Neuron* 46.5, 703–713. doi:10.1016/j.neuron.2005.05.002 [page 114].

Loeb, Gerald E. (2012). Optimal isn't good enough. In: *Biol Cybern* 106.11-12, 757–765. doi:10.1007/s00422-012-0514-6 [pages 54, 56, 60, 61, 74, 125, 145].

Loewenstein, George (1994). The psychology of curiosity: A review and reinterpretation. In: *Psychological Bulletin* 116.1, 75–98. doi:10.1037/0033-2909.116.1.75 [page 105].

Lopes, Manuel, Lang, Tobias, Toussaint, Marc and Oudeyer, Pierre-Yves (2012). Exploration in Model-based Reinforcement Learning by Empirically Estimating Learning Progress. Anglais. In: *Neural Information Processing Systems (NIPS)*. Lake Tahoe, États-Unis [page 111].

Lopes, Manuel, Melo, Francisco, Montesano, Luis and Santos-Victor, José (2010). Abstraction Levels for Robotic Imitation: Overview and Computational Approaches. In: *Studies in Computational Intelligence*. Springer Science + Business Media, 313–355. doi:10.1007/978-3-642-05181-4_14 [page 151].

Lopes, Manuel and Montesano, Luis (2014). Active Learning for Autonomous Intelligent Agents: Exploration, Curiosity, and Interaction. In: *CoRR* abs/1403.1497 [page 93].

Lopes, Manuel and Oudeyer, Pierre-Yves (2010). Guest Editorial Active Learning and Intrinsically Motivated Exploration in Robots: Advances and Challenges. In: *IEEE Trans. Auton. Mental Dev.* 2.2, 65–69. doi:10.1109/tamd.2010.2052419 [page 93].

Lopes, Manuel and Oudeyer, Pierre-Yves (2012). The strategic student approach for life-long exploration and learning. In: *Proc. ICDL-Epirob 2012*. IEEE. doi:10.1109/devlrn.2012.6400807 [pages 163, 164, 198].

Loveland, K. (1986). Discovering the Affordances of a Reflecting Surface. In: *Developmental Review* 6.1, 1–24. doi:10.1016/0273-2297(86)90001-8 [page 72].

Lungarella, Max, Metta, Giorgio, Pfeifer, Rolf and Sandini, Giulio (2003). Developmental robotics: a survey. In: *Connection Science* 15.4, 151–190. doi:10.1080/09540090310001655110 [pages 53, 66, 68].

Mahdavi, Siavash Haroun and Bentley, Peter J. (2003). An Evolutionary Approach to Damage Recovery of Robot Motion with Muscles. In: *Advances in Artificial Life*. Springer Science + Business Media, 248–255. doi:10.1007/978-3-540-39432-7_27 [page 210].

Mántaras Badia, Ramon López de (2013). Computational creativity. In: *Arbor* 189.764, a082. doi:10.3989/arbor.2013.764n6005 [page 120].

March, James G. (1991). Exploration and Exploitation in Organizational Learning. In: *Organization Science* 2.1, 71–87. doi:10.1287/orsc.2.1.71 [page 120].

Markou, Markos and Singh, Sameer (2003a). Novelty Detection: A Review—Part 1: Statistical Approaches. In: *Signal Processing* 83.12, 2481–2497. doi:10.1016/j.sigpro.2003.07.018 [page 115].

Markou, Markos and Singh, Sameer (2003b). Novelty Detection: A Review—Part 2: Neural Network Based Approaches. In: *Signal Processing* 83.12, 2499–2521. doi:10.1016/j.sigpro.2003.07.019 [page 115].

Marques, Hugo Gravato, Imtiaz, Farhan, Iida, Fumiya and Pfeifer, Rolf (2012). Self-organization of reflexive behavior from spontaneous motor activity. In: *Biol Cybern* 107.1, 25–37. doi:10.1007/s00422-012-0521-7 [page 102].

Marques, Hugo Gravato, Völk, Kristin, König, Stefan and Iida, Fumiya (2012). Self-organization of Spinal Reflexes Involving Homonymous, Antagonist and Synergistic Interactions. In: *From Animals to Animats 12*. Springer Science + Business Media, 269–278. doi:10.1007/978-3-642-33093-3_27 [page 102].

Marshall, James B., Blank, Douglas and Meeden, Lisa (2004). An Emergent Framework for Self-Motivation in Developmental Robotics. In: *Proc. of ICDL 2004*. Salk Institute, 104–111 [page 114].

Marsland, Stephen, Shapiro, Jonathan and Nehmzow, Ulrich (2002). A self-organising network that grows when required. In: *Neural Networks* 15.8-9, 1041–1058. doi:10.1016/s0893-6080(02)00078-3 [page 115].

Martin, L. B. (2005). A taste for novelty in invading house sparrows, Passer domesticus. In: *Behavioral Ecology* 16.4, 702–707. doi:10.1093/beheco/ari044 [page 109].

Martinez-Cantin, Ruben, Lopes, Manuel and Montesano, Luis (2010). Body schema acquisition through active learning. In: *ICRA 2010*. IEEE. doi:10.1109/robot.2010.5509406 [page 95].

Martius, Georg, Der, Ralf and Ay, Nihat (2013). Information Driven Self-Organization of Complex Robotic Behaviors. In: *PLoS ONE* 8.5. Ed. by Josh Bongard, e63400. doi:10.1371/journal.pone.0063400 [pages 100, 111].

Mataric, Maja J (1994). Reward functions for accelerated learning. In: *Machine Learning: Proceedings of the Eleventh international conference*, 181–189 [pages 116, 220].

Matheron, Georges (1962). *Traité de géostatistique appliquée*. Editions Technip [page 95].

Maturana, H. R. and Varela, F. J. (1973). Autopoiesis: The Organization of the Living. In: *Autopoiesis and Cognition: The Realization of the Living (Maturana & Varela 1980)*. First published in 1972 in Chile under the title *De Maquinas y Seres Vivos*, Editorial Universitaria S.A., pp. 59–138 [pages 51, 55].

Maturana, Humberto (1987). Everything is said by an observer. In: *Gaia, a way of knowing : political implications of the new biology*. Ed. by William Thompson. Great Barrington, MA Rochester, VT: Lindisfarne Press Distributed by Inner Traditions International, pp. 65–82. isbn: 9780940262232 [page 97].

Maurice, M and Gioanni, H (2004). Eye–neck coupling during optokinetic responses in head-fixed pigeons (Columba livia): influence of the flying behaviour. In: *Neuroscience* 125.2, 521–531. doi:10.1016/j.neuroscience.2004.01.054 [page 65].

Mautner, Craig and Belew, Richard K. (2000). Evolving robot morphology and control. In: *Artif Life Robotics* 4.3, 130–136. doi:10.1007/bf02481333 [page 116].

McClelland, David C, Atkinson, John W, Clark, Russell A and Lowell, Edgar L (1953). The Achievement Motive. In: [page 104].

McCrea, D. A. and Rybak, I. A. (2008). Organization of mammalian locomotor rhythm and pattern generation. In: *Brain Res Rev* 57.1, 134–146 [page 145].

McGeer, Tad (1990). Passive Dynamic Walking. In: *The International Journal of Robotics Research* 9.2, 62–82. doi:10.1177/027836499000900206 [page 64].

McGeer, Tad (1992). Principles of walking and running. In: *Advances in comparative and environmental physiology* 11, 113–139 [page 64].

McGraw, Myrtle B. (1945). *The neuromuscular maturation of the human infant*. New York: Hafner [page 101].

Meagher, Rebecca K. and Mason, Georgia J. (2012). Environmental Enrichment Reduces Signs of Boredom in Caged Mink. In: *PLoS ONE* 7.11. Ed. by Nei Moreira, e49180. doi:10.1371/journal.pone.0049180 [page 108].

Meer, A. van der, Weel, F. van der and Lee, D. (1995). The functional significance of arm movements in neonates. In: *Science* 267.5198, 693–695. doi:10.1126/science.7839147 [page 102].

Meer, Audrey L van der (1997). Keeping the arm in the limelight: Advanced visual control of arm movements in neonates. In: *European Journal of Paediatric Neurology* 1.4, 103–108. doi:10.1016/s1090-3798(97)80040-2 [page 103].

Meng, Q. and Lee, M. H. (2005). Novelty and Habituation: The Driving Forces in Early Stage Learning for Developmental Robotics. In: *Biomimetic Neural Learning for Intelligent Robots*. Springer Science + Business Media, 315–332. doi:10.1007/11521082_19 [page 114].

Mennerick, Steven and Zorumski, CharlesF. (2000). Neural activity and survival in the developing nervous system. English. In: *Molecular Neurobiology* 22.1-3, 41–54. issn: 0893-7648. doi:10.1385/MN:22:1-3:041 [page 100].

Merrick, K. and Maher, M. L. (2009). Motivated Learning from Interesting Events: Adaptive, Multitask Learning Agents for Complex Environments. In: *Adaptive Behavior* 17.1, 7–27. doi:10.1177/1059712308100236 [page 76].

Merrick, Kathryn E. (2012). Novelty and Beyond: Towards Combined Motivation Models and Integrated Learning Architectures. In: *Intrinsically Motivated Learning in Natural and Artificial Systems*. Springer Science + Business Media, 209–233. doi:10.1007/978-3-642-32375-1_9 [page 110].

Meyer, Jean-Arcady, Husbands, Phil and Harvey, Inman (1998). Evolutionary robotics: A survey of applications and problems. In: *Evolutionary Robotics*. Springer Science + Business Media, 1–21. doi:10.1007/3-540-64957-3_61 [page 69].

Miall, R. C. and Wolpert, D. M. (1996). Forward Models for Physiological Motor Control. In: *Neural Networks* 9.8, 1265–1279. doi:10.1016/s0893-6080(96)00035-4 [page 54].

Milh, M., Kaminska, A., Huon, C., Lapillonne, A., Ben-Ari, Y. and Khazipov, R. (2006). Rapid Cortical Oscillations and Early Motor Activity in Premature Human Neonate. In: *Cerebral Cortex* 17.7, 1582–1594. doi:10.1093/cercor/bhl069 [page 102].

Mirtich, Brian and Canny, John (1995). Impulse-based Simulation of Rigid Bodies. In: *Proceedings of the 1995 Symposium on Interactive 3D Graphics*. I3D '95. Monterey, California, USA: ACM, 181–ff. isbn: 0-89791-736-7. doi:10.1145/199404.199436 [page 212].

Misslin, René and Cigrang, Marc (1986). Does neophobia necessarily imply fear or anxiety? In: *Behavioural Processes* 12.1, 45–50. issn: 0376-6357. doi:http://dx.doi.org/10.1016/0376-6357(86)90069-0 [page 107].

Mitchell, Tom Michael (2006). *The Discipline of Machine Learning*. Carnegie Mellon University, School of Computer Science, Machine Learning Department [page 45].

Moessinger, A C (1983). Fetal akinesia deformation sequence: an animal model. In: *Pediatrics* 72.6, 857–63 [page 101].

Montgomery, K. C. (1954). The role of the exploratory drive in learning. In: *Journal of Comparative and Physiological Psychology* 47.1, 60–64. doi:10.1037/h0054833 [page 104].

Montgomery, K. C. and Segall, Marshall (1955). Discrimination learning based upon the exploratory drive. In: *Journal of Comparative and Physiological Psychology* 48.3, 225–228. doi:10.1037/h0047087 [page 104].

Montgomery, Kay C. (1951a). "Spontaneous alternation" as a function of time between trials and amount of work. In: *Journal of Experimental Psychology* 42.2, 82–93. doi:10.1037/h0059834 [page 104].

Montgomery, Kay C. (1951b). The relation between exploratory behavior and spontaneous alternation in the white rat. In: *Journal of Comparative and Physiological Psychology* 44.6, 582–589. doi:10.1037/h0063576 [page 104].

Montgomery, Kay C. (1952a). A test of two explanations of spontaneous alternation. In: *Journal of Comparative and Physiological Psychology* 45.3, 287–293. doi:10.1037/h0058118 [page 104].

Montgomery, Kay C. (1952b). Exploratory behavior and its relation to spontaneous alternation in a series of maze exposures. In: *Journal of Comparative and Physiological Psychology* 45.1, 50–57. doi:10.1037/h0053570 [page 104].

Morales, José Luis and Nocedal, Jorge (2011). Remark on "algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound constrained optimization". In: *ACM Trans. Math. Softw.* 38.1, 1–4. doi:10.1145/2049662.2049669 [page 244].

Morgan, C. Lloyd (1894). *An introduction to comparative psychology*. American Psychological Association (APA). doi:10.1037/11344-000 [page 103].

Morrongiello, Barbara A., Walpole, Beverly and Lasenby, Jennifer (2007). Understanding children's injury-risk behavior: Wearing safety gear can lead to increased risk taking. In: *Accident Analysis & Prevention* 39.3, 618–623. doi:10.1016/j.aap.2006.10.006 [page 211].

Moulin-Frier, Clément, Nguyen, Sao M. and Oudeyer, Pierre-Yves (2014). Self-organization of early vocal development in infants and machines: the role of intrinsic motivation. In: *Frontiers in Psychology* 4. doi:10.3389/fpsyg.2013.01006 [pages 76, 99].

Moulin-Frier, Clément, Rouanet, Pierre and Oudeyer, Pierre-Yves (2014). Explauto: an open-source Python library to study autonomous exploration in developmental robotics. In: *ICDL-Epirob - International Conference on Development and Learning, Epirob*. Genoa, Italy [page 89].

Mouret, J. -B. and Doncieux, S. (2012). Encouraging Behavioral Diversity in Evolutionary Robotics: An Empirical Study. In: *Evolutionary Computation* 20.1, 91–133. doi:10.1162/evco_a_00048 [page 117].

Mouret, Jean-Baptiste (2011). Novelty-Based Multiobjectivization. In: *Studies in Computational Intelligence*. Springer Science + Business Media, 139–154. doi:10.1007/978-3-642-18272-3_10 [pages 117, 121].

Mouret, Jean-Baptiste and Doncieux, Stephane (2009a). Overcoming the bootstrap problem in evolutionary robotics using behavioral diversity. In: *2009 IEEE Congress on Evolutionary Computation*. Institute of Electrical & Electronics Engineers (IEEE). doi:10.1109/cec.2009.4983077 [pages 116, 117].

Mouret, Jean-Baptiste and Doncieux, Stéphane (2009b). Using behavioral exploration objectives to solve deceptive problems in neuro-evolution. In: *Proceedings of the 11th Annual conference on Genetic and evolutionary computation - GECCO '09*. Association for Computing Machinery (ACM). doi:10.1145/1569901.1569988 [page 117].

Munzer, T., Stulp, F. and Sigaud, O. (2014). Non-linear regression algorithms for motor skill acquisition: a comparison. In: *Proceedings JFPDA*, 1–16 [page 243].

Nagai, Yukie, Asada, Minoru and Hosoda, Koh (2006). Learning for joint attention helped by functional development. In: *Advanced Robotics* 20.10, 1165–1181. doi:10.1163/156855306778522497 [page 148].

Nehring, Klaus and Puppe, Clemens (2002). A Theory of Diversity. In: *Econometrica* 70.3, 1155–1198. doi:10.1111/1468-0262.00321 [pages 25, 236].

Nemhauser, G. L., Wolsey, L. A. and Fisher, M. L. (1978). An analysis of approximations for maximizing submodular set functions. In: *Mathematical Programming* 14.1, 265–294. doi:10.1007/bf01588971 [page 164].

Neto, Hugo Vieira and Nehmzow, Ulrich (2005a). Automated Exploration and Inspection: Comparing Two Visual Novelty Detectors. In: *Int J Adv Robotic Sy*, 1. doi:10.5772/5770 [page 115].

Neto, Hugo Vieira and Nehmzow, Ulrich (2005b). Incremental PCA: An alternative approach for novelty detection. In: *Towards Autonomous Robotic Systems* [page 115].

Neto, Hugo Vieira and Nehmzow, Ulrich (2007a). Real-time Automated Visual Inspection using Mobile Robots. In: *J Intell Robot Syst* 49.3, 293–307. doi:10.1007/s10846-007-9146-9 [page 115].

Neto, Hugo Vieira and Nehmzow, Ulrich (2007b). Visual novelty detection with automatic scale selection. In: *Robotics and Autonomous Systems* 55.9, 693–701. doi:10.1016/j.robot.2007.05.012 [page 115].

Nguyen, D. M. and Wong, K. P. (2003). Controlling diversity of evolutionary algorithms. In: *Proceedings of the 2003 International Conference on Machine Learning and Cybernetics (IEEE Cat. No.03EX693)*. Institute of Electrical & Electronics Engineers (IEEE). doi:10.1109/icmlc.2003.1259581 [page 117].

Nguyen, Sao Mai and Oudeyer, Pierre-Yves (2012). Active choice of teachers, learning strategies and goals for a socially guided intrinsic motivation learner. In: *Paladyn* 3.3, 136–146. doi:10.2478/s13230-013-0110-z [pages 163, 164].

Nissen, Henry W. (1930). A Study of Exploratory Behavior in the White Rat by Means of the Obstruction Method. In: *The Pedagogical Seminary and Journal of Genetic Psychology* 37.3, 361–376. doi:10.1080/08856559.1930.9944162 [page 103].

Nocera, Dario Di, Finzi, Alberto, Rossi, Silvia and Staffa, Mariacarla (2014). The role of intrinsic motivations in attention allocation and shifting. In: *Frontiers in Psychology* 5. doi:10.3389/fpsyg.2014.00273 [page 120].

Nolfi, Stefano (2000). *Evolutionary robotics : the biology, intelligence, and technology of self-organizing machines*. Cambridge, Mass: MIT Press. isbn: 9780262640565 [pages 53, 69].

Nolfi, Stefano, Floreano, Dario, Miglino, Orazio and Mondada, Francesco (1994). How to evolve autonomous robots: Different approaches in evolutionary robotics. In: *Artificial life IV: Proceedings of the 4th International Workshop on Artificial Life*. LIS-CONF-1994-002. MA: MIT Press, 190–197 [page 225].

Ollion, Charles and Doncieux, Stéphane (2011). Why and how to measure exploration in behavioral space. In: *Proceedings of the 13th annual conference on Genetic and evolutionary computation - GECCO '11*. Association for Computing Machinery (ACM). doi:10.1145/2001576.2001613 [page 85].

Olorunda, Olusegun and Engelbrecht, Andries P. (2008). Measuring exploration/exploitation in particle swarms using swarm diversity. In: *2008 IEEE Congress on Evolutionary Computation (IEEE World Con-*

gress on Computational Intelligence). Institute of Electrical & Electronics Engineers (IEEE). doi:10.1109/cec.2008.4630938 [pages 87, 118].

O'Regan, J. Kevin (1992). Solving the "real" mysteries of visual perception: The world as an outside memory. In: *Canadian Journal of Psychology/Revue canadienne de psychologie* 46.3, 461–488. doi:10.1037/h0084327 [page 65].

Orrell, David (2007). *The future of everything : the science of prediction : from wealth and weather to chaos and complexity*. New York: Basic Books. isbn: 9781568583693 [page 99].

Otmakhova, Nonna, Duzel, Emrah, Deutch, Ariel Y. and Lisman, John (2012). The Hippocampal-VTA Loop: The Role of Novelty and Motivation in Controlling the Entry of Information into Long-Term Memory. In: *Intrinsically Motivated Learning in Natural and Artificial Systems*. Springer Science + Business Media, 235–254. doi:10.1007/978-3-642-32375-1_10 [page 114].

Oudeyer, Pierre-Yves (2004). *Intelligent Adaptive Curiosity: a source of Self-Development*. Ed. by Luc Berthouze, Hideki Kozima, Christopher G. Prince, Giulio Sandini, Georgi Stojanov, Giorgio Metta and Christian Balkenius [pages 111, 134].

Oudeyer, Pierre-Yves (2006). *Self-Organization in the Evolution of Speech*. Vol. 6. Studies in the Evolution of Language. Oxford University Press, p. 177. isbn: 9780199289158. doi:10.1093/acprof:oso/9780199289158.001.0001 [page 99].

Oudeyer, Pierre-Yves (2013). *Aux sources de la parole : auto-organisation et évolution*. Paris: O. Jacob. isbn: 9782738129482 [page 99].

Oudeyer, Pierre-Yves, Baranes, Adrien and Kaplan, Frédéric (2013). Intrinsically Motivated Learning of Real World Sensorimotor Skills with Developmental Constraints. Anglais. In: *Intrinsically Motivated Learning in Natural and Artificial Systems*. Ed. by Gianluca Baldassarre and Marco Mirolli. Springer [pages 56, 57, 95].

Oudeyer, Pierre-Yves and Kaplan, F. (2007). What is Intrinsic Motivation? A Typology of Computational Approaches. In: *Frontiers in neurorobotics* 1, 6. issn: 1662-5218. doi:10.3389/neuro.12.006.2007 [page 131].

Oudeyer, Pierre-Yves, Kaplan, F. and Hafner, V.V. (2007). Intrinsic Motivation Systems for Autonomous Mental Development. In: *IEEE Transactions on Evolutionary Computation* 11.2, 265–286. issn: 1089-778X. doi:10.1109/TEVC.2006.890271 [pages 76, 111, 113, 134].

Oudeyer, Pierre-Yves and Kaplan, Frederic (2008). How can we define intrinsic motivation? In: *proceedings of the 8th international conference on epigenetic robotics: modeling cognitive development in robotic systems*. lund university cognitive studies [page 112].

Overveld, Thijs van, Careau, Vincent, Adriaensen, Frank and Matthysen, Erik (2013). Seasonal- and sex-specific correlations between dispersal and exploratory behaviour in the great tit. In: *Oecologia* 174.1, 109–120. doi:10.1007/s00442-013-2762-0 [page 109].

Oyama, Susan (2000). *The ontogeny of information developmental systems and evolution*. Durham: Duke University Press. isbn: 9780822324669 [pages 68, 99, 153].

Özöğür-Akyüz, Süreyya, Windeatt, Terry and Smith, Raymond (2014). Pruning of Error Correcting Output Codes by optimization of accuracy–diversity trade off. In: *Machine Learning*. doi:10.1007/s10994-014-5477-5 [page 118].

Oztop, Erhan, Kawato, Mitsuo and Arbib, Michael A. (2013). Mirror neurons: Functions, mechanisms and models. In: *Neuroscience Letters* 540, 43–55. doi:10.1016/j.neulet.2012.10.005 [pages 54, 63].

Page, Scott (2011). *Diversity and complexity*. Princeton, N.J: Princeton University Press. isbn: 9780691137674 [page 88].

Palmer, M, Miller, D and Blackwell, T (2009). An evolved neural controller for bipedal walking: Transitioning from simulator to hardware. In: *Proc. of IROS 2009 Workshop on Exploring new horizons in Evolutionary Design of Robots* [page 224].

Pan, Sinno Jialin, Kwok, James T. and Yang, Qiang (2008). Transfer Learning via Dimensionality Reduction. In: *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 2*. AAAI'08. Chicago, Illinois: AAAI Press, 677–682. isbn: 978-1-57735-368-3 [page 188].

Pan, Sinno Jialin and Yang, Qiang (2010). A Survey on Transfer Learning. In: *Knowledge and Data Engineering, IEEE Transactions on* 22.10, 1345–1359. issn: 1041-4347. doi:10.1109/TKDE.2009.191 [pages 183, 185, 187, 188].

Paolo, Ezequiel A. Di and Iizuka, Hiroyuki (2008). How (not) to model autonomous behaviour. In: *Biosystems* 91.2, 409–423. doi:10.1016/j.biosystems.2007.05.016 [page 52].

Partridge, L. D. (1982). The good enough calculi of evolving control systems: evolution is not engineering. In: *American Journal of Physiology - Regulatory, Integrative and Comparative Physiology* 242.3, R173–R177 [page 54].

Parzen, Emanuel (1962). On Estimation of a Probability Density Function and Mode. In: *Ann. Math. Statist.* 33.3, 1065–1076. doi:10.1214/aoms/1177704472 [page 131].

Pascal, Blaise (1662). *Pensées* [page 104].

Paul, Chandana (2004). Investigation of Morphology and Control in Biped Locomotion. PhD thesis [page 64].

Paul, Chandana (2006). Morphological computation. In: *Robotics and Autonomous Systems* 54.8, 619–630. doi:10.1016/j.robot.2006.03.003 [page 64].

Pavlov, I. P. (1904). О психической секреции слюнных желез. In: *Поли. собр. соч* 3. About the Psychic Secretion of the Salivary Glands, 40–57 [page 103].

Pavlov, I. P. (1927). *Conditioned Reflexes.* New York, NY: Oxford University Press [page 103].

Pea, Roy D. and Kurland, D. Midian (1984). On the cognitive effects of learning computer programming. In: *New Ideas in Psychology* 2.2, 137–168. doi:10.1016/0732-118x(84)90018-7 [page 185].

Pearson, Karl (1924). Historical Note on the Origin of the Normal Curve of Errors. In: *Biometrika* 16.3/4, 402. doi:10.2307/2331714 [page 98].

Peet, Robert K (1974). The measurement of species diversity. In: *Annual review of ecology and systematics*, 285–307 [page 87].

Pellegrino, G. di, Fadiga, L., Fogassi, L., Gallese, V. and Rizzolatti, G. (1992). Understanding motor events: a neurophysiological study. In: *Exp Brain Res* 91.1, 176–180. doi:10.1007/bf00230027 [page 63].

Penrose, R. and Todd, J. A. (1955). A generalized inverse for matrices. In: *Mathematical Proceedings of the Cambridge Philosophical Society* 51.03, 406. doi:10.1017/s0305004100030401 [page 244].

Perfiliev, S., Isa, T., Johnels, B., Steg, G. and Wessberg, J. (2010). Reflexive Limb Selection and Control of Reach Direction to Moving Targets in Cats, Monkeys, and Humans. In: *Journal of Neurophysiology* 104.5, 2423–2432. doi:10.1152/jn.01133.2009 [pages 65, 66, 145].

Perkins, Theodore J and Precup, Doina (1999). Using options for knowledge transfer in reinforcement learning. In: *University of Massachusetts, Amherst, MA, USA, Tech. Rep* [page 188].

Pfeifer, R., Lungarella, M. and Iida, F. (2007). Self-Organization, Embodiment, and Biologically Inspired Robotics. In: *Science* 318.5853, 1088–1093. doi:10.1126/science.1145803 [pages 53, 55].

Pfeifer, Rolf and Bongard, Josh C. (2006). *How the Body Shapes the Way We Think: A New View of Intelligence*. The MIT Press. isbn: 0262162393 [pages 63, 66, 70, 156].

Pfeifer, Rolf and Iida, Fumiya (2005). Morphological computation: Connecting body, brain and environment. In: *Japanese Scientific Monthly* [page 64].

Pfeifer, Rolf and Scheier, Christian (1999). *Understanding Intelligence*. Cambridge, Mass: MIT Press. isbn: 0262161818 [page 62].

Piaget, Jean and Cook, Peggy (1953). *The Origin of Intelligence in the Child.(Translated by Margaret Cook.)*. translated from 'La naissance de l'intelligence chez l'enfant', 1936. London: Routledge & Kegan Paul [pages 68, 71, 104].

Pianka, Eric R. (1966). Latitudinal Gradients in Species Diversity: A Review of Concepts. In: *The American Naturalist* 100.910, 33. doi:10.1086/282398 [page 87].

Pilipski, Mark and Diessner Pilpiski, Jacob (2006). A Treatise on the Preponderance of Designs Over Historic and Measured Snowfalls, or No Two Snowflakes Are Alike: Considerations About the Formation of Snowflakes and the Possible Numbers and Shapes of Snowflakes. In: *Proc. of the 63rd Eastern Snow Conference*, 283–289 [page 98].

Polani, Daniel (2008). Foundations and Formalizations of Self-organization. In: *Advanced Information and Knowledge Processing*. Springer Science + Business Media, 19–37. doi:10.1007/978-1-84628-982-8_2 [pages 97, 111].

Poole, Trevor B (1998). Meeting a mammal's psychological needs: Basic principles. In: *Second Nature: Environmental Enrichment for Captive Animals. Smithsonian Institution Press, Washington, DC*, pp. 83–94 [page 107].

Prechtl, Heinz F. R. (1985). Ultrasound studies of human fetal behaviour. In: *Early Human Development* 12.2, 91–98. doi:10.1016/0378-3782(85)90173-2 [page 100].

Prechtl, Heinz F. R. and Hopkins, Brian (1986). Developmental transformations of spontaneous movements in early infancy. In: *Early Human Development* 14.3-4, 233–238. doi:10.1016/0378-3782(86)90184-2 [page 100].

Precup, Doina (2000). Temporal Abstraction in Reinforcement Learning. AAI9978540. PhD thesis. University of Massachusetts Amherst. isbn: 0-599-84488-4 [page 188].

Pukelsheim, Friedrich (2006). *Optimal Design of Experiments*. Society for Industrial and Applied Mathematics. doi:10.1137/1.9780898719109 [page 94].

Purvis, Andy and Hector, Andy (2000). Getting the measure of biodiversity. In: *Nature* 405.6783, 212–219. doi:10.1038/35012221 [page 87].

Quiñonero-Candela, Joaquin, Sugiyama, Masashi, Schwaighofer, Anton and Lawrence, Neil D., eds. (2008). *Dataset Shift in Machine Learning*. MIT Press - Journals. doi:10.7551/mitpress/9780262170055.001.0001 [pages 184, 187, 188].

Raina, Rajat, Ng, Andrew Y. and Koller, Daphne (2006). Constructing informative priors using transfer learning. In: *Proceedings of the 23rd international conference on Machine learning - ICML '06*. Association for Computing Machinery (ACM). doi:10.1145/1143844.1143934 [page 188].

Ranganath, Charan and Rainer, Gregor (2003). Cognitive neuroscience: Neural mechanisms for detecting and remembering novel events. In: *Nature Reviews Neuroscience* 4.3, 193–202. doi:10.1038/nrn1052 [page 114].

Rankin, Catharine H., Abrams, Thomas, Barry, Robert J., Bhatnagar, Seema, Clayton, David F., Colombo, John, Coppola, Gianluca, Geyer, Mark A., Glanzman, David L., Marsland, Stephen, McSweeney, Frances K., Wilson, Donald A., Wu, Chun-Fang and Thompson, Richard F. (2009). Habituation revisited: An updated and revised description of the behavioral characteristics of habituation. In: *Neurobiology of Learning and Memory* 92.2, 135–138. doi:10.1016/j.nlm.2008.09.012 [page 114].

Raphael, G., Tsianos, G. A. and Loeb, G. E. (2010). Spinal-Like Regulator Facilitates Control of a Two-Degree-of-Freedom Wrist. In: *Journal of Neuroscience* 30.28, 9431–9444. doi:10.1523/jneurosci.5537-09.2010 [pages 60, 145].

Rechenberg, Ingo (1973). *Ingo Rechenberg Evolutionsstrategie Optimierung technischer Systeme nach Prinzipien der biologishen Evolution*. Fromman-Holzboog Verlag [page 69].

Redgrave, Peter and Gurney, Kevin (2006). The short-latency dopamine signal: a role in discovering novel actions? In: *Nature Reviews Neuroscience* 7.12, 967–975. doi:10.1038/nrn2022 [page 114].

Redgrave, Peter, Gurney, Kevin, Stafford, Tom, Thirkettle, Martin and Lewis, Jen (2012). The Role of the Basal Ganglia in Discovering Novel Actions. In: *Intrinsically Motivated Learning in Natural and Artificial*

Systems. Springer Science + Business Media, 129–150. doi:10.1007/978-3-642-32375-1_6 [page 114].

Regan, Will, Breugel, Floris van and Lipson, Hod (2006). Towards Evolvable Hovering Flight on a Physical Ornithopter. In: *Artificial Life X : Proceedings of the Tenth International Conference on the Simulation and Synthesis of Living Systems*. Ed. by Luis M. Rocha, Larry S. Yaeger, Mark A. Bedau, Dario Floreano, Robert L. Goldstone and Alessandro Vespignani. International Society for Artificial Life. The MIT Press (Bradford Books), 241–247 [page 224].

Riget, J. and Vesterstrøm, J. S. (2002). *A Diversity-Guided Particle Swarm Optimizer—the ARPSO*. Tech. rep. EVALife Technical Report no. 2002-02. EVALife Project Group, Department of Computer Science, Aarhus Universitet [page 118].

Risi, Sebastian, Vanderbleek, Sandy D., Hughes, Charles E. and Stanley, Kenneth O. (2009). How novelty search escapes the deceptive trap of learning to learn. In: *Proceedings of the 11th Annual conference on Genetic and evolutionary computation - GECCO '09*. Association for Computing Machinery (ACM). doi:10.1145/1569901.1569923 [page 117].

Rizzolatti, Giacomo, Fogassi, Leonardo and Gallese, Vittorio (2001). In: *Nature Reviews Neuroscience* 2.9, 661–670. doi:10.1038/35090060 [page 63].

Robbins, Herbert (1952). Some Aspects of the Sequential Design of Experiments. In: *Bull. Amer. Math. Soc.* 58.5, 527–535 [page 162].

Robison, Scott R. and Kelven, Gale A. (2005). Learning to Move Before Birth. In: *Prenatal development of postnatal functions*. Ed. by Brian Hopkins and Scott P. Johnson (Eds). Praeger Publishers, Westport, Oxford. Chap. 5, 131–175. isbn: 0275981266 [page 101].

Rodríguez, José I., Palacios, José, Ruiz, Antonio, Sanchez, Miguel, Alvarez, Ignacio and Demiguel, Enrique (1992). Morphological changes in long bone development in fetal akinesia deformation sequence: An experimental study in curarized rat fetuses. In: *Teratology* 45.2, 213–221. doi:10.1002/tera.1420450215 [page 101].

Rolf, M., Steil, J.J. and Gienger, M. (2011). Online Goal Babbling for rapid bootstrapping of inverse models in high dimensions. In: *Proc. ICDL 2011*. Vol. 2, 1–8. doi:10.1109/DEVLRN.2011.6037368 [page 131].

Rolf, Matthias (2013). Goal Babbling with Unknown Ranges: A Direction-Sampling Approach. In: *IEEE Int. Conf. on Development and Learning and on Epigenetic Robotics (ICDL)*. Osaka, Japan, 1–7. doi:10.1109/DevLrn.2013.6652526 [pages 76, 138].

Rolf, Matthias, Steil, Jochen J. and Gienger, Michael (2010). Mastering Growth while Bootstrapping Sensorimotor Coordination. English. In: Int. Conf. on Epigenetic Robotics. Örenäs Slott, Sweden [page 103].

Rosenblatt, Murray (1956). Remarks on Some Nonparametric Estimates of a Density Function. In: *Ann. Math. Statist.* 27.3, 832–837. doi:10.1214/aoms/1177728190 [page 131].

Ruiz-del-Solar, J., Palma-Amestoy, R., Marchant, R., Parra-Tsunekawa, I. and Zegers, P. (2009). Learning to fall: Designing low damage fall sequences for humanoid soccer robots. In: *Robotics and Autonomous Systems* 57.8, 796–807. doi:10.1016/j.robot.2009.03.011 [page 211].

Rushen, J (1993). Exploration in the pig may not be endogenously motivated. In: *Animal Behaviour* 45.1, 183–184. doi:10.1006/anbe.1993.1016 [page 107].

Russell, James C., McMorland, Angus J. C. and MacKay, Jamie W. B. (2010). Exploratory behaviour of colonizing rats in novel environments. In: *Animal Behaviour* 79.1, 159–164. doi:10.1016/j.anbehav.2009.10.020 [page 109].

Russell, P. A. (1983). Psychological Studies of Exploration in Animals: A Reappraisal. In: *Exploration in Animals and Humans*. Ed. by J. Archer & L. I. A. Birke. Cambridge, U. K.: Van Nostrand Reinhold, 22–54 [page 109].

Rutkowska, J. C. (1994). Scaling Up Sensorimotor Systems: Constraints from Human Infancy. In: *Adaptive Behavior* 2.4, 349–373. doi:10.1177/105971239400200402 [page 147].

Ryan, Richard M. and Deci, Edward L. (2000). Intrinsic and Extrinsic Motivations: Classic Definitions and New Directions. In: *Contemporary Educational Psychology* 25.1, 54–67. doi:10.1006/ceps.1999.1020 [page 112].

Sacks, Jerome, Welch, William J., Mitchell, Toby J. and Wynn, Henry P. (1989). Design and Analysis of Computer Experiments. In: *Statist. Sci.* 4.4, 409–423. doi:10.1214/ss/1177012413 [page 225].

Saleh, Babak, Abe, Kanako, Arora, Ravneet Singh and Elgammal, Ahmed (2014). Toward automated discovery of artistic influence. In: *Multimedia Tools and Applications*. doi:10.1007/s11042-014-2193-x [page 47].

Salge, Christoph, Glackin, Cornelius and Polani, Daniel (2014a). Changing the Environment Based on Empowerment as Intrinsic Motivation. In: *Entropy* 16.5, 2789–2819. doi:10.3390/e16052789 [page 111].

Salge, Christoph, Glackin, Cornelius and Polani, Daniel (2014b). Empowerment — An Introduction. In: *Guided Self-Organization: Inception*. Ed. by Mikhail Prokopenko. Dordrecht: Springer, 67–114. isbn: 978-3-642-53733-2 [page 111].

Salomon, Gavriel and Perkins, D. N. (1987). Transfer of Cognitive Skills from Programming: When and How? In: *Journal of Educational Computing Research* 3.2, 149–169. doi:10.2190/6f4q-7861-qwa5-8pl1 [page 185].

Santucci, Vieri Giuliano, Baldassarre, Gianluca and Mirolli, Marco (2013). Which is the best intrinsic motivation signal for learning multiple skills? In: *Frontiers in Neurorobotics* 7.22. issn: 1662-5218. doi:10.3389/fnbot.2013.00022 [page 121].

Sareni, B. and Krahenbuhl, L. (1998). Fitness sharing and niching methods revisited. In: *IEEE Transactions on Evolutionary Computation* 2.3, 97–106. doi:10.1109/4235.735432 [page 117].

Sasaki, Ryosuke, Yamada, Yasunori, Tsukahara, Yuki and Kuniyoshi, Yasuo (2013). Tactile stimuli from amniotic fluid guides the development of somatosensory cortex with hierarchical structure using human fetus simulation. In: *Proc. ICLD-Epirob 2013*. IEEE. doi:10.1109/devlrn.2013.6652530 [page 102].

Sasamoto, Yuki, Nishijima, Naoto and Minoru, Asada (2013). Towards understanding the origin of infant directed speech: A vocal robot with infant-like articulation. In: *Proc. ICDL-Epirob 2013*. IEEE. doi:10.1109/devlrn.2013.6652562 [page 48].

Saunders, Rob (2002). Curious Design Agents and Artificial Creativity. PhD thesis. University of Sydney [page 120].

Schaal, S. and Atkeson, C. G. (1994). Robot juggling: implementation of memory-based learning. In: *IEEE Control Syst.* 14.1, 57–71. doi:10.1109/37.257895 [page 138].

Schaal, Stefan (1999). Is Imitation Learning the Route to Humanoid Robots? In: *Trends in Cognitive Sciences* 3.6, 233–242. doi:10.1016/s1364-6613(99)01327-3 [page 151].

Schmidhuber, Jürgen (1990). A Possibility for Implementing Curiosity and Boredom in Model-building Neural Controllers. In: *Proceedings of the First International Conference on Simulation of Adaptive Behavior on From Animals to Animats*. Paris, France: MIT Press, 222–227. isbn: 0-262-63138-5 [page 110].

Schmidhuber, Jürgen (1991). Curious Model-Building Control Systems. In: *Proc. 1991 IEEE Int. Joint Conf. on Neural Networks*. IEEE. doi:10.1109/ijcnn.1991.170605 [page 110].

Schmidhuber, Jürgen (1994). *On Learning How to Learn Learning Strategies*. Tech. rep. Technical Report FKI-198-94, Fakultät für Informatik, Technische Universität München [page 163].

Schmidhuber, Jürgen (2009). Driven by Compression Progress: A Simple Principle Explains Essential Aspects of Subjective Beauty, Novelty, Surprise, Interestingness, Attention, Curiosity, Creativity, Art, Science, Music, Jokes. In: *Lecture Notes in Computer Science*. Springer Science + Business Media, 48–76. doi:10.1007/978-3-642-02565-5_4 [page 110].

Schmidhuber, Jürgen (2010). Formal Theory of Creativity, Fun, and Intrinsic Motivation (1990-2010). In: *IEEE Trans. Auton. Mental Dev.* 2.3, 230–247. doi:10.1109/tamd.2010.2056368 [page 110].

Schulkey, Claire E., Regmi, Suk D., Magnan, Rachel A., Danzo, Megan T., Luther, Herman, Hutchinson, Alayna K., Panzer, Adam A., Grady, Mary M., Wilson, David B. and Jay, Patrick Y. (2015). The maternal-age-associated risk of congenital heart disease is modifiable. In: *Nature*. doi:10.1038/nature14361 [page 101].

Schulz, Laura E. and Bonawitz, Elizabeth Baraff (2007). Serious fun: Preschoolers engage in more exploratory play when evidence is confounded. In: *Developmental Psychology* 43.4, 1045–1050. doi:10.1037/0012-1649.43.4.1045 [pages 72, 95, 96].

Schulz, Laura E., Gopnik, Alison and Glymour, Clark (2007). Preschool children learn about causal structure from conditional interventions. In: *Developmental Science* 10.3, 322–332. doi:10.1111/j.1467-7687.2007.00587.x [page 95].

Schütze, Oliver, Esquivel, Xavier, Lara, Adriana and Coello, Carlos A Coello (2010). *Measuring the Averaged Hausdorff Distance to the Pareto Front of a Multi-objective Optimization Problem*. Tech. rep. Technical Report TR-OS-2010-02, CINVESTAV [page 81].

Scribner, Sylvia (1981). *The psychology of literacy*. Cambridge, Mass: Harvard University Press. isbn: 0674721152 [page 185].

Sequeira, Pedro (2013). Socio-Emotional Reward Design for Intrinsically Motivated Learning Agents. In: *PhD Thesis* [page 110].

Settles, Burr (2012). Active Learning. In: *Synthesis Lectures on Artificial Intelligence and Machine Learning* 6.1, 1–114. doi:10.2200/s00429ed1v01y201207aim018 [pages 93, 94].

Seung, H. S., Opper, M. and Sompolinsky, H. (1992). Query by committee. In: *Proceedings of the fifth annual workshop on Computational learning theory - COLT '92*. Association for Computing Machinery (ACM). doi:10.1145/130385.130417 [page 94].

Shalizi, Cosma Rohilla (2001). Causal Architecture, Complexity and Self-Organization in Time Series and Cellular Automata. PhD thesis. University of Wisconsin–Madison [page 97].

Shannon, C. E. (1948). A Mathematical Theory of Communication. In: *Bell System Technical Journal* 27.3, 379–423. doi:10.1002/j.1538-7305.1948.tb01338.x [page 87].

Sherstov, Alexander A. and Stone, Peter (2005). Improving Action Selection in MDP's via Knowledge Transfer. In: *Proceedings of the 20th National Conference on Artificial Intelligence*. Vol. 2. AAAI'05. Pittsburgh, Pennsylvania: AAAI Press. isbn: 1-57735-236-x [page 188].

Shi, Y. and Eberhart, R. (1998). A modified particle swarm optimizer. In: *1998 IEEE International Conference on Evolutionary Computation Proceedings. IEEE World Congress on Computational Intelligence (Cat. No.98TH8360)*. Institute of Electrical & Electronics Engineers (IEEE). doi:10.1109/icec.1998.699146 [page 118].

Shi, Yuhui and Eberhart, Russell C. (2008). Population diversity of particle swarms. In: *2008 IEEE Congress on Evolutionary Computation (IEEE World Congress on Computational Intelligence)*. Institute of Electrical & Electronics Engineers (IEEE). doi:10.1109/cec.2008.4630928 [page 118].

Shim, YoonSik and Husbands, Phil (2007). Feathered Flyer: Integrating Morphological Computation and Sensory Reflexes into a Physically Simulated Flapping-Wing Robot for Robust Flight Manoeuvre. In: *Advances in Artificial Life*. Springer Science + Business Media, 756–765. doi:10.1007/978-3-540-74913-4_76 [page 65].

Shimodaira, Hidetoshi (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. In: *Journal of Statistical Planning and Inference* 90.2, 227–244. doi:10.1016/s0378-3758(00)00115-4 [page 187].

Shohamy, Daphna (2011). Learning and motivation in the human striatum. In: *Current Opinion in Neurobiology* 21.3, 408–414. doi:10.1016/j.conb.2011.05.009 [page 120].

Simon, Herbert A. (1969). *The Sciences of the Artificial*. Cambridge, Massachusetts: M.I.T. Press [page 60].

Simpson, E. H. (1949). Measurement of Diversity. In: *Nature* 163.4148, 688–688. doi:10.1038/163688a0 [page 88].

Sims, Karl (1994). Evolving 3D Morphology and Behavior by Competition. In: *Artificial Life* 1.4, 353–372. doi:10.1162/artl.1994.1.4.353 [page 69].

Şimşek, Özgür, Wolfe, Alicia P. and Barto, Andrew G. (2005). Identifying useful subgoals in reinforcement learning by local graph partitioning. In: *Proceedings of the 22nd international conference on Machine learning - ICML '05*. Association for Computing Machinery (ACM). doi:10.1145/1102351.1102454 [page 188].

Singh, Satinder, Barto, Andrew and Chentanez, Nuttapong (2005). Intrinsically Motivated Reinforcement Learning. In: *Proc. of the 18th Annual Conf. on Neural Information Processing Systems (NIPS'04)* [page 110].

Singh, Satinder, Lewis, Richard L. and Barto, Andrew G. (2009). *Where Do Rewards Come From?* [Pages 109, 110, 112].

Singh, Satinder, Lewis, Richard L, Barto, Andrew G and Sorg, Jonathan (2010). Intrinsically Motivated Reinforcement Learning: An Evolutionary Perspective. In: *IEEE Trans. Auton. Mental Dev.* 2.2, 70–82. doi:10.1109/tamd.2010.2051031 [pages 109–112, 121].

Skinner, B. F. (1938). *The behavior of organisms: an experimental analysis*. Oxford, England: Appleton-Century, p. 457 [page 103].

Skinner, B. F. (1957). *Verbal behavior.* American Psychological Association (APA). doi:10.1037/11256-000 [page 103].

Slonaker, James Rollin (1912). The normal activity of the albino rat from birth to natural death, its rate of growth and the duration of life. In: *Journal of Animal Behavior* 2.1, 20–42. doi:10.1037/h0072764 [page 103].

Small, Willard S. (1899). Notes on the Psychic Development of the Young White Rat. English. In: *The American Journal of Psychology* 11.1, 80–100. issn: 00029556 [page 103].

Smith, Linda B. and Thelen, Esther (2003). Development as a dynamic system. In: *Trends in Cognitive Sciences* 7.8, 343–348. doi:10.1016/s1364-6613(03)00156-6 [page 66].

Smith, R. C. and Cheeseman, P. (1986). On the Representation and Estimation of Spatial Uncertainty. In: *The International Journal of Robotics Research* 5.4, 56–68. doi:10.1177/027836498600500404 [pages 74, 120].

Smith, Randall, Self, Matthew and Cheeseman, Peter (1990). Estimating Uncertain Spatial Relationships in Robotics. In: *Autonomous Robot Vehicles*. Springer New York, 167–193. doi:10.1007/978-1-4613-8997-2_14 [page 120].

Smithers, Tim (1994). On Why Better Robots Make It Harder. In: *Proceedings of the Third International Conference on Simulation of Adaptive Behavior : From Animals to Animats 3: From Animals to Animats 3*. SAB94. Brighton, United Kingdom: MIT Press, 64–72. isbn: 0-262-53122-4 [page 64].

Smithers, Tim (1997). Autonomy in Robots and Other Agents. In: *Brain and Cognition* 34.1, 88–106. doi:10.1006/brcg.1997.0908 [page 52].

Snyder, Kristy M., Ashitaka, Yuki, Shimada, Hiroyuki, Ulrich, Jana E. and Logan, Gordon D. (2013). What skilled typists don't know about the QWERTY keyboard. In: *Attention, Perception, & Psychophysics* 76.1, 162–171. doi:10.3758/s13414-013-0548-4 [page 50].

Sommerville, Jessica A., Woodward, Amanda L. and Needham, Amy (2005). Action experience alters 3-month-old infants' perception of others' actions. In: *Cognition* 96.1, B1–B11. doi:10.1016/j.cognition.2004.07.004 [pages 71, 103].

Sparling, Joyce W, Van Tol, Julia and Chescheir, Nancy C (1999). Fetal and Neonatal Hand Movement. In: *Physical Therapy* 79.1, 24–39 [page 101].

Spence, Kenneth W. (1952). Mathematical formulations of learning phenomena. In: *Psychological Review* 59.2, 152–160. doi:10.1037/h0058010 [page 103].

Spitzer, Nicholas C. (2006). Electrical activity in early neuronal development. In: *Nature* 444.7120, 707–712. doi:10.1038/nature05300 [page 100].

Stachniss, C. and Burgard, W. (2003). Mapping and exploration with mobile robots using coverage maps. In: *Proc. IROS 2003*. IEEE. doi:10.1109/iros.2003.1250673 [page 119].

Stanley, Kenneth O. (2011). Why Evolutionary Robotics Will Matter. In: *Studies in Computational Intelligence*. Springer Science + Business Media, 37–41. doi:10.1007/978-3-642-18272-3_3 [page 116].

Stanley, Kenneth O. and Lehman, Joel (2015). *Why Greatness Cannot Be Planned*. Springer Science + Business Media. doi:10.1007/978-3-319-15524-1 [pages 20, 232].

Stead, John D. H., Clinton, Sarah, Neal, Charles, Schneider, Johanna, Jama, Abas, Miller, Sue, Vazquez, Delia M., Watson, Stanley J. and Akil, Huda (2006). Selective Breeding for Divergence in Novelty-seeking Traits: Heritability and Enrichment in Spontaneous Anxiety-related Behaviors. In: *Behavior Genetics* 36.5, 697–712. doi:10.1007/s10519-006-9058-7 [page 109].

Steels, Luc (2012). Introduction. Self-organization and selection in cultural language evolution. In: *Experiments in Cultural Language Evolution*. John Benjamins Publishing Company, 1–37. doi:10.1075/ais.3.02ste [page 99].

Stephens, P. A., Sutherland, W. J. and Freckleton, R. P. (1999). What Is the Allee Effect? In: *Oikos* 87.1, 185. doi:10.2307/3547011 [page 109].

Stewart, D. E. and Trinkle, J. C. (1996). An Implicit Time-Stepping Scheme for Rigid Body Dynamics with Coulomb Friction. In: *International Journal for Numerical Methods in Engineering* 39.15, 2673–2691. doi:10.1002/(sici)1097-0207(19960815)39:15<2673::aid-nme972>3.0.co;2-i [page 212].

Stout, Andrew, Konidaris, George D. and Barto, Andrew G. (2005). Intrinsically Motivated Reinforcement Learning: A Promising Framework for Developmental Robot Learning. In: *In The AAAI Spring Symposium on Developmental Robotics* [page 110].

Stulp, Freek (2014). `DmpBbo` *– A C++ library for black-box optimization of dynamical movement primitives.* [Page 203].

Stulp, Freek and Oudeyer, Pierre-Yves (2012). Adaptive exploration through covariance matrix adaptation enables developmental motor learning. In: *Paladyn* 3.3, 128–135. doi:10.2478/s13230-013-0108-6 [page 76].

Suh, Il Hong, Lim, Gi Hyun, Hwang, Wonil, Suh, Hyowon, Choi, Jung-Hwa and Park, Young-Tack (2007). Ontology-based multi-layered robot knowledge framework (OMRKF) for robot intelligence. In: *2007 IEEE/RSJ International Conference on Intelligent Robots and Systems*. Institute of Electrical & Electronics Engineers (IEEE). doi:10.1109/iros.2007.4399082 [pages 59, 68].

Sutton, Richard (1998). *Reinforcement learning an introduction*. Cambridge, Mass: MIT Press. isbn: 9780262193986 [page 110].

Sutton, Richard S., Precup, Doina and Singh, Satinder (1999). Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. In: *Artificial Intelligence* 112.1-2, 181–211. doi:10.1016/s0004-3702(99)00052-1 [page 188].

Tang, E. K., Suganthan, P. N. and Yao, X. (2006). An analysis of diversity measures. In: *Machine Learning* 65.1, 247–271. doi:10.1007/s10994-006-9449-2 [page 118].

Taylor, Caz M. and Hastings, Alan (2005). Allee effects in biological invasions. In: *Ecology Letters* 8.8, 895–908. doi:10.1111/j.1461-0248.2005.00787.x [page 109].

Taylor, Charles (1995). *Philosophical Arguments*. Cambridge, Mass: Harvard University Press. isbn: 9780674664760 [page 63].

Taylor, Matthew E. and Stone, Peter (2009). Transfer Learning for Reinforcement Learning Domains: A Survey. In: *J. Mach. Learn. Res.* 10, 1633–1685. issn: 1532-4435 [pages 183, 186, 189].

Tenorth, M. and Beetz, M. (2013). KnowRob: A knowledge processing infrastructure for cognition-enabled robots. In: *The International Journal of Robotics Research* 32.5, 566–590. doi:10.1177/0278364913481635 [pages 59, 68].

Thelen, Esther (1979). Rhythmical stereotypies in normal human infants. In: *Animal Behaviour* 27, 699–715. doi:10.1016/0003-3472(79)90006-x [page 101].

Thelen, Esther (1981a). Kicking, rocking, and waving: Contextual analysis of rhythmical stereotypies in normal human infants. In: *Animal Behaviour* 29.1, 3–11. doi:10.1016/s0003-3472(81)80146-7 [page 101].

Thelen, Esther (1981b). Rhythmical behavior in infancy: An ethological perspective. In: *Developmental Psychology* 17.3, 237–257. doi:10.1037/0012-1649.17.3.237 [page 101].

Thelen, Esther (1995). Mind As Motion. In: ed. by Robert F. Port and Timothy van Gelder. Cambridge, MA, USA: Massachusetts Institute of Technology. Chap. Time-scale Dynamics and the Development of an Embodied Cognition, 69–100. isbn: 0-262-16150-8 [page 66].

Thelen, Esther and Smith, Linda B. (1996). *A Dynamic Systems Approach to the Development of Cognition and Action*. Cambridge, Mass: MIT Press. isbn: 9780262700597 [pages 66, 68].

Thelen, Esther and Smith, Linda B. (2007). Dynamic Systems Theories. In: *Handbook of Child Psychology*. Ed. by William Damon and Richard M. Lerner. Wiley-Blackwell, 258–312. isbn: 9780470147658. doi:10.1002/9780470147658.chpsy0106 [pages 66, 99].

Thompson, D'Arcy Wentworth (1917). *On growth and form.* Smithsonian Institution. doi:10.5962/bhl.title.11332 [page 63].

Thorndike, E. L. and Woodworth, R. S. (1901). The influence of improvement in one mental function upon the efficiency of other functions: III. Functions involving attention, observation and discrimination. In: *Psychological Review* 8.6, 553–564. doi:10.1037/h0071363 [page 185].

Thorndike, Edward L. (1911). *Animal Intelligence: Experimental Studies*. Smithsonian Institution. doi:10.5962/bhl.title.55072 [page 103].

Thorndike, Edward L (1923). The Influence of First-Year Latin Upon Ability to Read English. In: *School & Society* 17, 165–168 [page 185].

Thrun, Sebastian (2005). *Probabilistic Robotics*. Cambridge, Mass: MIT Press. isbn: 9780262201629 [pages 61, 120].

Thrun, Sebastian and Mitchell, Tom M. (1995). Lifelong robot learning. In: *Robotics and Autonomous Systems* 15.1-2, 25–46. doi:10.1016/0921-8890(95)00004-y [page 94].

Thrun, Sebastian and Pratt, Lorien (1998). *Learning to Learn*. Springer Science + Business Media. doi:10.1007/978-1-4615-5529-2 [pages 183, 188].

Till, Mirco S. and Ullmann, G. Matthias (2009). McVol - A program for calculating protein volumes and identifying cavities by a Monte Carlo algorithm. In: *Journal of Molecular Modeling* 16.3, 419–429. doi:10.1007/s00894-009-0541-y [page 240].

Todorov, Emanuel (2014). Convex and analytically-invertible dynamics with contacts and constraints: Theory and implementation in MuJoCo. In: *2014 IEEE International Conference on Robotics and Automation (ICRA)*. Institute of Electrical & Electronics Engineers (IEEE). doi:10.1109/icra.2014.6907751 [page 212].

Todorov, Emanuel and Jordan, Michael I. (2002). Optimal feedback control as a theory of motor coordination. In: *Nature Neuroscience* 5.11, 1226–1235. doi:10.1038/nn963 [page 60].

Trevarthen, Colwyn and Aitken, K. J. (2003). Regulation of Brain Development and Age-Related Changes in Infants' Motives: The Developmental Function of Regressive Periods. In: *Regression Periods in Human Infancy*. Ed. by Mikael Heimann. Mahwah, NJ: Erlbaum, pp. 107–184 [page 122].

Trivedi, Deepak, Rahn, Christopher D., Kier, William M. and Walker, Ian D. (2008). Soft robotics: Biological inspiration, state of the art, and future research. In: *Applied Bionics and Biomechanics* 5.3, 99–117. doi:10.1080/11762320802557865 [page 49].

Trujillo, Leonardo, Olague, Gustavo, Lutton, Evelyne and Vega, Francisco Fernández de (2008). Discovering Several Robot Behaviors through Speciation. In: *Applications of Evolutionary Computing*. Springer Science + Business Media, 164–174. doi:10.1007/978-3-540-78761-7_17 [page 117].

Turing, A. M. (1950). Computing Machinery and Intelligence. In: *Mind* LIX.236, 433–460. doi:10.1093/mind/lix.236.433 [page 66].

Turkewitz, Gerald and Kenny, Patricia A. (1982). Limitations on input as a basis for neural organization and perceptual development: A preliminary theoretical statement. In: *Developmental Psychobiology* 15.4, 357–368. doi:10.1002/dev.420150408 [page 147].

Ulaş, Aydın, Semerci, Murat, Yıldız, Olcay Taner and Alpaydın, Ethem (2009). Incremental construction of classifier and discriminant ensembles. In: *Information Sciences* 179.9, 1298–1318. doi:10.1016/j.ins.2008.12.024 [page 118].

Umiltà, M. A., Kohler, E., Gallese, V., Fogassi, L., Fadiga, L., Keysers, C. and Rizzolatti, G. (2001). I Know What You Are Doing. In: *Neuron* 31.1, 155–165. doi:10.1016/s0896-6273(01)00337-3 [page 63].

Urzelai, Joseba, Floreano, Dario, Dorigo, Marco and Colombetti, Marco (1998). Incremental Robot Shaping. In: *Connection Science* 10.3-4, 341–360. doi:10.1080/095400998116486 [pages 116, 220].

Vanhoutte, Peter and Bading, Hilmar (2003). Opposing roles of synaptic and extrasynaptic NMDA receptors in neuronal calcium signalling and BDNF gene regulation. In: *Current Opinion in Neurobiology* 13.3, 366–371. doi:10.1016/s0959-4388(03)00073-4 [page 100].

Varela, Francisco (1991). *The embodied mind : cognitive science and human experience*. Cambridge, Mass: MIT Press. isbn: 9780262220422 [page 62].

Vargas, Saúl (2014). Novelty and diversity enhancement and evaluation in recommender systems and information retrieval. In: *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval - SIGIR '14*. Association for Computing Machinery (ACM). doi:10.1145/2600428.2610382 [page 119].

Vargas, Saúl and Castells, Pablo (2011). Rank and relevance in novelty and diversity metrics for recommender systems. In: *Proceedings of the fifth ACM conference on Recommender systems - RecSys '11*. Association for Computing Machinery (ACM). doi:10.1145/2043932.2043955 [page 119].

Vaughan, Eric D., Paolo, Ezequiel Di and Harvey, Inman R. (2004). The Evolution of Control and Adaptation in a 3D Powered Passive Dynamic Walker. In: *In Proceedings of the Ninth International Conference on the Simulation and Synthesis of Living Systems, ALIFE'9*. MIT Press, 139–145 [page 64].

Vincent, Julian F. V., Bogatyreva, Olga A., Bogatyrev, Nikolaj R., Bowyer, Adrian and Pahl, Anja-Karina (2006). Biomimetics: its practice and theory. In: *Journal of The Royal Society Interface* 3.9, 471–482. doi:10.1098/rsif.2006.0127 [page 53].

Vygotsky, Lev S (1978). *Mind and society: The development of higher mental processes*. Cambridge, Massachusetts: Harvard University Press [page 105].

Wahby, Mostafa and Hamann, Heiko (2015). On the Tradeoff between Hardware Protection and Optimization Success: A Case Study in Onboard Evolutionary Robotics for Autonomous Parallel Parking. In: *EvoStar 2015* [pages 207, 224].

Waibel, Markus, Beetz, Michael, Civera, Javier, D'Andrea, Raffaello, Elfring, Jos, Gálvez-López, Dorian, HÃ¤ussermann, Kai, Janssen, Rob, Montiel, J. M. M., Perzylo, Alexander, Schießle, BjĂśrn, Tenorth, Moritz, Zweigle, Oliver and Molengraft, René De (2011). RoboEarth. In: *IEEE Robot. Automat. Mag.* 18.2, 69–82. doi:10.1109/mra.2011.941632 [page 189].

Walker, Edward L (1964). Psychological complexity as a basis for a theory of motivation and choice. In: *Nebraska symposium on motivation*. University of Nebraska Press [page 105].

Wang, Gai-yun and Han, Dong-xue (2009). Particle Swarm Optimization Based on Self-adaptive Acceleration Factors. In: *2009 Third International Conference on Genetic and Evolutionary Computing*. Institute of Electrical & Electronics Engineers (IEEE). doi:10.1109/wgec.2009.55 [page 118].

Wang, Zheng, Song, Yangqiu and Zhang, Changshui (2008). Transferred Dimensionality Reduction. In: *Machine Learning and Knowledge Discovery in Databases*. Springer Science + Business Media, 550–565. doi:10.1007/978-3-540-87481-2_36 [page 188].

Warrick, D. R. (2002). Bird Maneuvering Flight: Blurred Bodies, Clear Heads. In: *Integrative and Comparative Biology* 42.1, 141–148. doi:10.1093/icb/42.1.141 [page 65].

Watson, John B (1913). Psychology as the behaviorist views it. In: *Psychological review* 20.2, 158 [page 103].

Watson, Sheree L., Shively, Carol A. and Voytko, Mary Lou (1999). Can puzzle feeders be used as cognitive screening instruments? Differential performance of young and aged female monkeys on a puzzle feeder task. In: *American Journal of Primatology* 49.2, 195–202. doi:10.1002/(sici)1098-2345(199910)49:2<195::aid-ajp9>3.0.co;2-j [page 108].

Webb, Barbara (2000). What does robotics offer animal behaviour? In: *Animal Behaviour* 60.5, 545–558. doi:10.1006/anbe.2000.1514 [page 52].

Webb, Barbara (2001). Can robots make good models of biological behaviour? In: *Behavioral and Brain Sciences* 24.06. doi:10.1017/s0140525x01000127 [page 52].

Webb, Barbara (2002). Robots in invertebrate neuroscience. In: *Nature* 417.6886, 359–363. doi:10.1038/417359a [page 52].

Weng, J. (2001). Artificial Intelligence: Autonomous Mental Development by Robots and Animals. In: *Science* 291.5504, 599–600. doi:10.1126/science.291.5504.599 [pages 53, 66].

White, Robert W. (1959). Motivation Reconsidered: The Concept of Competence. In: *Psychological Review* 66.5, 297–333. doi:10.1037/h0040934 [pages 105, 107].

Whittaker, R. H. (1972). Evolution and Measurement of Species Diversity. In: *Taxon* 21.2/3, 213. doi:10.2307/1218190 [page 87].

Whittle, P. (1988). Restless Bandits: Activity Allocation in a Changing World. In: *Journal of Applied Probability* 25, 287. doi:10.2307/3214163 [page 163].

Wilson, T. D., Reinhard, D. A., Westgate, E. C., Gilbert, D. T., Ellerbeck, N., Hahn, C., Brown, C. L. and Shaked, A. (2014). Just think: The challenges of the disengaged mind. In: *Science* 345.6192, 75–77. doi:10.1126/science.1250830 [page 104].

Wisse, Martijn, Keliksdal, Guillaume, Frankenhyyzen, Jan and Moyer, Brian (2007). Passive-Based Walking Robot. In: *IEEE Robot. Automat. Mag.* 14.2, 52–62. doi:10.1109/mra.2007.380639 [page 64].

Wolf, Tom De and Holvoet, Tom (2005). Emergence Versus Self-Organisation: Different Concepts but Promising When Combined. In: *Lecture Notes in Computer Science*. Springer Science + Business Media, 1–15. doi:10.1007/11494676_1 [pages 96–98].

Wolff, Peter (1987). *The development of behavioral states and the expression of emotions in early infancy : new proposals for investigation*. Chicago: University of Chicago Press. isbn: 9780226905204 [page 70].

Wolpert, D. M. and Kawato, M. (1998). Multiple paired forward and inverse models for motor control. In: *Neural Networks* 11.7-8, 1317–1329. doi:10.1016/s0893-6080(98)00066-5 [page 54].

Wood-Gush, D. G. M. and Vestergaard, K. (1991). The seeking of novelty and its relation to play. In: *Animal Behaviour* 42.4, 599–606. doi:10.1016/s0003-3472(05)80243-x [page 107].

Wood-Gush, D. G. M., Vestergaard, K. and Petersen, H. V. (1990). The significance of motivation and environment in the development of exploration in pigs. In: *Biology of Behaviour* 15.1, 39–52 [page 107].

Wood-Gush, David G. M. and Vestergaard, Klaus (1993). Inquisitive exploration in pigs. In: *Animal Behaviour* 45.1, 185–187. doi:10.1006/anbe.1993.1017 [page 107].

Woodworth, Robert S (1947). Reenforcement of perception. In: *The American journal of psychology*, 119–124 [page 105].

Woodworth, Robert S (1958). Dynamics of behavior. In: [page 105].

Wright, Sewall (1932). The Roles of Mutation, Inbreeding, Crossbreeding, and Selection in Evolution. In: vol. 1, 355–366 [page 98].

Wright, T. F., Eberhard, J. R., Hobson, E. A., Avery, M. L. and Russello, M. A. (2010). Behavioral flexibility and species invasions: the adaptive flexibility hypothesis. In: *Ethology Ecology & Evolution* 22.4, 393–404. doi:10.1080/03949370.2010.505580 [page 109].

Xing, Dikan, Dai, Wenyuan, Xue, Gui-Rong and Yu, Yong (2007). Bridged Refinement for Transfer Learning. In: *Knowledge Discovery in Databases: PKDD 2007*. Springer Science + Business Media, 324–335. doi:10.1007/978-3-540-74976-9_31 [page 188].

Xu, Bo, Min, Huaqing and Xiao, Fangxiong (2014). A brief overview of evolutionary developmental robotics. In: *Industrial Robot* 41.6, 527–533. doi:10.1108/ir-04-2014-0324 [page 69].

Yamada, Yasunori, Fujii, Keiko and Kuniyoshi, Yasuo (2013). Impacts of Environment, Nervous System and Movements of Preterms on Body Map Development: Fetus Simulation with Spiking Neural Network. In: *Proc. of ICDL-Epirob 2013*. IEEE. doi:10.1109/devlrn.2013.6652548 [page 102].

Yen, G. G. and Daneshyari, M. (2006). Diversity-based Information Exchange among Multiple Swarms in Particle Swarm Optimization. In: *2006 IEEE International Conference on Evolutionary Computation*. Institute of Electrical & Electronics Engineers (IEEE). doi:10.1109/cec.2006.1688511 [page 118].

Zagal, Juan Cristobal, Ruiz-del-Solar, Javier and Palacios, Adrian Galo (2008). Fitness Based Identification of a Robot Structure. In: *Artificial Life* 11, 733 [page 225].

Zeng, Ming and Ren, Jiangtao (2012). Domain Transfer Dimensionality Reduction via Discriminant Kernel Learning. In: *Lecture Notes in Computer Science*. Springer Science + Business Media, 280–291. doi:10.1007/978-3-642-30220-6_24 [page 188].

Zhou, T., Kuscsik, Z., Liu, J. -G., Medo, M., Wakeling, J. R. and Zhang, Y. -C. (2010). Solving the apparent diversity-accuracy dilemma of recommender systems. In: *Proceedings of the National Academy of Sciences* 107.10, 4511–4515. doi:10.1073/pnas.1000488107 [page 119].

Zhu, Ciyou, Byrd, Richard H., Lu, Peihuang and Nocedal, Jorge (1997). Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. In: *ACM Trans. Math. Softw.* 23.4, 550–560. doi:10.1145/279232.279236 [pages 243, 244].

Ziegler, Cai-Nicolas, McNee, Sean M., Konstan, Joseph A. and Lausen, Georg (2005). Improving recommendation lists through topic diversification. In: *Proceedings of the 14th international conference on World Wide Web - WWW '05*. Association for Computing Machinery (ACM). doi:10.1145/1060745.1060754 [page 119].

Zito, Karen and Svoboda, Karel (2002). Activity-Dependent Synaptogenesis in the Adult Mammalian Cortex. In: *Neuron* 35.6, 1015–1017. doi:10.1016/s0896-6273(02)00903-0 [page 100].

Zykov, Viktor, Bongard, Josh and Lipson, Hod (2004). Evolving Dynamic Gaits on a Physical Robot. In: *Proceedings of Genetic and Evolutionary Computation Conference, Late Breaking Paper, GECCO'04* [pages 69, 224].