Curriculum Vitae – Pedro A. Ortega

Contact Information

pedro.ortega@gmail.com
http://www.adaptiveagents.org
Updated: November 2022

Summary

Artificial intelligence and machine learning research, with expertise in reinforcement learning, Bayesian statistics, deep learning, causality, and safety.

Education

PhD in Machine Learning, U. of Cambridge (2006-2011). Thesis supervisor: Zoubin Ghahramani. Thesis: *A Unified Framework for Resource-Bounded Autonomous Agents Interacting with Unknown Environments.* Cambridge, United Kingdom.

BSc Computer Engineering, U. of Chile (1998-2005). Dissertation: *Design of Interactive Processing Mechanisms for the Analysis of Brain Waves* (in Spanish). Summa Cum Laude (relative position 13 of 542). Santiago, Chile.

Recent Work

2022 VP of Research, Kosen Labs, London, U.K.

2016-2022 Lead of AI Safety Analysis Research, DeepMind, London, U.K.

2013-2016 Postdoc, U. of Pennsylvania. Work with Prof. Daniel Lee and Prof. Alan A. Stocker. Philadelphia, U.S.A.

2013 Postdoc, Hebrew University in Jerusalem. Work with Prof. Naftali Tishby. Jerusalem, Israel.

2011-2013 Postdoc, Max Planck Institute for Intelligent Systems and Biological Cybernetics. Work with Daniel A. Braun, Tübingen, Germany.

2010-2011 Postdoc, University of Cambridge. Work with Prof. Simon Godsill and Prof. Patrick Langdon. Cambridge, U.K.

Scholarships and Distinctions

2006-2010 PhD Scholarship, Government of Chile.

1998-2002 Honor roll student (top 5%), School of Engineering, University of Chile.

1997 Scholarship “Liceo Manuel de Salas”, Santiago, Chile.

1996 Winner of National Physics Olympics, Chile.

1996 Winner of Regional Physics Olympics, Santiago, Chile.

Languages

Spanish and **Swiss German** (Native);
English, **German** and **Italian** (Full Proficiency).

Organization

Organization: NIPS Program Manager (2015); NIPS Workshop on Novel Trends and Applications in Reinforcement Learning (2014); NIPS Workshop on Planning with Information Constraints (2013); Max Planck on Intelligent Systems Colloquium Series (2012). Univeristy of Chile, “Artificial Life Group” at the Laboratory of Electrotechnologies (2002).

Program Committee Member: NIPS 2019, WHY 2019, NIPS 2018, NIPS 2017, ICML 2016, NIPS 2016, UAI 2016, ICML 2016, AAAI 2015, AISTATS 2015, ICML 2015, RLDM 2015, COLT 2015, AAAI 2014, EWRL 2012, Max Planck–Chile Research Seminar 2012.

Journal and Grant Reviewer: Annals of Statistics, Journal of Machine Learning Research, Philosophical Transactions of the Royal Society, PLOS Computational Biology, Transactions on Intelligent Systems and Technology, Pattern Recognition Letters, Journal of Artificial General Intelligence, Neural Computation, Cognitive Science, Fondecyt (National Science & Technology Foundation of Chile), Frontiers, Entropy.

Invited Talks & Panels

Information-Theoretic Bounded Rationality. Decision theory and the future of AI. Speaker. Cambridge, UK, 2017.

Information-Theoretic Bounded Rationality. Workshop on Decision-Making and Behavioral Variability, Keynote Speaker, Columbia University, 2017.

Agency and Causality. NIPS Workshop on Imperfect Decision-Makers, Barcelona, 2016.

NIPS Workshop on Bounded Optimality and Rational Meta-Reasoning, panel member, 2015.

Ortega, P.A. *Information-Theoretic Bounded Rationality.* Dept. of Aerospace Engineering, Georgia Tech; Laboratory for Information & Decision Systems, MIT; EU Workshop on Information & Decision Making, University of Hertfordshire; Gatsby Computational Neuroscience Unit; Google Deepmind, 2015.

Ortega, P.A. and Braun, D.A. *Adaptive Coding of Actions and Observations.* NIPS Workshop on Information Theory of Perception and Action, Lake Tahoe, 2012.

Ortega, P.A. and Braun, D.A. *Free Energy and Bounded Rationality.* Workshop on Statistical Physics of Inference and Control, Granada, 2012.

Ortega, P.A. and Braun, D.A. *Free Energy and Bounded Rationality.* Combining Information Theory and Game Theory, Santa Fe Institute, 2012.

Ortega, P.A. *Mathematical Foundations of Bounded Rationality.* Max Planck - Chile Research Seminar, 2012.

Ortega, P.A. and Braun, D.A. *Free Energy and the Generalized Optimality Equations for Sequential Decision Making.* European Workshop on Reinforcement Learning, 2012.

Ortega, P.A. *Bayesian Causal Induction.* NIPS Workshop on Philosophy and Machine Learning, 2011.

Publications

Beyond Bayes-optimality: meta-learning what you know you don't know Grau-Moya J., Delétang G., Kunesch M., Genewein T., Catt E., Li W.K., Ruoss A., Cundy C., Veness J., Wang J.X., Hutter M., Summerfield C., Legg S., Ortega P.A. ArXiv:2207.02098, 2022.

Neural Networks and the Chomsky Hierarchy. Delétang G., Ruoss A., Grau-Moya J., Genewein T., Wenliang L.K., Catt E., Hutter M., Legg S., Ortega P.A. ArXiv:2207.02098, 2022.

Your Policy Regularizer is Secretly an Adversary. Brekelmans R., Genewein T., Grau-Moya J., Delétang G., Kunesch M., Legg S., Ortega P.A. Transactions on Machine Learning Research, 2022.

Model-Free Risk-Sensitive Reinforcement Learning. Delétang G., Grau-Moya J., Kunesch M., Genewein T., Brekelmans R., Legg S., Ortega P.A. DeepMind Technical Report, ArXiv:2111.02907, 2021.

Shaking the foundations: delusions in sequence models for interaction and control. Ortega P.A., Kunesch M., Delétang G., Genewein T., Grau-Moya J., Veness J., Buchli J., Degrave J., Piot B., Perolat J., Everitt T., Tallec C., Parisotto E., Erez T., Chen Y., de Freitas, N., Legg S. DeepMind Technical Report, ArXiv:2110.10819, 2021.

Causal Analysis of Agent Behavior for AI Safety. Déletang G., Grau-Moya J., Martic M., Genewein T., McGrath T., Mikulik V., Kunesch M., Legg S., Ortega P.A. ArXiv:2010.12237, 2020.

From Poincaré Recurrence to Convergence in Imperfect Information Games: Finding Equilibrium via Regularization. Perolat J., Munos R., Lespiau J.-B., Omidshafiei S., Rowland M., Ortega P.A., Burch N., Anthony T., Balduzzi D., De Vylder B., Piliouras G., Lanctot M., Tuyls K. International Conference on Machine Learning (ICML), 2021.

Agent Incentives: A Causal Perspective. Everitt T., Carey R., Langlois E., Ortega P.A., Legg S. AAAI Conference on Artificial Intelligence, 2020

Meta-trained agents implement Bayes-optimal agents. Mikulik V., Delétang G., McGrath T., Genewein T., Martic M., Legg S., Ortega P.A. Neural Information Processing Systems (NIPS), 2020

Algorithms for Causal Reasoning in Probability Trees. Genewein T., McGrath T., Delétang G., Mikulik V., Martic M., Legg S., Ortega P.A. ArXiv:2010.12237, 2020

Action and Perception as Divergence Minimization. Hafner D., Ortega P.A., Ba J., Parr T., Friston K., Heess N. arXiv:2009.01791, 2020

Meta reinforcement learning as task inference. Humplik J., Galashov A., Hasenclever L., Ortega P.A., Teh Y.W., Heess N. arXiv:1905.06424, 2019

Intrinsic Social Motivation via Causal Influence in Multi-Agent RL. Jaques N., Lazaridou A., Hughes E., Gulcehre C., Ortega P.A., Strouse D.J., Leibo J.Z., de Freitas N. International Conference on Machine Learning (ICML), 2019

Understanding Agent Incentives using Causal Influence Diagrams. Part I: Single Action Settings. Everitt T., Ortega P.A., Barnes E., Legg S. arXiv:1902.09980, 2019

Meta-learning of Sequential Strategies. Ortega P.A., Wang J.X., Rowland M., Genewein T., Kurth-Nelson Z., Pascanu R., Heess N., Veness J., Pritzel A., Sprechmann P., Jayakumar S.M., McGrath T., Miller K., Azar M., Osband I., Rabinowitz N., György A., Chiappa S., de Freitas N., Osindero S., Teh Y.W., van Hasselt H., Botvinick M., Legg S. To appear, 2019

Causal Reasoning from Meta-reinforcement Learning. Dasgupta I., Wang J., Chiappa S., Mitrovic J., Ortega P.A., Raposo D., Hughes E., Battaglia E., Botvinick M., Kurth-Nelson Z. arXiv:1901.08162, 2019.

Bayesian Optimistic Kullback-Leibler Exploration. Lee K., Kim G.-H., Ortega P.A., Lee D.D., and Kim K.-E. Machine Learning, 2018.

Modelling Friends and Foes. Ortega, P.A. and Legg, S. ArXiv:1807.00196, 2018.

AI safety gridworlds. Leike, J., Martic, M., Krakovna, V., Ortega, P.A., Everitt, T., Lefrancq, A., Orseau, L. and Legg, S. ArXiv:1711.09883, 2017.

Memory controls time perception and intertemporal choices. Ortega, P.A. and Tishby, N. ArXiv:1604.05129, 2016.

Human Decision-Making under Limited Time. Ortega, P.A., and Stocker, A. Neural Information Processing Systems (NIPS), 2016.

Bayesian Reinforcement Learning with Behavioral Feedback. Hong, T., Lee, J., Kim, K.-E., Ortega, P.A., and Lee, D.D. International Joint Conference on Artificial Intelligence (IJCAI), 2016.

Information-Theoretic Bounded Rationality. Ortega, P.A., Braun, D.A., Dyer, J.S., Kim, K.-E., and Tishby, N. ArXiv:1512.06789, 2015.

Decision-making under ambiguity is modulated by visual framing, but not by motor vs. non-motor context. Experiments and an information-theoretic ambiguity model. Grau-Moya, J. and Ortega, P.A. and Braun, D.A. PLoS One, 2015.

Commentary: What is epistemic value in free energy models of learning and acting? A bounded rationality perspective. Ortega, P.A. and Braun, D.A. Cognitive Neuroscience, 2015.

Subjectivity, Bayesianism, and causality. Ortega, P.A. Special Issue on Philosophical Aspects of Pattern Recognition, Pattern Recognition Letters, pp. 63–70, 2015.

Causal reasoning in a prediction task with hidden causes. Ortega, P.A. and Lee, D.D. and Stocker, A.A. 37th Annual Cognitive Science Society Meeting (CogSci), 2015.

Reactive bandits with attitude. Ortega, P.A. and Kim, K.-E. and Lee, D.D. The 18th International Conference on Artificial Intelligence and Statistics (AISTATS), 2015.

Belief flows for robust online learning. Ortega, P.A. and Crammer, K. and Lee, D.D. Information Theory and Applications (ITA), 2015.

Perceptual adaptation: getting ready for the future. Wei, X.-X. and Ortega, P.A. and Stocker, A.A. Computational and Systems Neuroscience (Cosyne), 2015.

Information-theoretic bounded rationality and ε-optimality. Braun, D.A. and Ortega, P.A. Entropy 16(8), 4662-4676, 2014.

An adversarial interpretation of information-theoretic bounded rationality. Ortega, P.A. and Lee, D.D. Twenty-Eighth AAAI Conference on Artificial Intelligence (AAAI '14), 2014.

Generalized Thompson sampling for sequential decision-making and causal inference. Ortega, P.A. and Braun, D.A. Complex Adaptive Systems Modeling 2:2, 2014

Dynamic belief state representations. Lee, D.D. and Ortega, P.A. and Stocker, A. Current Opinion in Neurobiology, 2014.

Monte Carlo methods for exact & efficient solution of the generalized optimality equations. Ortega, P.A. and Braun, D.A. and Tishby, N. IEEE International Conference on Robotics and Automation (ICRA), 2014.

Thermodynamics as a theory of information processing costs. Ortega, P.A. and Braun, D.A. Proceedings of the Royal Society A, 20120683, 2013.

Metabolic cost as an organizing principle for cooperative learning. Balduzzi, D., Ortega, P.A. and Besserve, M. Advances in Complex Systems, 2013.

Adaptive coding of actions and observations. Ortega, P.A. and Braun, D.A. NIPS Workshop on Information in Perception and Action, 2012.

A nonparametric conjugate prior distribution for the maximizing argument of a noisy function. Ortega, P.A., Grau-Moya, J., Genewein, T., Balduzzi, D. and Braun, D.A. Neural Information Processing Systems (NIPS) 2012.

Risk-sensitivity in Bayesian sensorimotor integration. Grau-Moya, J., Ortega, P.A. and Braun, D.A. PLOS Computational Biology 8(9): e1002698, 2012.

Free energy and the generalized optimality equations for sequential decision making. Ortega, P.A. and Braun, D.A. European Workshop on Reinforcement Learning, 2012.

Bayesian causal induction. Ortega, P.A. NIPS Workshop on Philosophy and Machine Learning, 2011.

Information, utility and bounded rationality. Ortega, P.A. and Braun, D.A. The fourth conference on artificial general intelligence, Springer-Verlag, 2011.

Reinforcement learning and the Bayesian control rule. Ortega, P.A. and Braun, D.A. and Godsill, S.J. The fourth conference on artificial general intelligence, Springer-Verlag, 2011.

Motor coordination: when two have to act as one. Braun, D.A., Ortega, P.A. and Wolpert, D.M. Special issue of Experimental Brain Research on Joint Action, 2011.

A unified framework for resource-bounded autonomous agents interacting with unknown environments. Ortega, P.A. PhD dissertation. Department of Engineering, University of Cambridge, 2011.

Path integral control and bounded rationality. Braun, D.A., Ortega, P.A., Theodorou, E. and Schaal, S. SSCI 2011 ADPRL - 2011 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning, 2011.

A minimum relative entropy principle for learning and acting. Ortega, P.A. and Braun, D.A. Journal of Artificial Intelligence Research 38, pp. 475-511, 2010.

A minimum relative entropy principle for adaptive control in linear quadratic regulators. Braun, D.A. and Ortega, P.A. Proceedings of the 7th international conference on informatics in control, automation and robotics, pp. 103-108, 2010.

A conversion between utility and information. Ortega, P.A. and Braun, D.A. The third conference on artificial general intelligence, pp. 115-120, 2010.

A Bayesian rule for adaptive control based on causal interventions. Ortega, P.A. and Braun, D.A. The third conference on artificial general intelligence, pp. 121-126, 2010.

Nash equilibria in multi-agent motor interactions. Braun D.A., Ortega P.A. & Wolpert D.M. PLoS Computational Biology 5 (8):e1000468, 2010.

Error backpropagation with generalized functional composition. Bassi, A. and Ortega, P.A. Technical Report, Department of Computer Science, University of Chile, 2006.

A medical claim fraud/abuse detection system based on data mining: a case study in Chile. Ortega, P.A. and Figueroa, C. and Ruz, G. DMIN 2006:224-231, 2006.

Diseño de mecanismos de procesamiento interactivos para el anásis de ondas cerebrales (design of interactive processing mechanisms for the analysis of brain waves). Ortega, P.A. BSc dissertation. School of Engineering, University of Chile, 2005.