Intentional commitment through an internalized theory of mind: Acting in the eyes of an imagined observer

Shaozhe Cheng[1] shaozhecheng@zju.edu.cn
Minglu Zhao[3] minglu.zhao@ucla.edu
Jingyin Zhu[1] zhujy@zju.edu.cn
Jifan Zhou[1] jifanzhou@zju.edu.cn
Mowei Shen[1] mwshen@zju.edu.cn
Tao Gao[2,3] tao.gao@stat.ucla.edu

[1] Department of Psychology and Behavioral Sciences, Zhejiang University
[2] Department of Communication, UCLA [3] Department of Statistics, UCLA

Abstract

The ancient Greek hero Ulysses chose to bind himself to resist the temptation of Sirens, highlighting the fact that humans may voluntarily sacrifice their freedom of choice to achieve committed goals. In this work, we propose a computational model for such commitment under the framework of Bayesian Theory of Mind. The model is based on the idea that even when alone, humans act to better demonstrate their intentions to an imagined third-party observer (ITO) censoring their actions. Our model successfully captures the Ulysses-constraint of freedom, as the freedom confuses the ITO's inference of their intention. We further show that, trajectories generated both by human actors and actors modeled with ITO censorship are easy to interpret both in the eyes of an actual human observer and an ITO. The results demonstrate that under conflicting desires, humans achieve commitment by spontaneously censoring their actions with an internalized theory of mind.

Keywords: conflicting desires; intention; commitment; metacognition; internalized theory of mind

Introduction

Humans are purposeful agents who act to fulfill desires. Yet, human minds are often full of desires incompatible with each other. The ancient Greek hero Ulysses wanted to hear the Siren's song, yet he was also eager to safely return to homeland without being seduced by the Siren — in the end he voluntarily surrendered his freedom and bound himself to a mast to resist the temptation. We mundane people also constantly experience this contradiction within ourselves: People suffer from conflicting desires as if they have multiple selves; part of you wants longevity while another part is addicted to alcohol (Schelling, 1984). This multiple-selves dilemma has long been discussed in philosophy (Elster, 1987) and psychology (Freud, 1923), while the cognitive and computational mechanisms of why humans can generally act coherently under conflicting desires are still unclear.

The traditional action theory in philosophy asserts that desires, despite their complexity, are sufficient for directly generating any action when combined with beliefs (Audi, 1974; Davidson, 1963). Rational actions can be defined as the ones that are expected to fulfill desires (Dennett, 1987). In decision theory and artificial intelligence, this insight has been formulated as designing algorithms to maximize expected utilities (MEU) (Von Neumann & Morgenstern, 1953; Russell & Norvig, 2002). Complex desires can be simultaneously maintained by defining all of them as part of the reward function. Agents do not need to make the hard choice among desires and can simply act to maximize expected utility, where the expectation is jointly evaluated by the probability of all future states and how well the states can satisfy all desires.

The power of MEU has been demonstrated by modern AI such as deep reinforcement learning, which is able to generate complex intelligent behaviors, reaching human-expert level performance in games like Atari (Mnih et al., 2015) and Go (Silver et al., 2016, 2017). The importance of modeling desires has also been widely accepted in cognitive psychology, such as Theory of Mind (ToM), assuming that humans spontaneously explain others' actions by attributing them to the combination of beliefs and desires (Wellman, 2014). Combining Bayesian inference and MEU, it has been shown that human decision making can be considered as a naive utility calculus with positive rewards and negative costs (Jara-Ettinger, Gweon, Schulz, & Tenenbaum, 2016).

On the other hand, intention models are based on the assumption that other than beliefs and desires, intention is also an indispensable mental state. Intention is the "deliberation" of what to do based on belief as the information and desire as the motivation (Bratman, 1987). Desires do not directly drive human actions but are instead mediated by intentions (Harman, 1986; Searle, Willis, et al., 1983). Intention-based actions do not consider the expectation of all future states evaluated by all desires but requires deliberation of what desires to choose. Intention then serves as a proactive commitment to a fixed plan towards a specific goal (Bandura, 2001).

Therefore, any conflicting nature of desires must be "filtered out" before forming an intention to execute actions: An agent is allowed to desire conflicting things but not to intend conflicting things (Bratman, 1987). The philosophical theory of intention has been supported by human empirical studies using introspection and self-reports (Malle & Knobe, 2001; Perugini & Bagozzi, 2004; Schult, 2002). The feasibility of modeling intention computationally has also been demonstrated: Early work of logical AI formalizes intention as selecting a goal for persistent pursuit (Cohen & Levesque, 1990). More recently, intention has been modeled as optimizing the order of destinations with different rewards, which allows the model to focus on one destination at a time (Jara-Ettinger, Schulz, & Tenenbaum, 2020).

Another study focused more on the psychophysics of intentional actions in adults (Cheng et al., 2021). In a 2D navigation task, it compared humans with an optimal Markov Decision Process (MDP) model. The results showed that human actions qualitatively deviate from modeled actions with several behavioral signatures of intentional commitment: "Disruption resistance," with which humans persistently pursue a plan despite setbacks; "Temporal leap", where people commit to a distant future even before achieving the proximal

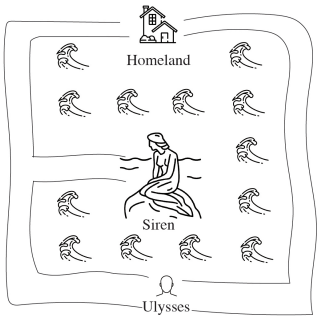


Figure 1: **Ulysses and Siren.** Ulysses faces two paths: one with the option to go to the Siren, and the other one with the fixed future of going home only.

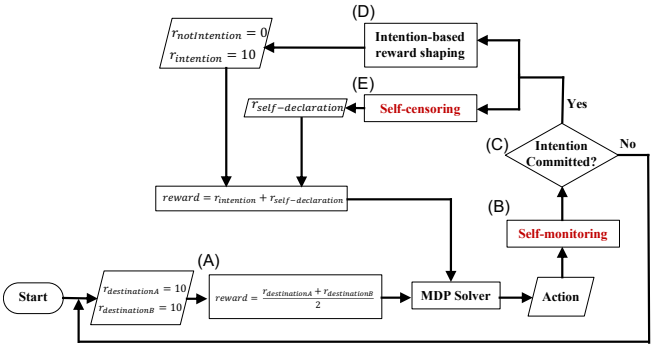


Figure 2: **A meta-cognitive model of intention.**

one; "Ulysses-constraint of freedom", which, especially related to the current work, refers to the proactive constraint of one's freedom by avoiding a path that could lead to many futures, similar to Ulysses's self-binding to resist the temptation of the Siren's song (Fig. 1). In the study, participants were given the opportunity to "self-bind" at a crossroad with two paths: an open-ended path that could lead to two destinations, or a fixed-future path that leads to only one destination. Similar to the spirit of Ulysses, participants are biased towards choosing a path that leads to one fixed future, instead of the one with the freedom to choose different destinations. Such behavior is essentially the participants' declaration to maintaining a fixed intention. Since participants finished the tasks individually, the Ulysses-constraint of freedom can be taken as a "self-declaration" without being influenced by others. Together, these behavioral results show the inflexibility of human actions, indicating that they are indeed driven by committed intentions, instead of the desire to optimize the expectation of multiple rewards.

Intention commitment through an internalized theory of mind

ToM serves as the basis for people to understand others' actions. People further use it to censor their own actions to present a better version of themselves to others (Goffman et al., 1978). Once such mindset is internalized, even when people are alone by themselves, they would still imagine an observer being present and act in a way that satisfies his expectations. The ability to view one's own mind from an objective perspective has been considered as a great achievement of human rationality (Dennett, 1996), which has also been highlighted as a type of meta-cognitive process (Flavell, 1979; Bandura, 1989). Applying the Ulysses-constraint of freedom phenomenon, people chose to abandon the choice with multiple destinations in order to let the imagined observer understand their intentions more easily.

Empirical studies on commitment

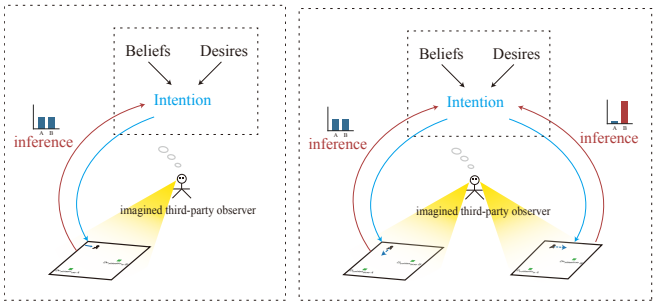
Commitment is the hallmark of intention. Empirical studies on commitment were first explored by economists, who found that humans are not fully rational as utility-maximizers due to the fact that their preferences may change over time—referred to as the "changing tastes" problem (Strotz, 1955). To forestall the changing tastes, commitment has been proposed as a regulation device to deal with the temporal fluctuations of preferences (Thaler, 1980; Schelling, 1980; Bryan, Karlan, & Nelson, 2010). However, economists' focuses are on explaining consumers' behavior and providing high-level commitment strategies, instead of the cognitive mechanisms of intentional commitment under ToM.

Psychological researchers, on the other hand, have been looking into how people control their desires executively since the time of Sigmund Freud. When faced with temporal fluctuations of preferences, a lack of self-control frequently leads to inconsistency in behavior (Ainslie, 1975). One classic demonstration of executive control is children's ability to suppress current impulses in exchange for larger future advantages (Mischel, Ebbesen, & Raskoff Zeiss, 1972; Choe, Keil, & Bloom, 2005). Although children do not understand internal conflicting desires until they are at least 7 years old (Choe et al., 2005), they appear to demonstrate persistence towards a specific goal as early as infancy (Leonard, Lee, & Schulz, 2017). These findings indicate people's capability to self-regulate to resolve conflicting desires. Yet, it is not clear how the commitment is achieved through self-control, especially how it can be computationally formalized through a modern Bayesian ToM (BToM) model (Baker, Saxe, & Tenenbaum, 2009).

A meta-cognitive model of intention

We focus on how the self-declaration of intention can be modeled as an internalized BToM, which works as an imagined third-party observer (ITO) constantly monitoring and censoring one's own actions. Here we only consider the scenario with two possible goals, while the model can be easily generalized to multi-goal situations (Fig. 2). The model has the following stages:

(A) Action planning before committing Before committing to an intention, an agent acts by considering all possible rewards and aims to maximize the expected utility. We solve the problem of planning with rewards as a Markov Decision Process (MDP) model whose terminal states are all desired goals. Consider there are two goal states, A and B, with terminal rewards r_A, r_B respectively. The agent's reward function is defined as the averaged reward:



(a) **Self-monitoring**. Intention inference from an imagined third-party observer (ITO) perspective.

(b) **Self-censoring**. Choosing action that best demonstrate a committed intention.

Figure 3: **Meta-cognitive processes of intention.**

$$r = (r_A + r_B)/2 \tag{1}$$

Given this reward function, the action policy can be obtained from classic dynamic programming algorithms such as Value Iteration (Bellman, 1957). This policy implies that the agent acts to maximize the expectation of all rewards without committing to a specific goal. The output of this uncommitted policy $P_\beta(action \mid state)$ is based on a soft-max function with a rationality parameter β. When $\beta \to 0$, the agent acts randomly; when $\beta \to +\infty$, the agent acts greedily by the optimal solution. Here we chose $\beta = 2.5$ following previous studies in modeling human actions with MDP (Baker et al., 2009).

(B) Self-monitoring. The observed actions generated in step (A) are constantly monitored by the ITO (Fig. 3a). Although the generation of an action is not based on any specific intention, the ITO always tries to attribute it to an intention through Bayesian goal inference (Baker et al., 2009).

$$\begin{aligned} &P_\beta\left(intention \mid action_{1:T}, state_{1:T}\right) \\ &\propto \prod_{t=1}^{T} P_\beta\left(action_t \mid state_t, intention\right) P(intention) \end{aligned} \tag{2}$$

The action likelihood, $P_\beta\left(\, action_t \mid state_t\right)$, is derived from an optimal policy of an MDP whose terminal state is the agent's intention with a reward function defined as $r = r_{intention}$. The output of this self-monitoring process is the posterior probability of all possible intentions.

(C) Commit or not? Once having a posterior distribution of intentions calculated from self-monitoring, the agent needs to decide whether to commit, and if so, which one of the two intentions to commit to. The decision is based on the difference of posterior probability between two intentions:

$$\Delta_{Posteriors} = |P(A) - P(B)| \tag{3}$$

This difference can be modeled as a Just-noticeable difference (*JND*), which is a standard protocol in modeling human perception in psychophysics (Weber's Law). If $\Delta_{Posteriors} < JND$, no intention will be selected, and the agent will continue to follow the uncommitted policy defined in Step (A); If $\Delta_{Posteriors} > JND$, the agent will commit by sampling an intention from the posterior distribution of the two intentions $[P(A), P(B)]$. Here we use sampling instead of deterministic decision rule to allow for re-planning (Bratman, 1987), as it is possible for agents to switch intentions after committing.

(D) Intention-based reward shaping. Once an intention has been committed, only the one reward term consistent with the intention will be maintained, while all others will be ignored for future action planning.

$$r = r_{intention} \tag{4}$$

(E) Self-censoring. If an intention is sampled in Step (C), the agent will commit to it through a self-censoring process (Fig. 3b). Unlike self-monitoring, which infers the agent's intention given an already generated action, self-censoring evaluates possible future states by simulating the consequences of different actions and how the future states will influence the ITO's inference of the agent's intention. The purpose of self-censoring is to discourage the agent from entering states that will confuse the ITO's in recognizing the already-committed intention. This is achieved by defining the desirability of a state $r_{self-declaration}$ to be the likelihood ratio of entering that state given the committed intention. Intuitively, this will encourage the agent to enter states that only the committed intention can lead to while others cannot. This likelihood is computed by integrating out the action probability given policy and transition uncertainty.

$$r_{self-declaration} = \frac{P\left(state_{t+1} \mid state_t, intention_{committed}\right)}{\sum_{intention \in I} P\left(state_{t+1} \mid state_t, intention\right)} \tag{5}$$

$$\begin{aligned} &P\left(state_{t+1} \mid state_t, intention\right) \\ &= \sum_{action \in A} P\left(state_{t+1} \mid s_t, action\right) P\left(action \mid state_t, intention\right) \end{aligned} \tag{6}$$

(F) Action planning with self-commitment. Considering both (D) and (E), the agent will take actions that 1) physically leads to the goal and 2) mentally facilitates the ITO's inference. The committed policy is solved by an MDP with reward function defined by a weighted sum of $r_{intention}$ and $r_{self-declaration}$:

$$r = r_{intention} + \alpha r_{self-declaration} \tag{7}$$

where $\alpha \in [0, +\infty)$ is a declaration parameter. When $\alpha \to +\infty$, the agent acts to best declare its commitment through facilitating the ITO's inference of its intention.

Experiments

Overview

Here we tested our theory of commitment by modeling a previous human experiment and running a new human experiment. In Experiment 1, we used our model to explain the behavioral signatures of Experiment 2 of the Cheng et al.'s (2021) study (stimuli can be found at https://osf.io/k5e69/) including "disruption resistance" and "Ulysses-constraint". Experiment 2 further tested the "Enhanced legibility" assumption that humans should purposefully demonstrate their intentions in order to make it easier for an observer (human or Bayesian model) to infer their intentions.

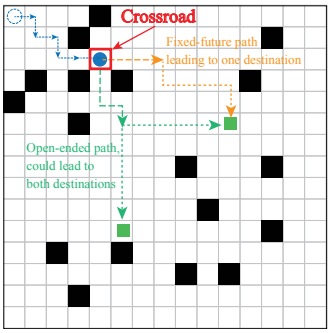


Figure 4: **Design of the crossroad.** An agent at a crossroad. Agents can either choose a fixed-future path that leads to one destination (orange arrow) or choose an open-ended path that could lead to both destinations (green arrow). This is an sample map of 6-steps-to-crossroad condition, defined as the length of the shortest path between the agent's starting position (dashed blue circle) and the crossroads was 6.

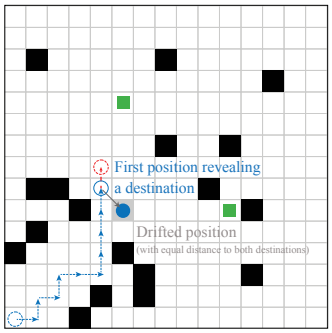
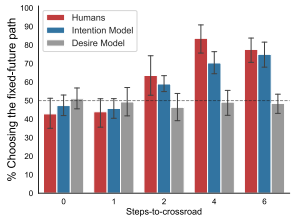


Figure 5: **Design of the deliberate disruption.** Design of the deliberate disruption. Once an agent revealed its destination, it was immediately pushed back to a position equally distanced between the two destinations.

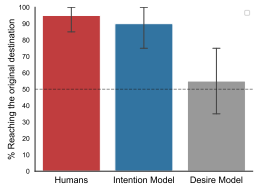
Experiment 1: Modeling Ulysses-constraint and disruption resistance

The goal of this modeling experiment is to use our model to explain behavioral signature of intentional commitment observed in Experiment 2 of the Cheng et al.'s (2021) study . In the experiment, humans are required to reach one of two goals in a 2D map, with carefully designed crossroads that lead to two paths: One that constrains agents towards one fixed destination, and the other one with options to switch (Fig. 4). The result demonstrated the Ulysses-constraint, indicated by the fact that humans prefer the fixed-destination path over the open-ended one, even though the expected utility of taking the two paths was identical from a reward maximization perspective. In addition, the number of steps for the humans to reach the crossroad was systematically manipulated, varying from [0, 1, 2, 4, 6] steps. The effects of Ulysses-constraint gradually increase as the number of steps to reach the crossroad increases, suggesting that intentional commitment does not emerge immediately but takes time and deliberation. To reveal the behavioral signature of "disruption resistance", there was noisy drift that would drag the agents away from their intended locations during the entire course of the experiment. The design was to cover for a special trial added at the end of the experiment. In this trial, the drift was not random but deliberately designed so that when the agent first revealed its destination to reach, the drift would drag the agent back to a position equally distanced from the two destinations (see Fig. 5). Instead of re-planning and choosing the two destinations with equal probability, humans strongly prefer to fight back at the disruption and resume their pursuit of the originally revealed destination.

In our study, we predict that an agent controlled by our intention model can demonstrate both the Ulysses-constraint and the disruption resistance phenomena. In contrast, a desire model without intentional commitment, implemented as including only step (A) of the intention model, should not be able to demonstrate either of the two.



(a) **Ulysses-constraint of freedom.** Percentage of choosing the fixed-future paths as a function of steps-to-crossroad.



(b) **Disruption resistance.** Percentage of reaching the original destination with different types of agent.

Figure 6: **Experimental results.** The error bars reflect 95% confidence intervals.

Results

Results of both intention model and desire model, together with human experiment results from Cheng et al. (2021) are shown in Fig. 6a. For the intention model, we fitted the threshold parameter to $JND = 0.08$ and the declaration parameter to participants' judgments to $\alpha = 5.5$ by minimizing the sum of squared errors between model's and humans' choices between the two paths across all steps-to-crossroad conditions; $r = 0.98$, $RMSE = 0.075$. For the desire model, there was no free parameter to fit. Both the intention model and the desire model used a fixed rationality parameter $\beta = 2.5$, which match the overall task performances of humans' 2D navigation. The desire model shows 50% under all conditions, regardless of the rationality parameter. Consequently, we cannot apply the same method of model fitting. Instead, we adjust the model parameters to match the number of steps humans need to reach their goals.

Ulysses-constraint of freedom Overall, the intention model preferred the fixed-future path (59%; 95% CI [0.56,0.62]) over the open-ended path (41%) (one sample t-test with a 50% baseline, $t(19) = 4.17$, two-tailed $p < 0.001$, $BF_{10} = 11150$, $d = 1.52$). The main effect of steps-to-crossroad was significant ($F(4, 95) = 19.34$, $p < 0.001$, $\eta_p^2 = 0.45$). They chose the fixed-future path more at steps

2, 4, 6, as revealed by one sample t-test with a 50% baseline (all $ps < 0.05$). Post-hoc analyses further indicated that this preference was much stronger at steps 4 and 6, compared with step 2 (all $ps < 0.05$).

This pattern is consistent with human data which shows no commitment at step 0 and 1. According to our model, this is because at these two steps, there is not enough evidence for the ITO to strongly favor one intention over the other, hence triggering the commitment. Once enough trajectory information has been obtained (after step 2), the ITO can infer the intentions with posterior high enough to trigger commitment. The agent will bias against the open-ended path, as here it is difficult for the ITO to infer the agent's intentions with both being able to explain the actions. As expected, the desire model shows no preference at all crossroads simply because the expected utility of taking these two paths are identical.

The intention model with the same parameters also explains disruption resistance in human behavior (90%; 95% CI [0.76,1]) by reaching the original destination much more than the chance level (Fig. 6b) (one sample t-test with a 50% baseline, $t(19) = 13$, two-tailed $p < 0.001$, $BF_{10} = 1.36e+08$, Cohen's $d = 2.91$), while the result of the desire model (55%; 95% CI [0.31,0.79]) showed no difference to a chance level. These results demonstrate that the intention model captured both the "Ulysses-constraint of freedom" and "disruption resistance" signatures of human intentional actions.

Experiment 2: Enhanced legibility

The core of our theory of intentional commitment is that humans act to make it easier for an ITO to infer their intentions. Beyond reporting the signatures in the Cheng et al.'s (2021) study, our model also generate predictions that can be tested. First, it predicts that human observers should actually find the destinations of trajectories generated by humans and the intention model easier to predict, compared with those generated by the desire model. We refer to this prediction as "enhanced legibility." Note that the result should hold even with the desire model following a policy that can reach the goal as efficiently as the human subjects, optimally solved by MDP. This prediction is tested in Experiment 2a.

Another core argument we have here is that ITO is functionally identical to an imagined human observer. If so, when replacing the human observer with an ITO to infer the destination of trajectories, we should produce similar enhanced legibility results. This prediction is tested in Experiment 2b.

Experiment 2a. Enhanced legibility in the eyes of real human observers

Participants A total of sixty participants (32 females, M_{age} = 22.17, SD = 4.17) were recruited for credits or payments. They were evenly split to three groups with human actors, intention model actors, desire model actors, respectively. All participants were given informed consent.

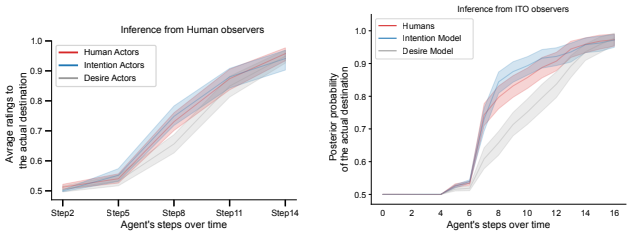
Method The human trajectories in our experiment were taken from Experiment 2 of Cheng et al. (2021), Trajectories of intention model and desire model came from Experiment 1. To prevent participants from detecting the regularities in the crossroads, 50 trajectories were randomly mixed with 25 trajectories with randomly generated maps. Data from the random trajectories were not analyzed. In all groups, participants were asked to watch the trajectories and then predict which one of the two destinations is the goal. Each trial started with a central fixation of 1300ms. After that, a 2D navigation game display was presented with a blue agent, two red destinations and black barriers. Each time step of the agent's motion was presented for 500ms. At the [2,5,8,11,14] steps, the display freezes, at which time participants were asked to identify the goal and rate their confidence using a 9-point scale by clicking with the mouse on an array of numbered boxes (from very unlikely to very likely), aligned horizontally at the bottom of the screen. After the last report, the trajectory continues until the agent reaches the destination.

Results When participants select the non-target as the goal with confidence p, the probability of the actual target is $1-p$. Each time the mean posterior of the actual destination over the group of participants is computed by averaging participants' confidence in judging the target as the true destination.

Fig. 7a shows how the posterior evolves over time. The difference in posterior is the most salient at step 8, which is around the middle of the trajectory. At this step, the posterior from the human actor (0.73, 95% CI, [0.71, 0.75], $t(38) = 3.29$, p = 0.002, BF_{10} = 16.43, Cohen's $d = 1.04$) and the intention actor (0.75, 95% CI, [0.73, 0.77], $t(38)$ = 3.86, p <0.001, BF_{10} = 62.84, Cohen'd = 1.22) are significantly higher than that of the desire actor (0.66, 95% CI, [0.64, 0.67]). There is no significant difference between judgments to trajectories from the human data and the intention model ($t(38)$ = -0.74, p = 0.464, BF_{10} = 0.38, Cohen'd = 0.23). These results show that indeed humans find trajectories from human actors and intention actors easier to interpret, indicating the "enhanced legibility" nature of self-declaration.

Experiment 2b. Enhanced legibility in the eyes of imagined third-party observers (ITO) We employed the BToM model as an imagined observer to test the "enhanced legibility". The ITO model is rationally derived from Bayesian inference over MDP policies for reaching each goal. This model is not trained on any data of how humans interpret trajectories. We are curious about whether this model can nevertheless function in a similar way as the human observers in Experiment 2a. Specifically, the experiment here is identical to Experiment 2a except that the human observer is replaced by the ITO model, which infers intentions from trajectories generated by the same human actors, intention actors, and desire actors. The ITO model infers the agent's destination by Equation 2. The initial $P_0(destination)$ was set to 0.5 for both potential destinations. The results are analyzed in the same way as Experiment 2a.

Results The posterior of the ITO inference of the actual destination (finally reached) was plotted in Fig. 7b. The ITO is able to infer the actual destination much faster than the desire model (cluster-based permutation tests identified significant gaps from steps 5 to 14, all ps <0.05). Overall, the results show that trajectories generated by human actors and intention actors are equally legible from the perspective of



(a) Enhanced legibility to human observer.

(b) Enhanced legibility to ITO.

Figure 7: **Results of enhanced legibility.** Posterior of actual destination inferred by human and ITO observers as a function of steps. Error shading denotes 95% confidence intervals.

the ITO observer. This strongly supports our assumption that when human actors generate actions, they are indeed trying to facilitate the ITO's inference of their intention.

ITO captures the human judgement qualitatively but is more extreme in its prediction. There are two possible reasons. One is that ITO assumes less randomness in actions so that the likelihood for the intentions become more distinct from each other. The other is that ITO has perfect memory and uses all past information as reference to make judgements. Humans observers, however, are limited by working memory capacity and experience memory decay when new information arrives (Gao, Baker, Tang, Xu, & Tenenbaum, 2019). In the future, the ITO may be improved by assuming a similar temporal decay in accumulated information over time.

These results collectively demonstrate that the "enhanced legibility" nature of intention was quantitatively captured.

Discussion

We proposed a meta-cognition model of intention based on an internalized ToM framework. Our model quantitatively captured several existing behavioral signatures of human intention and discovered new ones. Similar to humans, the model displays the "Ulysses-constraint of freedom" as gradually biasing against an open-ended path. It also explains why humans persistently pursue the same goal even after setbacks, as shown in the "disruption resistance" results. Moreover, results indicate that for both human observers and ITO, it is much easier to read intentions from trajectories of human actors and the intention model, as compared to the desire model which simply maximizes the expected utilities. These results support our core hypothesis that when facing conflicting desires, people act not just to maximize the expected reward, but also to better demonstrate their committed intentions.

One potential explanation for why humans act differently from a rational MEU-based model is to reduce the computational cost. While humans are indeed limited by computational resources (Lieder & Griffiths, 2020; Gershman, 2021), and committing to one intention can reduce the burden of planning, this alone does not suffice to explain human behavior in our experiments. First, when facing two equally desirable paths to one committed intention, the simplest solution is to randomly sample one, but humans were systematically biased to the path that binds themselves. Second, our results indicate that the emergence of humans' bias depends on when they arrive at the crossroad. If the bias was due to the computational limits, people should be less biased when arriving later, as they had more time to make a decision close to the optimal policy — however, our results suggested otherwise.

One surprising conclusion from the above experiments is that humans act to better demonstrate their intention even when alone. Here we offer two possible motivations for such behavior. One is to avoid temptation. From a computational perspective, although such consideration requires additional resources, it indeed helps save computational costs in the long run — by constraining the freedom of choice like Ulysses, the cognitive cost of resisting temptation were offloaded to the environment. Another potential motivation is to increase social legibility. From an evolutionary perspective, the self-demonstration behavior can be considered as an analogy to the unique morphology of the human eye. Unlike chimpanzees, humans have evolved with a high color contrast between the white sclera and the darker colored iris in order to better convey their attention with different displacements of the gaze (Kobayashi & Kohshima, 1997; Tomasello, Hare, Lehmann, & Call, 2007). This cooperative eye hypothesis is similar to Dennett's hypothesis: while ToM is originally developed to understand others, due to evolutionary pressures of cooperation and communication, it has been internalized to monitor one's own actions (Dennett, 1996). This internalized social evaluation is also the foundation of Vygotsky's theory (Vygotsky & Cole, 1978), which has been supported by empirical studies on children's understanding of commitment in collaboration (Tomasello, Carpenter, Call, Behne, & Moll, 2005). From this perspective, the imagined observer we proposed here is similar to concept of imagined audience in developmental and social psychology (Piaget, 1952; Elkind & Bowen, 1979). We would like to further highlight that both motivations mentioned here rely on the self-censoring mechanism to make the intention more obvious. The two motivations do not contradict each other, since reducing temptation through commitment helps the group focus on the same goal and facilitates cooperation (Tang et al., 2022), and finding a temptation-free path requires meta-cognition.

To model human intentional behavior in an individual task, we ended up with a model that is surprisingly similar to the work of showing versus doing in the context of multi-agent communication (Dragan, Lee, & Srinivasa, 2013; Shafto, Goodman, & Griffiths, 2014; Ho, Littman, MacGlashan, Cushman, & Austerweil, 2016). For instance, a teacher may want to demonstrate his pedagogical intention by carefully picking examples to facilitate the student's Bayesian inference of the teacher's intention. The resemblance between our model and theirs is striking given the differences in purpose of the two lines of work: We aim to model human intentions in a purely individual setting, while theirs focuses on communication between multiple agents. Such similarity shows that humans are indeed highly socialized creatures who, even when acting alone, still communicate with themselves. This resemblance also suggests that the meta-cognition model proposed here is hard to further simplify, as the demonstration model we incorporate here is the most appropriate to our knowledge.

References

Ainslie, G. (1975). Specious reward: a behavioral theory of impulsiveness and impulse control. *Psychological bulletin*, *82*(4), 463.

Audi, R. (1974). Intending. *The Journal of Philosophy*, *70*(13), 387–403.

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329–349.

Bandura, A. (1989). Human agency in social cognitive theory. *American psychologist*, *44*(9), 1175.

Bandura, A. (2001). Social cognitive theory: An agentic perspective. *Annual review of psychology*, *52*(1), 1–26.

Bellman, R. (1957). A markovian decision process. *Journal of mathematics and mechanics*, *6*(5), 679–684.

Bratman, M. (1987). Intention, plans, and practical reason.

Bryan, G., Karlan, D., & Nelson, S. (2010). Commitment devices. *Annu. Rev. Econ.*, *2*(1), 671–698.

Cheng, S., Tang, N., An, W., Zhao, Y., Zhou, J., Shen, M., & Gao, T. (2021). Intention beyond desire: Humans spontaneously commit to future actions. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 43).

Choe, K. S., Keil, F. C., & Bloom, P. (2005). Children's understanding of the ulysses conflict. *Developmental Science*, *8*(5), 387–392.

Cohen, P. R., & Levesque, H. J. (1990). Intention is choice with commitment. *Artificial intelligence*, *42*(2-3), 213–261.

Davidson, D. (1963). Actions, reasons, and causes. *The journal of philosophy*, *60*(23), 685–700.

Dennett, D. C. (1987). *The intentional stance*. MIT press.

Dennett, D. C. (1996). *Kinds of minds: Toward an understanding of consciousness*. Basic Books.

Dragan, A. D., Lee, K. C., & Srinivasa, S. S. (2013). Legibility and predictability of robot motion. In *2013 8th acm/ieee international conference on human-robot interaction (hri)* (pp. 301–308).

Elkind, D., & Bowen, R. (1979). Imaginary audience behavior in children and adolescents. *Developmental psychology*, *15*(1), 38.

Elster, J. (1987). *The multiple self*. Cambridge University Press.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American psychologist*, *34*(10), 906.

Freud, S. (1923). The ego and the id. *Standard Edition*, *19*, 3–59.

Gao, T., Baker, C. L., Tang, N., Xu, H., & Tenenbaum, J. B. (2019). The cognitive architecture of perceived animacy: Intention, attention, and memory. *Cognitive science*, *43*(8), e12775.

Gershman, S. (2021). *What makes us smart: The computational logic of human cognition*. Princeton University Press.

Goffman, E., et al. (1978). *The presentation of self in everyday life* (Vol. 21). Harmondsworth London.

Harman, G. (1986). *Change in view: Principles of reasoning*. The MIT Press.

Ho, M. K., Littman, M., MacGlashan, J., Cushman, F., & Austerweil, J. L. (2016). Showing versus doing: Teaching by demonstration. *Advances in neural information processing systems*, *29*, 3027–3035.

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

Jara-Ettinger, J., Schulz, L. E., & Tenenbaum, J. B. (2020). The naive utility calculus as a unified, quantitative framework for action understanding. *Cognitive Psychology*, *123*, 101334.

Kobayashi, H., & Kohshima, S. (1997). Unique morphology of the human eye. *Nature*, *387*(6635), 767–768.

Leonard, J. A., Lee, Y., & Schulz, L. E. (2017). Infants make more attempts to achieve a goal when they see adults persist. *Science*, *357*(6357), 1290–1294.

Lieder, F., & Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, *43*.

Malle, B. F., & Knobe, J. (2001). The distinction between desire and intention: A folk-conceptual analysis. *Intentions and intentionality: Foundations of social cognition*, *45*, 67.

Mischel, W., Ebbesen, E. B., & Raskoff Zeiss, A. (1972). Cognitive and attentional mechanisms in delay of gratification. *Journal of personality and social psychology*, *21*(2), 204.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... others (2015). Human-level control through deep reinforcement learning. *nature*, *518*(7540), 529–533.

Perugini, M., & Bagozzi, R. P. (2004). The distinction between desires and intentions. *European Journal of Social Psychology*, *34*(1), 69–84.

Piaget, J. (1952). Play, dreams and imitation in childhood.

Russell, S., & Norvig, P. (2002). Artificial intelligence: a modern approach.

Schelling, T. C. (1980). *The strategy of conflict: with a new preface by the author*. Harvard university press.

Schelling, T. C. (1984). *Choice and consequence*. Harvard University Press.

Schult, C. A. (2002). Children's understanding of the distinction between intentions and desires. *Child Development*, *73*(6), 1727–1747.

Searle, J. R., Willis, S., et al. (1983). *Intentionality: An essay in the philosophy of mind*. Cambridge university press.

Shafto, P., Goodman, N. D., & Griffiths, T. L. (2014). A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive psychology*, *71*, 55–89.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., ... others (2016). Mastering the game of go with deep neural networks and tree search. *nature*, *529*(7587), 484–489.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., ... others (2017). Mastering the game of go without human knowledge. *nature*, *550*(7676), 354–359.

Strotz, R. H. (1955). Myopia and inconsistency in dynamic utility maximization. *The review of economic studies*, *23*(3), 165–180.

Tang, N., Gong, S., Zhao, M., Gu, C., Zhou, J., Shen, M., & Gao, T. (2022). Exploring an imagined "we" in human collective hunting: Joint commitment within shared intentionality. In *Proceedings of the annual meeting of the cognitive science society.*

Thaler, R. (1980). Toward a positive theory of consumer choice. *Journal of economic behavior & organization*, *1*(1), 39–60.

Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and brain sciences*, *28*(5), 675–691.

Tomasello, M., Hare, B., Lehmann, H., & Call, J. (2007). Reliance on head versus eyes in the gaze following of great apes and human infants: the cooperative eye hypothesis. *Journal of human evolution*, *52*(3), 314–320.

Von Neumann, J., & Morgenstern, O. (1953). *Theory of games and economic behavior (3rd editon).* New York: John Wiley.

Vygotsky, L. S., & Cole, M. (1978). *Mind in society: Development of higher psychological processes*. Harvard university press.

Wellman, H. M. (2014). *Making minds: How theory of mind develops*. Oxford University Press.