
Modeling and Interpreting Real-world Human Risk Decision Making with Inverse Reinforcement Learning

Quanying Liu^{1,2} Haiyan Wu^{3,1} Anqi Liu¹

Abstract

We model human decision-making behaviors in a risk-taking task using inverse reinforcement learning (IRL) for the purposes of understanding real human decision making under risk. To the best of our knowledge, this is the first work applying IRL to reveal the implicit reward function in human risk-taking decision making and to interpret risk-prone and risk-averse decision-making policies. We hypothesize that the state history (e.g. rewards and decisions in previous trials) are related to the human reward function, which leads to risk-averse and risk-prone decisions. We design features that reflect these factors in the reward function of IRL and learn the corresponding weight that is interpretable as the importance of features. The results confirm the sub-optimal risk-related decisions of human-driven by the personalized reward function. In particular, the risk-prone person tends to decide based on the current pump number, while the risk-averse person relies on burst information from the previous trial and the average end status. Our results demonstrate that IRL is an effective tool to model human decision-making behavior, as well as to help interpret the human psychological process in risk decision-making.

1. Introduction

Human’s real-world decision making under reward and risk has been extensively studied with behavioral tasks and neuroimaging techniques in psychology and neuroscience field. Decision-making tasks are increasingly used to evaluate associations between risk-taking propensity and real-world risky behaviors. Unlike the traditional self-report assess-

ment of risk behavior, the Balloon Analogue Risk Taking Task (BART) is a widely used measure of risk-taking tendency in human. The BART task was developed by Lejuez et al. (Lejuez et al., 2002) as a task consists of balloons that have to be pumped up by individuals. In each trial of the task, the balloon is displayed on the screen with two options for the participant. One option is to stop the pump and secure the amount of money of the current balloon (i.e., cash). An alternative option is to take the risk to pump balloon (i.e., pump), resulting in larger balloon (more rewards) or exploded balloon (losing all accumulated rewards). The scores in this task have been shown to correlate with self-reported risk-taking behaviors such as alcohol abuse, drug abuse, gambling etc (Hopko et al., 2006; Lejuez et al., 2003).

To enhance our understanding of the psychological process involved in the BART, efforts have been made to decompose the psychological components in BART with computational models. For example, Wallsten et al. proposed the BART-model proposed to quantify the risk-taking, the speed of learning from experience and the behavior consistency (Wallsten et al., 2005). More recently, Van Ravenzwaaij, Dutilh, and Wagenmakers (2011) proposed a 2-parameter (drift rate and non-decision time) simplification diffusion model with empirical data on the BART (van Ravenzwaaij et al., 2012). However, these models do not directly capture the representation of the real reward function in risk-taking.

In this study, we use the inverse reinforcement learning (IRL) to study the real human decision making during the Balloon Analogue Risk Taking Task. IRL aims to find a reward function that explains the observed behavior (Abbeel & Ng, 2004; Ng et al., 2000). Here, we firstly apply IRL on the polling data from all subjects to estimate the weight of each feature as a baseline. We then split the subjects into two groups, risk-prone and risk-averse group, to examine the differences of reward functions of these two groups. Our model can predict the risk-prone person’s behavior better. We also find risk-prone person tends to decide based on the current pump number, while the risk-averse person relies on burst information from the previous trial and the average end status.

¹California Institute of Technology, Pasadena, USA
²Huntington Medical Research Institutes, Pasadena, USA ³IPCAS, Department of Psychology, University of Chinese Academy of Sciences, Beijing, China. Correspondence to: Anqi Liu <anqiliu@caltech.edu>, Haiyan Wu <wuhy@psych.ac.cn>.

2. Related Work

The risky decisions of human and animal are modulated by a variety of environmental and intrinsic contexts, such as risk propensity and risk perception. One traditional method considering computational factors for modeling risky behavior of human is under the reinforcement learning (RL) framework (Dayan & Niv, 2008). People typically learn the decision-making environments which are characterized by a few key concepts: a state space, a set of actions, and the outcomes from experience. Applying RL makes it possible to uncover the computational mechanisms and neural substrates for reward prediction error (RPE) and valuation (Daw et al., 2011) and show developmental difference (Palminteri et al., 2016).

Reinforcement learning aims to provide an optimal policy to maximize the rewards, which has been applied to predict the optimal decision policy based on the reward function (Kaelbling et al., 1996). However, the reward function in real-world human decision making under risk is usually unknown, for human often tries to avoid the risk states and therefore would not visit all the states. On the other hand, the reward function in our mind might not be the real reward function designed in the task. For instance, the mental status under the risk has to be taken into consideration in human decisions. As a consequence, the real reward function represented in humans might deviate from the true reward function designed in the decision making task. For these reasons, it is still debating in psychology whether humans make rational and optimal decisions under risk.

Inverse reinforcement learning provides a potential tool to estimate the reward function and learn demonstrator's decision policy (Ng et al., 2000). While IRL has been implemented to model the route choice (Ziebart et al., 2008), football players' strategies (Le et al., 2017), or robot navigation (Kretschmar et al., 2016), to the best of our knowledge, the IRL framework has not been used in psychology and decision-making neuroscience yet. To recover a reward function of human under risk and learning policies from risk-prone and risk-averse people, IRL predicts actions in states which even has not been observed/occurred.

3. Task and Behavioral Data

3.1. Balloon Analogue Risk Task

We used the Balloon Analogue Risk Task (BART), to assess the real-world human decision-making behavior through the conceptual frame of balancing the potential for reward versus risk (Lejuez et al., 2002). The instructions were the same as the ones described by Lejuez et al. (2002). Participants were asked to earn as many points as possible.

The experimental interface is in (Figure 1a). Subjects can

choose to pump the balloon (press “v”) or stop pump (press “n”). As a result of choosing to pump, the balloon may burst or may grow in size. If the balloon burst, the participant loses all the accumulated points in the current trial, otherwise the participant earns 10 more points and confronts with a next-round choice: cash out or pump. A new trial begins when the participant stops pump (cashes out) or the balloon bursts. The task consists of 1 practice trial and 30 formal task trials.

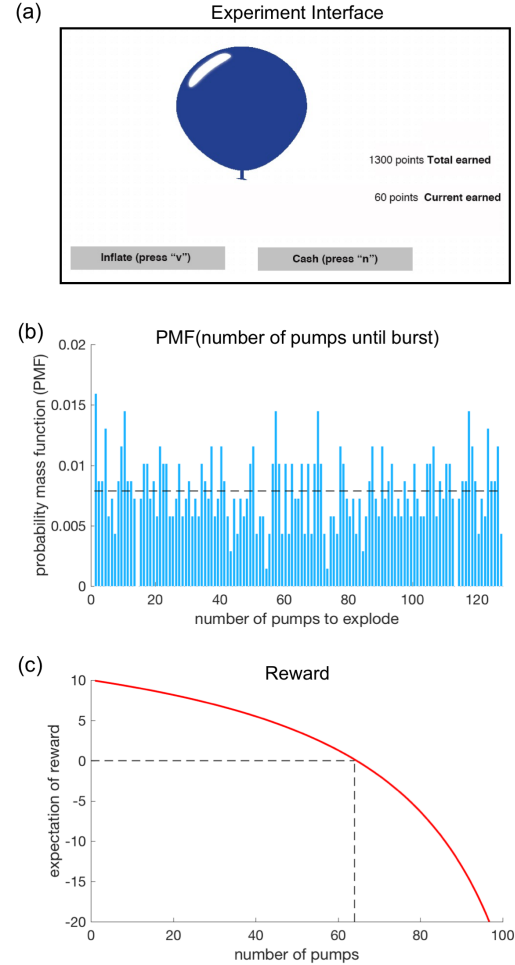


Figure 1. Experiment settings. (a) The experimental interface; (b) the probability mass function for the number of pumps until balloon explosion; (c) the true expectation of reward.

In each trial, the breakpoints are uniformly distributed from 1 to 128 (Figure 1b). It means the probability that a balloon would explode on the first pump is $1/128$. If the balloon did not blow on the first pump, the probability of a balloon burst on the second pump is $1/127$, and so on. Based on this setting, the true reward function at i^{th} pump is

$$R(i) = \begin{cases} -10(i-1), & \text{if exploded} \\ +10, & \text{otherwise} \end{cases} \quad (1)$$

Let $X_i = 1$ denote the event that the balloon burst at i^{th} trial. We have $P(X_i = 1 | X_{i-1} = 0) = \frac{1}{129-i}$, and $P(X_i = 0 | X_{i-1} = 0) = \frac{128-i}{129-i}$. Therefore, the expected reward at i^{th} pump is

$$E(i) = \frac{1290 - 20i}{129 - i} \quad (2)$$

The expected reward starts from 10 at first trial, and decreases with the increasing number of trials. The best strategy for this task is to cash out at 64th pump, when the reward reduces to 0 (Figure 1c).

3.2. Subjects and Human Behaviors

We collect experimental data with extensive coverage and report simple statistics. There are 23 university students (15 females with $M_{age} = 23.2$, $SD_{age} = 3.22$) from Beijing, aged 18 to 25, participating in the experiment and the monetary reward varies upon the task performance.

The behavioral results indicate a mean of pump number as 29.04, which is much smaller than a rational player would pump. The participants stopped pumping and cashed out before the balloon exploded in 74.14% of trials. The mean reward points in each test earned by the participants were 205.92. The histogram of the number of pumps for all the trials are in Figure 2a. It seems that the participants did not learn the true reward function during the task, for they stopped to pump before the expected reward decrease to 0. As a consequence, the more pumps, the more rewards. The final rewards are significantly correlated with the number of pumps (Figure 2b).

4. Algorithm

We model the human behavior in this risk-taking task using Markov Decision Process. We argue that the Markovian property of transition and action holds in this particular task with certain assumptions when setting up state and action space. Specifically, we set states as two dimensional vector $[i, s]$, where i represent the i^{th} pump and s represents the current status of the balloon as 0,1, meaning burst or not. Therefore, we can assume the next state, $[i + 1, s]$, meaning the $(i + 1)^{\text{th}}$ pump and the corresponding status of the balloon only depend on the previous status and current action. The transition relation shown in Eq. 1 also reflects this relation. The policy $P(a_t | s_1, s_2, \dots, s_t)$ can be also regarded as equivalent to $P(a_t | s_t)$ since the human subject always know how many inflation of the balloon has been executed. Our action space is binary with 0 indicating “stop” while 1 indicating “pump”.

In BART, human subjects resemble agents in a reinforcement learning task. The goal of the human subjects is to collect the largest reward from a single trial without knowing the probability of balloon burst. We can regard this

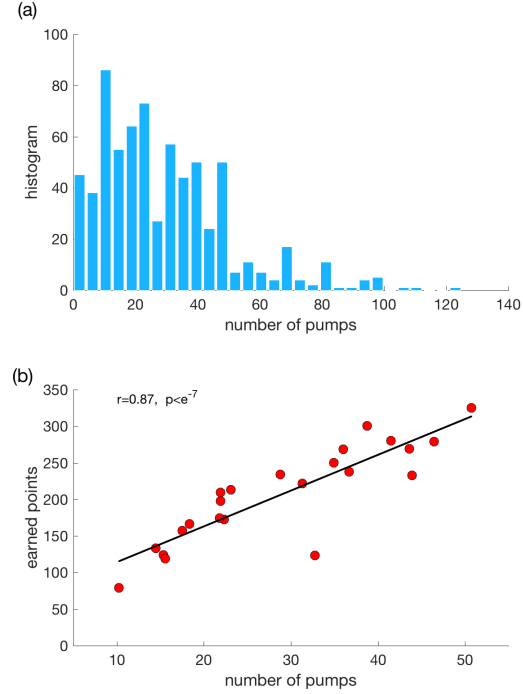


Figure 2. Behavioral data. (a) the histogram of number of pumps; (b) reward as a function of number of pumps.

process as human learning of an unknown reward function. Because of the irrational nature of human and their different risk-prone levels, the true reward function (burst probability) is never learned. Instead, the human decision making in this process provides data that demonstrate possible underlying influence of historical trials to current ones. In other words, the data is not independent with each other between different trials. However, we construct the state features in a way that reflects information from historical trials and we make the assumption that state features is independent among trials. Table 1 shows our state features. Note that by “stop” we mean ending the trial without balloon burst while by “end” we mean ending the trial with whatever status.

4.1. Maximum Entropy IRL

Traditional Markov decision process (MDP) assumes that we know the reward function of agents. However, in the real world, humans usually make decisions without knowing the reward function. In imitation learning, we observe demonstrations of expert trajectories. IRL reduces learning to a problem of recovering utility function that makes the behavior induced by a near-optimal policy closely mimic the demonstrated behavior (Abbeel & Ng, 2004; Ziebart et al., 2008). More specifically, IRL assumes reward function, R , has a linear relation with specific features that are

No.	Value	Meaning
1	Integer	Number of times being in this state
2	Binary	Whether this state was burst in the previous trial
3	Binary	Whether this state was stopped in the previous trial
4	Binary	Whether state was burst in the 2nd previous trial
5	Binary	Whether this state was stopped in the 2nd previous trial
6	Binary	Whether state was burst in the 3rd previous trial
7	Binary	Whether this state was stopped in the 3rd previous trial
8	Binary	Whether is the average burst status
9	Binary	Whether is the average stop status
10	Binary	Whether is the average end status
11	Integer	Number of steps in the current trial

Table 1. States Features with Value and Meaning.

associated with states representation, and the reward value of a trajectory, $f(\zeta)$, is simply the sum of state rewards $f(s_t)$:

$$R(f(\zeta)) = \theta^T f(\zeta) = \sum_{s_t \in \zeta} \theta^T f(s_t), \quad (3)$$

where θ is the reward weights. A deep neural network can be also used for representing the relation between features and reward function.

In the most prevalent IRL model, which is the Max-EntIRL, the model finds the policy that maximizes the entropy of trajectories: $\theta = \arg \max_{\theta} L(\theta) = \arg \max_{\theta} \sum_{\text{examples}} \log P(\tilde{\zeta}|\theta, T)$, where a convenient form of policy distribution is derived in inference, possessing the form:

$$P(a_t|s_t) \propto \exp(\theta^T f(s_t)), \quad (4)$$

where θ is the learning parameters (reward weights) and $f(s_t)$ is the state features. We follow the Algorithm 1 in (Ziebart et al., 2008) for inference of the state frequencies. The learning objective is the approximate likelihood of the samples collected from human experiments. Therefore, the maximum entropy IRL algorithm finds reward weights that makes human behavior appear near optimal under the learned reward function. It resolves the ambiguity from the real-world actions and resulting in a single stochastic policy. The gradient is the difference between expected features frequencies and empirical ones. We can use gradient descent to optimize for parameter θ .

5. Experiments

We conduct two sets of experiments to demonstrate how IRL can recover reward function underlying human decision-making behavior and help neuroscientist and psychologist analyze how historical observations and risk-taking habits can affect future decision making. We first use all the

	All subjects	Risk-Prone	Risk-Averse
Mean (# pumps)	29.04	38.78	18.42
Mean (RT/trial)	13.77	16.20	11.11
Mean (RT/pump)	0.53	0.46	0.60

Table 2. Number of bumps and reaction time for different groups.

merged data together and regard everyone as the same type of risk-taking pattern and learn the reward function. We then compare with the result we obtain in group-specific experiments where we split subjects into two groups: risk-prone group and risk-averse group, based on the median pump numbers in pooling data. Table 2 shows the different behavior patterns of these two groups of people. Risk-prone group tends to have longer trial length, meaning they pump balloon much more times and have longer total time. But they have a shorter reaction time per pump, meaning they make a decision quicker. However, the risk-averse group tends to have shorter trial length, shorter total time, but longer reaction time for making decisions. For each group, we use half of the data for training and the other half for testing.

5.1. Reward Function Analysis

The weights of features for all subjects, risk-averse subjects and risk-prone subjects, are in Figure 3. Overall, people show higher weight of “Number of times being in this state” (feature 1), indicating a general effect of past individual experience on the current risky decision making (Bechara et al., 2000; Hertwig & Erev, 2009). Interestingly, risk-prone subjects tend to weigh more on “Number of steps in the current trail” (feature 11), with less weight on “Number of times being in this state” (feature 1), “Whether burst in this state in the previous trial” (feature 2) and “Whether is the average end status” (feature 10), than risk-averse subjects. Our findings suggest that the history of “Number of times being in this state” serves as a general cue of contextual analysis when making a risky decision in outcome anticipation. Irrationally, the current number of pumps is more influential in risk-prone subjects than risk-averse individuals. Risk-averse subjects rely more on the previously learned information (from both the previous trial and the average end status). Thus, our results support the idea that people are learning to be risk-averse, with consideration of normal human as adaptive decision makers (Denrell, 2007; March, 1996).

5.2. Predictive Performance Analysis

We test the learned model on testing data and demonstrate the average likelihood of the test trajectories in Table 3. The likelihood of data sequence for a trial η is as following using the same model:

$$P(\eta) = P(s_0)P(a_0|s_0)P(s_1|a_0, s_0)P(a_1|s_1) \dots \quad (5)$$

	All subjects	Risk-Prone	Risk-Averse
Test LLD	-1.36	-0.89	-1.4

Table 3. Test Log Likelihood for Different Groups. The higher log likelihood, the better prediction of IRL is.

The predictive results (Table 3) indicate that the decisions from the risk-prone group are easier to predict than the risk-averse group. Convergetly, risk-prone individuals also show less hesitation (shorter reaction time per pump) when deciding whether to pump the balloon (Table 2), compared to the risk-averse people.

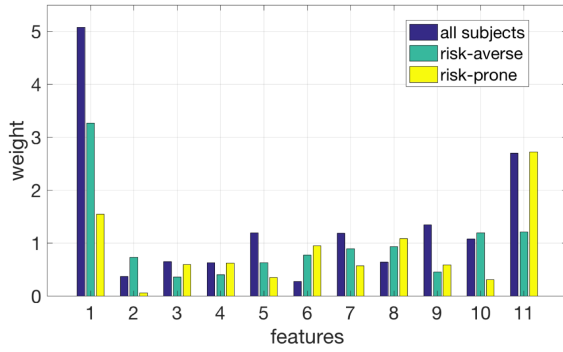


Figure 3. Weights of features from IRL model. The meaning for the 11 features are listed in Table 1

6. Discussions

This work aims to model human reward function in the BART under the framework of IRL. With the construction of IRL models for all subjects, risk-prone subjects and risk-averse subjects respectively, we reveal the weights for different features for each group.

6.1. Psychological Perspective

Our results (Figure 2) reveal a general risk averse behavioral pattern as most people did not reach the optimal pump point (64 pumps in our task). Such finding is consistent with existing results showing risk aversion in normal population (Fullenkamp et al., 2003; Holt & Laury, 2002; Loewenstein et al., 2001). It implies that human subjects did not search the entire state space and therefore did not learn the true reward function (Figure 3c).

Other alternative models have been used to learn and predict the human decision makings (Angela & Cohen, 2009; Guo & Angela, 2018; Ryal et al., 2018), such as dynamic belief model which requires knowledge or assumption about the models, or reinforcement learning which assume humans know the reward function. The key concept of IRL here is

that humans make decisions to optimize an unknown reward function which is linear in the features. Our results quantified the weight of features in the reward function represented in human’s decision, and show that risk-prone person tends to decide based on the current bump number, while the risk-averse person rely on burst information from previous trial and the average end status (Figure 3). Moreover, we found the IRL can predict the risk-prone subject’s decisions better than the risk-averse subject’s decisions (Table 3), suggesting the reward functions in risk-prone subjects are highly representative. This might result in high impulsivity and less variety in their decisions under risk, which has been shown in the literature (Adriani et al., 2009; Kashdan et al., 2009).

6.2. Machine Learning Perspective

Inverse Reinforcement Learning is a prevalent tool for imitation learning and learning from demonstration (Ng et al., 2000), which has been applied successfully to robotics, self-driving car, and assistive technology (Zhifei & Meng Joo, 2012). This work reveals the potential advantage for using IRL to study human decision making in psychological experiment with limited samples and large variability. Comparing with simpler models like behavior cloning (Ross et al., 2011), IRL can recover reward function even when only limited states are visited in data. But it also induces a slightly larger computational cost. The comparison between different machine learning models is an interesting research topic that we leave to future work. Importantly, this work is also an exploration towards interpretable machine learning when applying machine learning to fields that require high-level model interpretability. When we analyze predictive performance of the model, we do not seek the best model for highest accuracy but the best one providing the most reasonable explanation of data. Finally, the limited data of psychological experiments can heavily constrain the model choice for machine learning. This drives new developments of domain adaption and active learning techniques in machine learning community.

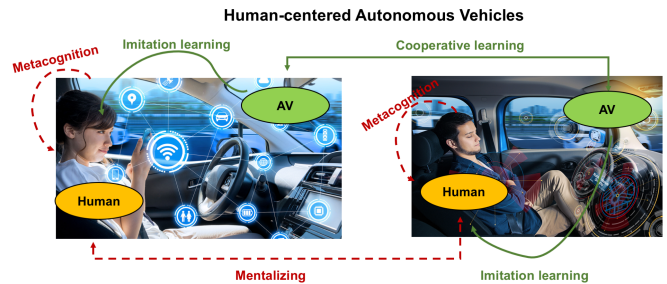


Figure 4. Learning human’s cost function in autonomous vehicles

6.3. Future Directions

In the future, we will improve the model by integrating the individual personalities from questionnaires (e.g. sensation seeking, and impulsivity) (Bornovalova et al., 2009; Donohew et al., 2000; Lauriola et al., 2014) and other contextual factors (such as risky cue learning) (Cohen et al., 2009) into our feature space. Moreover, we only tested 30 trials for each subject, which is not enough to train a model for each subject, and personalize the reward function. We will run more trials per subject, train IRL model and other computational models in the future. It will allow us to examine the cross-subject variation in model parameters. The implications of the IRL model in risk decisions can also benefit both understanding risk-taking behavior at an individual level or interactive mind level (e.g. metacognition or mentalizing process for the mental state inference within/between individuals), which can guide the design of human-centered autonomous vehicles (Figure 4).

Acknowledgements

QL gets supports from James Boswell Fellowship and FWO postdoctoral fellowship (12P6719N LV). HW gets supports from the National Natural Science Foundation of China (grant number: U1736125).

References

- Abbeel, P. and Ng, A. Y. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, pp. 1. ACM, 2004.
- Adriani, W., Boyer, F., Gioiosa, L., Macri, S., Dreyer, J.-L., and Laviola, G. Increased impulsive behavior and risk proneness following lentivirus-mediated dopamine transporter over-expression in rats' nucleus accumbens. *Neuroscience*, 159(1):47–58, 2009.
- Angela, J. Y. and Cohen, J. D. Sequential effects: superstition or rational behavior? In *Advances in neural information processing systems*, pp. 1873–1880, 2009.
- Bechara, A., Damasio, H., and Damasio, A. R. Emotion, decision making and the orbitofrontal cortex. *Cerebral cortex*, 10(3):295–307, 2000.
- Bornovalova, M. A., Cashman-Rolls, A., O'donnell, J. M., Ettinger, K., Richards, J. B., Lejuez, C., et al. Risk taking differences on a behavioral task as a function of potential reward/loss magnitude and individual differences in impulsivity and sensation seeking. *Pharmacology Biochemistry and Behavior*, 93(3):258–262, 2009.
- Cohen, M. X., Axmacher, N., Lenartz, D., Elger, C. E., Sturm, V., and Schlaepfer, T. E. Neuroelectric signatures of reward learning and decision-making in the human nucleus accumbens. *Neuropsychopharmacology*, 34(7):1649, 2009.
- Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., and Dolan, R. J. Model-based influences on humans' choices and striatal prediction errors. *Neuron*, 69(6):1204–1215, 2011.
- Dayan, P. and Niv, Y. Reinforcement learning: the good, the bad and the ugly. *Current opinion in neurobiology*, 18(2):185–196, 2008.
- Denrell, J. Adaptive learning and risk taking. *Psychological review*, 114(1):177, 2007.
- Donohew, L., Zimmerman, R., Cupp, P. S., Novak, S., Colon, S., and Abell, R. Sensation seeking, impulsive decision-making, and risky sex: Implications for risk-taking and design of interventions. *Personality and individual differences*, 28(6):1079–1091, 2000.
- Fullenkamp, C., Tenorio, R., and Battalio, R. Assessing individual risk attitudes using field data from lottery games. *Review of Economics and statistics*, 85(1):218–226, 2003.
- Guo, D. and Angela, J. Y. Why so gloomy? a bayesian explanation of human pessimism bias in the multi-armed bandit task. In *Advances in Neural Information Processing Systems*, pp. 5176–5185, 2018.
- Hertwig, R. and Erev, I. The description–experience gap in risky choice. *Trends in cognitive sciences*, 13(12):517–523, 2009.
- Holt, C. A. and Laury, S. K. Risk aversion and incentive effects. *American economic review*, 92(5):1644–1655, 2002.
- Hopko, D. R., Lejuez, C., Daughters, S. B., Aklin, W. M., Osborne, A., Simmons, B. L., and Strong, D. R. Construct validity of the balloon analogue risk task (bart): Relationship with mdma use by inner-city drug users in residential treatment. *Journal of Psychopathology and Behavioral Assessment*, 28(2):95–101, 2006.
- Kaelbling, L. P., Littman, M. L., and Moore, A. W. Reinforcement learning: A survey. *Journal of artificial intelligence research*, 4:237–285, 1996.
- Kashdan, T. B., McKnight, P. E., Richey, J. A., and Hofmann, S. G. When social anxiety disorder co-exists with risk-prone, approach behavior: Investigating a neglected, meaningful subset of people in the national comorbidity survey-replication. *Behaviour Research and Therapy*, 47(7):559–568, 2009.

- Kretzschmar, H., Spies, M., Sprunk, C., and Burgard, W. Socially compliant mobile robot navigation via inverse reinforcement learning. *The International Journal of Robotics Research*, 35(11):1289–1307, 2016.
- Lauriola, M., Panno, A., Levin, I. P., and Lejuez, C. W. Individual differences in risky decision making: A meta-analysis of sensation seeking and impulsivity with the balloon analogue risk task. *Journal of Behavioral Decision Making*, 27(1):20–36, 2014.
- Le, H. M., Yue, Y., Carr, P., and Lucey, P. Coordinated multi-agent imitation learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1995–2003. JMLR. org, 2017.
- Lejuez, C. W., Read, J. P., Kahler, C. W., Richards, J. B., Ramsey, S. E., Stuart, G. L., Strong, D. R., and Brown, R. A. Evaluation of a behavioral measure of risk taking: the balloon analogue risk task (bart). *Journal of Experimental Psychology: Applied*, 8(2):75, 2002.
- Lejuez, C. W., Aklin, W. M., Zvolensky, M. J., and Pedulla, C. M. Evaluation of the balloon analogue risk task (bart) as a predictor of adolescent real-world risk-taking behaviours. *Journal of adolescence*, 26(4):475–479, 2003.
- Loewenstein, G. F., Weber, E. U., Hsee, C. K., and Welch, N. Risk as feelings. *Psychological bulletin*, 127(2):267, 2001.
- March, J. G. Learning to be risk averse. *Psychological review*, 103(2):309, 1996.
- Ng, A. Y., Russell, S. J., et al. Algorithms for inverse reinforcement learning. In *Icml*, volume 1, pp. 2, 2000.
- Palminteri, S., Kilford, E. J., Coricelli, G., and Blakemore, S.-J. The computational development of reinforcement learning during adolescence. *PLoS computational biology*, 12(6):e1004953, 2016.
- Ross, S., Gordon, G., and Bagnell, D. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 627–635, 2011.
- Ryali, C., Reddy, G., and Angela, J. Y. Demystifying excessively volatile human learning: A bayesian persistent prior and a neural approximation. In *Advances in Neural Information Processing Systems*, pp. 2781–2790, 2018.
- van Ravenzwaaij, D., Dutilh, G., and Wagenmakers, E.-J. A diffusion model decomposition of the effects of alcohol on perceptual decision making. *Psychopharmacology*, 219(4):1017–1025, 2012.
- Wallsten, T. S., Pleskac, T. J., and Lejuez, C. W. Modeling behavior in a clinically diagnostic sequential risk-taking task. *Psychological review*, 112(4):862, 2005.
- Zhifei, S. and Meng Joo, E. A survey of inverse reinforcement learning techniques. *International Journal of Intelligent Computing and Cybernetics*, 5(3):293–311, 2012.
- Ziebart, B. D., Maas, A. L., Bagnell, J. A., and Dey, A. K. Maximum entropy inverse reinforcement learning. In *Aaai*, volume 8, pp. 1433–1438. Chicago, IL, USA, 2008.