## ACKNOWLEDGEMENT

This work has received funding from the European Commission H2020 framework program under the research Grant number 687831 (BabyRobot)

## REFERENCES

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450* (2016).
[2] Bram Bakker. 2002. Reinforcement learning with long short-term memory. In *Proc.of NIPS*.
[3] Senthilkumar Chandramohan, Matthieu Geist, Fabrice Lefevre, and Olivier Pietquin. 2011. User simulation in dialogue systems using inverse reinforcement learning. In *Interspeech 2011*. 1025–1028.
[4] Lu Chen, Runzhe Yang, Cheng Chang, Zihao Ye, Xiang Zhou, and Kai Yu. 2017. On-line Dialogue Policy Learning with Companion Teaching. *Proc. of EACLK* (2017).
[5] Sonia Chernova and Manuela Veloso. 2009. Interactive policy learning through confidence-based autonomy. *Journal of Artificial Intelligence Research* 34, 1 (2009), 1.
[6] Jeffery Allen Clouse. 1996. *On integrating apprentice learning and reinforcement learning*. Technical Report. Amherst, MA, USA.
[7] Jeffery A Clouse and Paul E Utgoff. 1992. A teaching method for reinforcement learning. In *Proc. of ICML*.
[8] Lucie Daubigney, Matthieu Geist, and Olivier Pietquin. 2012. Off-policy learning in large-scale POMDP-based dialogue systems. In *Proc. of ICASSP*.
[9] Layla El Asri, Jing He, and Kaheer Suleman. 2016. A sequence-to-sequence model for user simulation in spoken dialogue systems. In *Proc. of Interspeech*.
[10] Layla El Asri, Romain Laroche, and Olivier Pietquin. 2013. Reward shaping for statistical optimisation of dialogue management. In *Proc. of SLSP*.
[11] Layla El Asri, Bilal Piot, Matthieu Geist, Romain Laroche, and Olivier Pietquin. 2016. Score-based inverse reinforcement learning. In *Proc. of AAMAS*.
[12] Mehdi Fatemi, Layla El Asri, Hannes Schulz, Jing He, and Kaheer Suleman. 2016. Policy networks with two-stage training for dialogue systems. *Proc. of SigDial*.
[13] Milica Gašić, Filip Jurčíček, Simon Keizer, François Mairesse, Blaise Thomson, Kai Yu, and Steve Young. 2010. Gaussian processes for fast policy optimisation of POMDP-based dialogue managers. In *Proc. of SigDial*.
[14] Geoffrey J Gordon. 1995. Stable function approximation in dynamic programming. In *Proc. of ICML*.
[15] Shane Griffith, Kaushik Subramanian, Jonathan Scholz, Charles Isbell, and Andrea L Thomaz. 2013. Policy shaping: Integrating human feedback with reinforcement learning. In *Proc. of NIPS*.
[16] Anna Harutyunyan, Sam Devlin, Peter Vrancx, and Ann Nowé. 2015. Expressing Arbitrary Reward Functions as Potential-Based Advice.. In *Proc. of AAAI*.
[17] Matthew Henderson, Blaise Thomson, and Jason Williams. 2014. The second dialog state tracking challenge. In *Proc. of SigDial*.
[18] Todd Hester, Matej Vecerik, Olivier Pietquin, Marc Lanctot, Tom Schaul, Bilal Piot, Andrew Sendonaris, Gabriel Dulac-Arnold, Ian Osband, John Agapiou, et al. 2017. Learning from Demonstrations for Real World Reinforcement Learning. *arXiv preprint arXiv:1704.03732* (2017).
[19] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.
[20] Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proc. of ICML*.
[21] Hatim Khouzaimi, Romain Laroche, and Fabrice Lefevre. 2015. Optimising turn-taking strategies with reinforcement learning. In *Proc. of SigDial*.
[22] Beomjoon Kim, Amir massoud Farahmand, Joelle Pineau, and Doina Precup. 2013. Learning from limited demonstrations. In *Proc. of NIPS*.
[23] W Bradley Knox and Peter Stone. 2009. Interactively shaping agents via human reinforcement: The TAMER framework. In *Proc. of ICKC*.
[24] W Bradley Knox and Peter Stone. 2010. Combining manual feedback with subsequent MDP reward signals for reinforcement learning. In *Proc. of AAMAS*.
[25] J Zico Kolter and Andrew Y Ng. 2009. Regularization and feature selection in least-squares temporal difference learning. In *Proc. of ICML*.
[26] Michail G Lagoudakis and Ronald Parr. 2003. Least-squares policy iteration. *Journal of machine learning research* 4, Dec (2003), 1107–1149.
[27] Sascha Lange, Thomas Gabel, and Martin Riedmiller. 2012. Batch reinforcement learning. In *Reinforcement learning*. Springer, 45–73.
[28] Romain Laroche, Ghislain Putois, and Philippe Bretier. 2010. Optimising a hand-crafted dialogue system design.. In *Proc. of Interspeech*.
[29] Romain Laroche and Paul Trichelair. 2017. Safe Policy Improvement with Baseline Bootstrapping. *arXiv preprint arXiv:1712.06924* (2017).
[30] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature* 521, 7553 (2015), 436–444.
[31] Esther Levin and Roberto Pieraccini. 1997. A stochastic model of computer-human interaction for learning dialogue strategies.. In *Proc. of Eurospeech*.
[32] Lihong Li, Jason D Williams, and Suhrid Balakrishnan. 2009. Reinforcement learning for dialog management using least-squares Policy iteration and fast feature selection.. In *Proc. of Interspeech*.
[33] Ryan Thomas Lowe, Nissan Pow, Iulian Vlad Serban, Laurent Charlin, Chia-Wei Liu, and Joelle Pineau. 2017. Training end-to-end dialogue systems with the ubuntu dialogue corpus. *Dialogue & Discourse* 8, 1 (2017), 31–65.
[34] Richard Maclin and Jude W Shavlik. 1996. Creating advice-taking reinforcement learners. *Machine Learning* 22, 1-3 (1996), 251–281.
[35] Andrew Y Ng, Daishi Harada, and Stuart Russell. 1999. Policy invariance under reward transformations: Theory and application to reward shaping. In *Proc. of ICML*.
[36] Olivier Pietquin, Matthieu Geist, Senthilkumar Chandramohan, and Hervé Frezza-Buet. 2011. Sample-efficient batch reinforcement learning for dialogue management optimization. *ACM Transactions on Speech and Language Processing (TSLP)* 7, 3 (2011), 7.
[37] Olivier Pietquin and Steve Renals. 2002. ASR system modeling for automatic evaluation and optimization of dialogue systems. In *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, Vol. 1. IEEE, I–45.
[38] Bilal Piot, Matthieu Geist, and Olivier Pietquin. 2014. Boosted Bellman residual minimization handling expert demonstrations. In *Proc. of ECML*.
[39] Martin L Puterman. 2014. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons.
[40] Ghislain Putois, Romain Laroche, and Philippe Bretier. 2010. Online reinforcement learning for spoken dialogue systems: The story of a commercial deployment success. In *Proc. of SIGDIAL*.
[41] Jette Randlov and Preben Alstrom. 1998. Learning to drive a bicycle using reinforcement learning and shaping. In *Proc. of ICML*.
[42] Stéphane Ross, Geoffrey J Gordon, and Drew Bagnell. 2011. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proc. of AISTATS*.
[43] Jost Schatzmann, Karl Weilhammer, Matt Stuttle, and Steve Young. 2006. A survey of statistical user simulation techniques for reinforcement-learning of dialogue management strategies. *The knowledge engineering review* 21, 2 (2006), 97–126.
[44] Satinder P Singh, Michael J Kearns, Diane J Litman, and Marilyn A Walker. 2000. Reinforcement learning for spoken dialogue systems. In *Proc. of NIPS*.
[45] Florian Strub, Harm de Vries, Jeremie Mary, Bilal Piot, Aaron Courville, and Olivier Pietquin. 2017. End-to-end optimization of goal-driven and visually grounded dialogue systems. *Proc. of IJCAI*.
[46] Pei-Hao Su, David Vandyke, Milica Gasic, Nikola Mrksic, Tsung-Hsien Wen, and Steve Young. 2015. Reward shaping with recurrent neural networks for speeding up on-line policy learning in spoken dialogue systems. *Proc. of SigDial* (2015).
[47] Richard S Sutton and Andrew G Barto. 1998. *Reinforcement learning: An introduction*. Vol. 1. MIT press Cambridge.
[48] Blaise Thomson, Milica Gasic, Matthew Henderson, Pirros Tsiakoulis, and Steve Young. 2012. N-best error simulation for training spoken dialogue systems. In *Spoken Language Technology Workshop (SLT), 2012 IEEE*. 37–42.
[49] Tsung-Hsien Wen, Milica Gasic, Nikola Mrksic, Pei-Hao Su, David Vandyke, and Steve Young. 2015. Semantically conditioned lstm-based natural language generation for spoken dialogue systems. In *Proc. of SigDial*.
[50] Eric Wiewiora, Garrison Cottrell, and Charles Elkan. 2003. Principled methods for advising reinforcement learning agents. In *Proc. of ICML*.
[51] Jason D Williams. 2008. The best of both worlds: unifying conventional dialog systems and POMDPs.. In *Proc. of Interspeech*.
[52] Jason D Williams and Steve Young. 2003. Using Wizard-of-Oz simulations to bootstrap Reinforcement-Learning based dialog management systems. In *Proc. of SIGDIAL*.