Exploring and exploiting in social contexts

Erica J. Yoon*, **Kyle MacDonald***, **Mika Asaba**, **Hyowon Gweon**, and **Michael C. Frank**
{ejyoon, kylem4, masaba, hyo, mcfrank} @stanford.edu
Department of Psychology, Stanford University
*These authors contributed equally to this work.

Abstract

...

Keywords: Learning; social context; information gain; OED; self-presentation; goal tradeoff

Introduction

Imagine you had to cook a meal for yourself. Would you want to follow a recipe you have tried before, or venture a new recipe? The familiar recipe would ensure a good meal, but the new recipe might turn out to be even better, despite some risk that it might not. How would your decision change if the meal was for a date? Or for a charitable cooking teacher?

We often face such "exploration-exploitation dilemma" (Sutton & Barto, 1998), in which we have to choose between an overt, readily accessible reward based on what we already know (*exploitation*), and a search for discovery of knowledge that we do not yet have (*exploration*). How do we decide whether to explore or to exploit? We may think about our goals associated with each of the options, and how much we want to achieve each goal relative to others: Do I want to learn a new way of cooking, or secure a decent meal? We may also consider contextual factors that affect how much we prioritize each goal: For example, am I by myself, or is there a person whom I want to impress with a delicious meal? An agent can have a mixture of exploration and exploitation goals that conflict with each other. In this work, we seek to explain how people may behave given an exploration-exploitation dilemma, by (1) formalizing the tradeoff between an agent's goals for exploration versus exploitation, and (2) considering ways in which the social context (e.g. the presence of another person) may play a role in how much value is placed on each goal instantiation.

One prominent case study that shows people's exploratory behaviors is *active learning* (also known as "self-directed" learning), or active construction of their own learning experience by making informative queries or exploring unknown contingencies (Gureckis & Markant, 2012; Nelson, 2005; Sutton & Barto, 1998). Active learning research has focused on whether the context of self-directed learning presents advantages in acquiring and processing more information for the learner (see Gureckis & Markant (2012)). We are not always alone in a learning context, however; often there are teachers, peer learners, or other people around that might affect our learning process and outcome. Indeed, when adults and children receive pedagogical instructions from a teacher or view actions of another learner, they show different rate of learning or exploratory information-seeking behaviors (Bonawitz et al., 2011; Markant & Gureckis, 2014).

But what role does the *social* factor play on self-directed learning? For example, can a learner's action choices differ based on the presence of an observer, regardless of whether the observer provides relevant information for learning? Instead of a learning goal, a learner may prioritize reputation management: looking more attractive, knowledgeable, or competent to an observer. In such case, the learner will infer what the observer thinks about the learner based on what the learner did (cook a meal) and its outcome (a delicious meal). Then the learner will choose between exploration and exploitation that will likely maximize favorable impression of the learner to the observer ('That meal was delicious, so he must be a good cook!').

The link between self-directed learning and self-presentation to others is unclear. Self-presentational concerns affect people's prosocial and cheating behaviors (FIXME: cite). Socially elicited goals to prioritize presenting oneself as competent changes people's task choices and responses to difficulty (Dweck & Leggett, 1988; Elliott & Dweck, 1988). But no work has looked at how the social pressure based on another person's presence influences people's strategies to act in a self-directed learning environment.

We present a computational model formalizing the exploration-exploitation tradeoff in an active learning environment with social pressures. The model shows that the agent's goal preference may change depending on the social context, in the presence or absence of another person. We then look at empirical judgments for a simplistic representation of our model across different goals and social contexts (i.e. presence of another person), which are consistent with the predictions of the model.

Computational model

To start to examine people's exploration-exploitation tradeoff, we situate the model and paradigm in a simplistic learning environment. The learner in our model is to act on a toy, and can choose between two kinds of actions that will each lead to one outcome (new discovery) or the other (overt reward). The learner's action rests on his goals to explore versus exploit, and is determined in part by the presence or absence of another person he cares about (i.e. his boss)[1].

A key assumption underlying inferences in recent Bayesian models of human social cognition is that people act approximately optimally given a utility function (e.g. Goodman & Frank, 2016; Jara-Ettinger, Gweon, Schulz, & Tenenbaum, 2016). Our model adopts the same utility-theoretic approach,

[1]From here on, we use a male pronoun for Bob, the learner, and female pronoun for Ann, the boss and observer.

and assumes an approximately optimal agent, who reasons about the utility function that represents a combination of multiple goals. In a recent model of polite language production (Yoon, Tessler, Goodman, & Frank, 2017), the utility function comprised a weighted combination of multiple utilities (goals) considered by the speaker, reflecting a principled tradeoff between different communicative goals (e.g. to be informative, to be kind, and to appear to be a helpful speaker). We use a similarly structured utility function that reflects different goals that a learner has in a social learning context. Specifically, we model how a person may make a decision to act based on his desire to learn how a toy works (*learning utility*), to make the toy operate and perform a given function (*performance utility*), or to present himself as a competent individual who knows how to make the toy work (*presentational utility*; see the model diagram in Figure 1).

First, the *learning utility* symbolizes the goal to learn new information, which in our paradigm specifically is associated with figuring out how a given toy works. The learning utility is formally represented by an OED model (Lindley, 1956; "Optimal Experiment Design"; Nelson, 2005), which quantifies the *expected utility* of different information seeking actions. Here we follow the mathematical details of the OED approach as outlined in Coenen, Nelson, & Gureckis (2017) that was implemented in our model. The set of queries, each realized through taking an action, is defined as $Q_1, Q_2, ..., Q_n = Q$. The expected utility of each query ($EU(Q)$) is a function of two factors: (1) the probability of obtaining a specific answer $P(a)$ weighted by (2) the usefulness of that answer for achieving the learning goal $U(a)$.

$$EU(Q) = \sum_{a \in q} P(a)U(a)$$

There are a variety of ways to define the usefulness function to score each answer (for a detailed analysis of different approaches, see Nelson (2005)). One standard method is to use *information gain*, which is defined as the change in the learner's overall uncertainty (difference in entropy) before and after receiving an answer. This information gain is then the usefulness of the answer to the query, and thus is equal to the learning utility:

$$U_{learning} = U(a) = ent(H) - ent(H|a)$$

where $ent(H)$ is defined using Shannon entropy[2]. MacKay (2003), which provides a measure of the overall amount of uncertainty in the learner's beliefs about the candidate hypotheses.

$$ent(H) = -\sum_{a \in A} P(h) log_2 P(h)$$

[2]Shannon entropy is a measure of unpredictability or amount of uncertainty in the learner's probability distribution over hypotheses. Intuitively, higher entropy distributions are more uncertain and harder to predict. For example, if the learner believes that all hypotheses are equally likely, then they are in a state of high uncertainty/entropy. In contrast, if the learner firmly believes in one hypothesis, then uncertainty/entropy is low.

The conditional entropy computation is the same, but takes into account the change in the learner's beliefs after seeing an answer.

$$ent(H|a) = -\sum_{h \in H} P(h|a) log P(h|a)$$

To calculate the change in the learner's belief in a hypothesis $P(h|a)$, we use Bayes rule.

$$P(h|a) = \frac{P(h)P(a|h)}{P(a)}$$

The learner performs the expected utility computation for each query in the set of possible queries and picks the one that maximizes utility. In practice, the learner considers each possible answer, scores the answer with the usefulness function, and weights the score using the probability of getting that answer. In our paradigm, a learner thinking about the learning utility considers acting on the toy one way over another, and computes how informative a given answer should be in reducing uncertainty about how the toy works.

Second, the *performance utility* is the utility of successfully making the toy operate. Specifically within our current paradigm, the performance utility is the expected utility of music playing (m) given the learner's action a.

$$U_{performance} = P_L(m|a)$$

Thus, performance utility is maximized by taking an action that is most likely to make the toy "go" and play music, which is the operation of interest.

When there is no observer present, the learner considers the tradeoff between the learning utility and performance utility, and he determines his action based on a weighted combination of the two utilities:

$$U(a; \phi; obs = no) = \phi \cdot U_{learning} + (1 - \phi) \cdot U_{performance},$$

where ϕ is a mixture parameter governing the extent to which the learner prioritizes information gain over making the toy play music.

When there is another person present to observe the learner's action, this observer O reasons about the competence c of the learner L which is equal to whether the learner was able to make the toy work.

$$P_O(c) \propto P_L(m|a)$$

The learner thinks about how the observer infers the learner's competence, and his *presentational* utility is based on maximizing the apparent competence inferred by the observer.

$$U_{presentation} = P_O(c)$$

Thus, when there is an observer present, the learner considers the tradeoff between the learning utility and presentational utility:

$$U(m;a;\phi;obs=yes) = \phi \cdot U_{learning} + (1-\phi) \cdot U_{presentational}$$

Based on the utility functions above, the learner (L) chooses his action a approximately optimally (as per optimality parameter λ) given his goal weight and observer presence.

$$P_L(a|\phi,obs) \propto \exp(\lambda \cdot \mathbb{E}[U(a;\phi;obs)])$$

Experiment

Method

Participants FIXME participants with IP addresses in the United States were recruited on Amazon's Mechanical Turk. We excluded FIXME partcipants who failed to answer manipulation check questions correctly (See Procedure section for details on the manipulation check questions), and thus the remaining FIXME participants were included in our final analysis.

Stimuli and Design We presented three different toys that look very similar but each work in different ways, and provided instructions for them. The "ButtonMusic" toy instructions were: *"Press the button on the right to play music. Pull the handle on the left to turn on the light. Doing both produces both effects."* The "HandleMusic" toy instuctions were: *"Pull the handle on the left to play music. Press the button on the right to turn on the light. Doing both produces both effects."* The "BothMusicLight" toy instuctions were: *"Pull the handle on the left AND press the button on the right to turn on the light and play music at the same time. The button press or handle pull on its own doesn't produce any effect."* Each toy had a label showing its name.

We presented a story to the participants that motivated a goal the participants must achieve by acting on a given toy. Importantly, the given toy was missing its label, such that partcipants could not know whether the toy was a ButtonMusic, HandleMusic, or BothMusicLight toy. For all the participants, we asked them to imagine that they were a children's toy developer and that one day, their boss approached them and said: "That must be one of the new toys that you've been working on. But it looks like you forgot to put on the label! Can you figure out whether this toy is a ButtonMusic toy, HandleMusic toy, or BothMusicLight toy?" (*learning* condition); "That must be one of the new toys that you've been working on. I want to hear the music it plays." (*performance* condition); or "That must be one of the new toys that you've been working on. How does it work?" followed by the prompt "... [Imagine] you only had one chance to impress your boss and show that you're competent ..." (*presentation* condition). We asked participants to select an action they would like to try out on the toy in order to accomplish the specified goal, out of three possible actions: to "press the button", "pull the handle", or "press the button and pull the handle." We randomly assigned each participant to one of the three goal conditions, and randomized the order of actions to choose from.

Procedure Participants were first introduced to the task with a picture of a toy with labels on its parts. Then they read instructions for each of the three toys, after which they were asked what they would do to make the toy operate as manipulation check (e.g. "How would you make the toy play music?"). We asked participants to rate prior likelihood that an unknown toy is a ButtonMusic toy, HandleMusic toy, or BothMusicLight toy, to use as priors for our model. Participants then read a scenario for one of the three goal conditions, followed by the question: "If you only had one chance to try a SINGLE action to [goal], which action would you want to take? You will get a 10 cent bonus after submitting the HIT if you [goal]." After selecting one of three possible actions to perform on the toy and seeing that the toy successfully played music, participants were asked again to rate the likelihood that the unlabeled toy was each of the three possible toys. The experiment can be viewed at `https://langcog.stanford.edu/expts/EJY/soc-info/goal_actions_ver2/soc_info_goals.html`.

Results

Action selections

Response times

Entropy change

Discussion

Acknowledgements

This work was supported by NSERC postgraduate doctoral scholarship PGSD3-454094-2014 to EJY ... FIXME.

References

Bonawitz, E., Shafto, P., Gweon, H., Goodman, N. D., Spelke, E., & Schulz, L. (2011). The double-edged sword of pedagogy: Instruction limits spontaneous exploration and discovery. *Cognition*, *120*(3), 322–330.

Coenen, A., Nelson, J. D., & Gureckis, T. M. (2017). Asking the right questions about human inquiry.

Dweck, C. S., & Leggett, E. L. (1988). A social-cognitive approach to motivation and personality. *Psychological Review*, *95*(2), 256.

Elliott, E. S., & Dweck, C. S. (1988). Goals: An approach to motivation and achievement. *Journal of Personality and Social Psychology*, *54*(1), 5.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, *20*(11), 818–829.

Gureckis, T. M., & Markant, D. B. (2012). Self-directed learning: A cognitive and computational perspective. *Perspectives on Psychological Science*, *7*(5), 464–481.

Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The nave utility calculus: Computational principles underlying commonsense psychology. *Trends in Cognitive Sciences*, *20*(8), 589–604.

Lindley, D. V. (1956). On a measure of the information

provided by an experiment. *The Annals of Mathematical Statistics*, 986–1005.

MacKay, D. J. (2003). *Information theory, inference and learning algorithms*. Cambridge university press.

Markant, D. B., & Gureckis, T. M. (2014). Is it better to select or to receive? Learning via active and passive hypothesis testing. *Journal of Experimental Psychology: General*, *143*(1), 94.

Nelson, J. D. (2005). Finding useful questions: On bayesian diagnosticity, probability, impact, and information gain. *Psychological Review*, *112*(4).

Sutton, R. S., & Barto, A. G. (1998). *Introduction to reinforcement learning* (Vol. 135). MIT Press Cambridge.

Yoon, E. J., Tessler, M. H., Goodman, N. D., & Frank, M. C. (2017). "I won't lie, it wasn't amazing": Modeling polite indirect speech. In *Proceedings of the thirty-ninth annual conference of the Cognitive Science Society*.

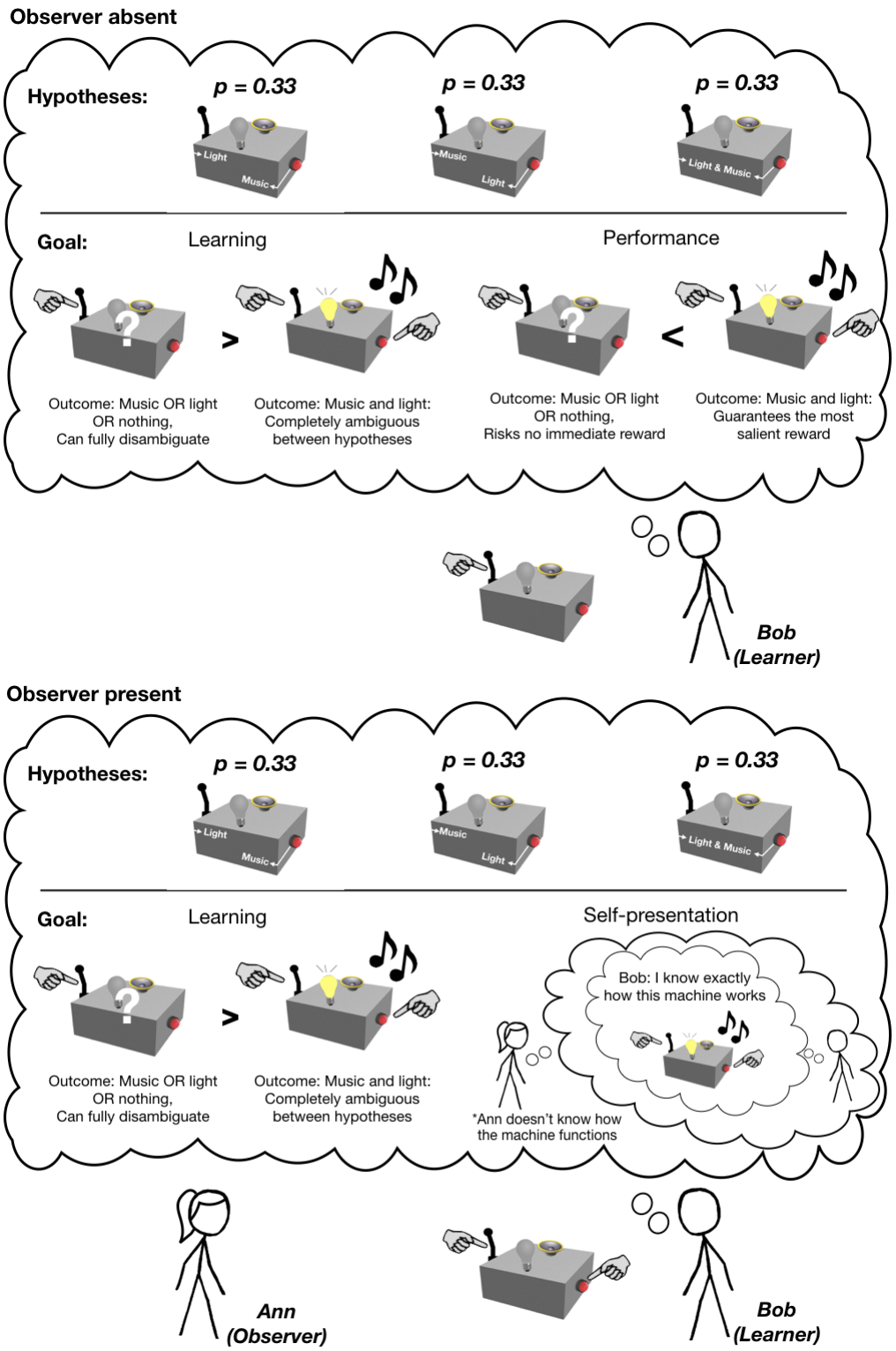

Figure 1: Diagram of the computational model: The learner considers possible hypotheses and his contextual goals. When an observer is absent, he considers his learning goal (to maximize information gain) and performance goal (e.g. to play music) and decides on an action. When an observer is present, his decision for an action is based on his learning goal vs. presentational goal (to have the observer infer his competence).