An Adaptive Model of Gaze-based Selection

Xiuli Chen
xiuli.chen@aalto.fi
Department of Communications and Networking,
Aalto University, Finland
Finnish Center for Artificial Intelligence, Finland

Antti Oulasvirta
antti.oulasvirta@aalto.fi
Department of Communications and Networking,
Aalto University, Finland
Finnish Center for Artificial Intelligence, Finland

Aditya Acharya
aditya.acharya@aalto.fi
Department of Computer Science, Aalto University,
Finland
Finnish Center for Artificial Intelligence, Finland

Andrew Howes
howesa@bham.ac.uk
Department of Communications and Networking,
Aalto University, Finland
Finnish Center for Artificial Intelligence, Finland
School of Computer Science, University of Birmingham
Birmingham, UK

ABSTRACT
Gaze-based selection has received significant academic attention over a number of years. While advances have been made, it is possible that further progress could be made if there were a deeper understanding of the adaptive nature of the mechanisms that guide eye movement and vision. Control of eye movement typically results in a sequence of movements (saccades) and fixations followed by a ‘dwell’ at a target and a selection. To shed light on how these sequences are planned, this paper presents a computational model of the control of eye movements in gaze-based selection. We formulate the model as an optimal sequential planning problem bounded by the limits of the human visual and motor systems and use reinforcement learning to approximate optimal solutions. The model accurately replicates earlier results on the effects of target size and distance and captures a number of other aspects of performance. The model can be used to predict number of fixations and duration required to make a gaze-based selection. The future development of the model is discussed.

CCS CONCEPTS
• Human-centered computing → HCI theory, concepts and models; User models.

KEYWORDS
Reinforcement learning, gaze-based selection, adaptive model, computational rationality

1 INTRODUCTION
Gaze-based selection has been studied in human-computer interaction (HCI) since the 1990s [29, 69] and interest has been maintained throughout the intervening years [36, 39, 43, 48, 51, 59, 68]. Recent technological advances, such as the addition of eye tracking capabilities to virtual reality and augmented reality (AR), have sparked another wave of interest [4, 22, 49]. Despite the great potential of faster, more intuitive and hands-free interaction, the comparison of performance between gaze input and other input methods, such as a mouse, touch input, or head-mounted sights has yielded conflicting results. Gaze input was found to be either faster than head-pointing but slower than mouse pointing [22], or even slower than head-pointing [49]. A different study compared gaze and head pointing for different simulated AR fields of view (FOV) and found a clear advantage for gaze pointing especially at larger FOV [4].

One reason these discrepancies may be hard to reconcile is the fact that it is only relatively recent that effort has been put into modeling eye movement and gaze selection for HCI [64, 72]. Drawing robust conclusions from data obtained across diverse empirical conditions is difficult, partly because, in the absence of a robust model, results can appear to be sensitive to relatively minor aspects of an experimental design, such as viewing distance or latency in tracking. Efforts to build theories of gaze-based selection could help reduce these difficulties by illuminating the underpinning reasons for the differences. So far, modeling efforts for gaze-based selection have largely centered on motor constraints and neglected constraints imposed by vision (although see [64]). Fitts’s Law has been prominent in these efforts [28, 37, 43, 69], despite a number of inadequacies [20, 44, 56, 61, 72].

In the current paper, we propose an adaptive model of gaze-based selection that uses a reinforcement learning algorithm to find an optimal policy given human-like bounds on motor movement and visual observations. The core assumption is that the control of eye movements is boundedly optimal [18, 27, 34, 50, 54]. That is, the sequence of saccades and fixations is governed by an attempt to optimize a utility function (e.g., selection time, accuracy or a trade-off between the two) constrained by the properties of both the motor system and the visual system.
A key feature of our model is that the predicted eye movement strategies are an adaptive consequence of the following constraints: (1) target distance, (2) target size, (3) distant dependent ocular-motor saccade noise, (4) distant dependent location estimation noise in peripheral vision, and (5) fixation jitter (which makes it impossible to hold the gaze location exactly still). The joint effect of these constraints on gaze-based selection is partially illustrated in Figure 1. The interaction between the target size, target eccentricity, signal-dependent ocular motor noise, and eccentricity-dependent estimation noise results, in this case, in a two-step gaze-based selection. Two-step selections are typical in humans but under some circumstances either one-step (large target) or 3-plus-steps (very small targets) can be observed.

A central goal for our modeling effort has been to replicate a standard set of empirical findings that are relevant to practical efforts to develop better gaze-based selection techniques. Therefore, the focus has been on reproducing the effects of target size and distance, studied most recently in detail by Schuetz et al. [56], as well as phenomena affecting dwell-based selection. In particular, the model we report captures the following five phenomena:

1. The effect of target size on selection time [56, 68, 72];
2. The effect of target size on dwell time [29, 72];
3. The effect of target size on the number of saccades [56, 70];
4. The effect of target size on the saccade duration of main saccades [56, 70];
5. The effect of Index of Difficulty on the number of saccades [56, 68].

In what follows, we show how these phenomena emerge through reinforcement learning. The learner solves a Partially Observable Markov Decision Process (POMDP) with an observation function that is bounded by the constraints described above. The model learns by trial-and-error in an artificial world. The model predicts both selection times and also the average number of saccades given target size and eccentricity. The work builds on previous work using reinforcement learning and deep learning to model users in human-computer interaction [5, 7, 9, 31, 35, 55].

Our research question is: Can human eye movement control during gaze-based selection be predicted as an emergent consequence of adaptation to bounds in the ocular motor and perceptual systems? This question, which links our work to the growing body of work on RL-based modeling of human performance, can be answered in the affirmative if the five empirical objectives listed above are met.

The paper makes the following contributions: (1) It proposes a new POMDP-based model of gaze-based selection which not only predicts total task performance time (aggregate) but the moment-by-moment process of eye movement control. (2) It reports a test of the model’s predictions with data from previous studies reported in [72] and [56]. (3) It illustrates how machine learning can be used to model human cognition through trial-and-error learning in an artificial world, rather than through generalization from large amounts of human data (similar to [58]’s learning through self-play).

2 RELATED WORK
2.1 Gaze-based selection: empirical findings
The time required to make a gaze-based selection is usually composed of (1) the time to search for the target (Eye Movement Time (EMT)), and (2) the dwell time to commit the selection (Dwell time). The EMTs are usually dependent on target size and target eccentricity (distance). For example, the number of saccades increases with target eccentricity [70, 72] and with target size [56, 70, 72]. That is, EMT increases with the smaller target sizes and further targets. Studies have shown that the increase in the selection time is due to the fact that it is more likely to need corrective saccades for smaller and further targets [56, 70].

The dwell time must be at least as long as the dwell criterion. Most gaze-based interaction relies on some form of dwell criterion to effect a system response [21, 38]. An icon is not selected until the eyes have been held on it for at least the duration of the dwell criterion and in practice sometimes longer because of difficulties in holding the eyes on the target (more on this below).

The performance of gaze-based selection is heavily dependent on the selection of a suitable dwell criterion. A short dwell criterion can lead to a relatively fast selection once the target is identified. However, it can also lead to the selection of many things glanced at on the display (known as the ‘Midas Touch problem’) [29]. A long dwell criterion can cause frustration while waiting for the target to be selected. Also, the long dwell criterion might make the selection of the small target very hard or impossible, due to uncontrolled jitter in fixations. Reported dwell criteria used in experiments include 50ms [56], 150 [57], 200 ms [44] and up to 800ms [72]. Studies attempted to minimize the dwell time by adjusting the button size [48], expanding the target region once a fixation on that target is identified [44], by user’s prior usage experience of the given button and button type [46], by using next-letter prediction for typing [45], and by using Fitts’ law to predict the movement time [28].

Figure 1: Illustration of the control problem in gaze-based selection: The user needs to move their gaze to a target (a black rectangle) via a number of saccade–fixation steps. In the figure the eyes start fixating at location 1. During this fixation an observation made using low acuity peripheral vision and a noisy and inaccurate estimate is made indicating that the target is at location 2. Then an eye movement is aimed at location 2 but noise in the motor system results in the eyes actually ending up fixated at location 3. The fixation is now closer to the target and another, more accurate, estimate is made of the target location (location 4). The eyes are aimed at location 4 but end up fixated at location 5 and after a short dwell, the target is selected.
2.2 Fitts’ Law for eye gaze

Fitts’ Law [15] formulates movement time (MT) for manually acquiring a target of width (W) at distance (D) as quickly and accurately as possible. The formulation is: \( MT = a + b \log_2(2D/W) \), where both a and b are regression coefficients. The framework has been shown to apply under a variety of conditions: manual selection [15], foot selection [24], wrist [41], and head mounted sights [60]. However, here we focus on the applicability of Fitts’ law for eye gaze movements.

Studies are more divided on the applicability of Fitts’ Law to modeling gaze duration [43, 57, 68, 69]. For example, while Miniotas et al., 2000 [43] demonstrated an excellent fit using Fitts’ law on gaze based selection, Sibert and Jacob 2000 [57] showed that movement times were better predicted by just saccade amplitudes. Recent studies have suggested that this discrepancy might be due to the relationship between the number of corrective saccades required to reach the target and the Index of Difficulty (ID) [56, 70]. Specifically, the increase of total movement time with ID is due to an increasing likelihood of secondary saccades at greater IDs, while latency and duration of each main saccade was amplitude dependent [70]. In addition, theoretical work has suggested that as few as two submovements is able to achieve a Fitts’ Law like relationship between movement time and ID [25, 41].

More generally, there is a fundamental difference between saccadic eye movements and many other sequential motor behaviors. In motor behaviors, such as tapping or wrist rotations, sensory feedback can be processed during the movement. However, it is known that eye movements cannot be reprogrammed on the basis of new visual information once initiated [10]. Indeed a number of papers have shown that Fitts’ Law is not a good model of data from eye-gaze tasks [17, 44, 61, 68, 71].

Together, these findings imply that a model which adaptively makes more corrective saccades for targets with higher IDs would offer a better explanation of gaze-based selection than Fitts’ Law. Such model is able to reveal a Fitts’ Law like relationship between movement time and ID during the cases where multiple saccades are needed. Furthermore, for cases where the target can be reached with one saccade, an adaptive model would make only a single saccade, in which the movement time is only amplitude dependent – a phenomenon well documented in the literature [2, 3]. Importantly, the number of saccades required for a certain task is a joint consequence of not only the task environment (e.g., target size and distance), but also the constraints imposed by human motor and visual systems. The model presented below has the capacity to make these adaptations.

2.3 POMDP models of visual search

There are a number of studies demonstrating that human visual search can be understood as the solution to a Partially Observable Markov Decision Process (POMDP) [1, 5–9, 26, 31, 35, 55]. POMDPs provide a mathematical framework that captures the interaction between an agent and a stochastic environment [30]. A crucial constraint on a POMDP is that it only has partial access to the state of the environment. It might, for example, be able to read only one word of a document at a time, and not read all words in a single cycle. Chen et al. 2017[9] reported a POMDP for the problem of using eye movements to gather information from a data visualisation in order to support decision making. The model made a sequence of partial observations (through visual fixations) until enough information had been gathered from the visualisation to make a decision. It predicted human performance on a simple decision making task.

A random dots motion discrimination task was framed as a POMDP problem by [50]. In this task, there was a group of dots moving on the screen. Some of the dots were moving to the left. Others were moving to the right. The task difficulty was manipulated by changing the proportion of dots moving in each direction. Participants were asked to determine the movement direction of the majority of dots. Rao’s (2010) model showed that primate decision times could be modeled as the solution to a POMDP that determined the optimal threshold for switching from information gathering to choice.

3 THEORY AND MODEL

The theoretical claim made in this paper is that gaze-based search and selection strategies are an emergent consequence of both the properties of the environment (the target distance, the target width) and of the constraints imposed by human visual-motor mechanisms (the spatial uncertainty of peripheral vision, the motor noise in the human ocular motor system, the ballistic nature of the eye saccade, and the dwell jitter during fixation).

In order to model adaptation to these constraints, we view gaze-based selection as a sequential decision making problem, and formulate it as a POMDP, solving this problem with machine learning to find an optimal gaze search and selection strategy. We use a deep reinforcement learning method to find an approximately optimal policy to solve the POMDP. The assumption is that where to look next is informed by the reward/cost structure of the task. Better gaze search patterns will receive higher rewards, which will reinforce good eye movement strategies. For an overview and broader background of this approach, see [26, 27].

In the following paragraphs, we report the problem formulation and model, followed by the implementation details.

3.1 Theoretical assumptions

3.1.1 Saccade duration and onset latency. Human eye movements consist of an alternating sequence of saccades and fixations. During fixations, the eye positions are relatively stable (with some small ‘jitter’ [12, 72]) and during these periods the eyes are able to extract visual information from the environment. Fast eye movements between fixations are called ‘saccades’. For amplitudes up to 15° or 20°, there is a relatively fixed relationship between the saccade amplitude and duration: \( \text{Duration} = 2.7 \times \text{Amplitude} + 37 \), where the saccade amplitude is in degrees, and the duration is in milliseconds (ms) [2, 3]. Most saccades terminate within a few tens of milliseconds (e.g., 60 ms in [68]). It is generally accepted that saccades cannot be reprogrammed on the basis of new visual information once initiated [2, 10, 66].

When the observer is asked to search and fixate at a target as soon as it appears, the onset latency is the time from when the target appears to the time when the first saccade movement begins. The onset latency for the primary saccade is usually around 200 ms [14].
However, it can be as low as 100ms, or as high as 350ms [14]. The onset latency is in general shorter when targets are in fixed and known locations, longer when there are distractors on the display [70] or multiple potential target locations [42]. Corrective saccades can occur a short interval (30 to 60 ms) after the primary saccade, the eyes do not have to be completely stopped during these very short intersaccadic intervals [16].

3.1.2 Spatial visual uncertainty. Human vision has the highest clarity in the fovea, which only covers around 1° - 2° of visual angle [12]. The acuity drops drastically with eccentricity from the fovea. The perception of colour, shape, and size deteriorate in peripheral vision to different extents [32]. While participants were asked to estimate a target location in their peripheral vision, the endpoint diversity (standard deviation) was found to be linear with eccentricity [42, 65]; an insight that we use in our model. More details can be found in 3.2 (Observation function).

3.1.3 Ocular motor noise. So as to model ocular model noise, we used a commonly accepted physiological assumption, signal dependent noise (SDN). SDN is zero-mean, white Gaussian noise in the magnitude of the signal with a standard deviation proportional to the magnitude of the signal. SDN has been used extensively in modeling studies [23, 62, 63, 66, 67]. More details can be found in 3.2 (Observation function).

3.1.4 Fixational eye movements. It is known that even during fixations, human eyes do not remain absolutely still. The eyes ‘jitter’ in a small area. These ‘jitters’ are referred to as fixational eye movement.

There are three categories of fixational eye movements: microsaccades, ocular drifts, and ocular microtremor [12]. Their causes have received a great deal of attention[12, 52, 53]. For example, microsaccades have been identified as potential indicators of mental fatigue [12]. A recent study and review on fixational eye movements can be found in [52]. Here, we focus on how the fixational eye movements influence the dwell criterion used to commit a target selection. For example, when dwelling on a small target region (1.2°) [72], this jitter causes the gaze to leave and reenter the target area several times, resulting in an actual target dwell time of 1075 ms for a target with a dwell criterion of 800 ms.

In our model, after the agent first fixated the target (first entry in the target region), the dwell drift was implemented by taking noisy location samples from a Gaussian distribution centered on the target (assuming that the participants were trying to maintain the gaze at the target center so as to reach the minimum dwell time and activate the selection). The standard deviation \( \rho_{drift} \) is a hyperparameter that we fit to the human data (Table 1).

3.2 Problem formulation

As we have said, we model the gaze based selection problem as a Partially Observable Markov Decision Process (POMDP) [30]. In this section, we give a formal definition of this problem. First, examine the overview given in Figure 2. On each time step, the environment is in one of the possible states. The state of the environment is initially unknown to the agent, and only partially observable. It starts with an initial belief about the state, and takes an action that causes a state change in the environment. One consequence of the action is that the agent receives a new observation, which is used to update the agent’s current belief about the state to a new belief. The other consequence of the action is that the agent receives a reward signal. The reward depends both on the state of the environment and the action taken. This action-observation-reward cycle repeats until the selection has occurred and the task is completed. Learning is achieved by adjusting the mapping between the belief and the action so as to maximize the reward (i.e., optimal policy).

A formal description of our task as a POMDP [30] is given in the following.

- **State** \( S \): At each time step \( t \), the environment is occupied at a state \( s_t \), \( s_t \in S \). A state represents a possible target position, and denotes as \( s = [x, y] \) where \( x, y \in [-1, 1] \), with -1 and 1 being the edge of the display. Specifically, at each trial (episode), the agent starts with a fixation at the starting location \( [0, 0] \) (as with the participants in the experiments). A circular target (with diameter \( W \)) is then presented at a \( D \) away from the starting location at a random angle \( \theta \). That is, the state consists of a real-valued \( x, y \) coordinate pair representing the target position, where \( x = D \times \cos \theta, y = D \times \sin \theta \). The state remains the same within each episode, and is chosen randomly across episodes (i.e., random angle \( \theta \)). Note that, the number of possible target locations can be set according to the task to be modeled.

- **Action** \( A \) : An action, \( a_t \), is taken at each time step \( t \). On each step the agent decides where to saccade to on the display. An action denotes as \( a = [x, y] \) where \( x, y \in [-1, 1] \). The aim saccade is corrupted by the ocular motor noise. Specifically, the actual landing position after the saccade is sampled from \( \tilde{a}_t \sim N(a, \sigma_{ocular}(t)) \). The ocular motor noise is linearly dependent on the saccadic amplitude [23]: \( \sigma_{ocular}(t) = P_{ocular} \times \text{Amplitude}(t) \), where \( P_{ocular} \) is a hyper-parameter of the model that we chosen by fitting to human data (Table 1).

- **Reward function** \( R(S,A) \): At each time step \( t \), the environment (in one of the states \( s_t \)) generates a reward (cost if the
value is negative), \( r(s_t, a_t) \), in response to the action taken \( a_t \). The reward is 0 if the agent reaches within the target region. The reward is \(-1\) for each extra saccade made. Maximizing this reward regime encourages the model to find the target as fast as possible given the constraints stated above.

- **Transition function** \( T(S_{t+1}|S_t, A_t) \): Another consequence of the action is that the environment switches to a new state according to the transition function. In the current task the states (i.e. target positions) stay unchanged across time steps within one trial. Therefore, \( T(S_{t+1}|S_t, A_t) \) equals to 1 only when \( S_{t+1} = S_t \). \( T(S_t + 1|S_t, A_t) \) equals 0 otherwise.

- **Observation** \( O \) and **observation function** \( O = f(S,A) \): After taking the action (i.e. saccade to and fixate at a new position on the display), a new observation is received, which is a function of state and action \( o_t = f(s_t, a_t) \). The observation of the target position is dependent on the true target location (state) and the current fixation location (action).

- **Discount rate** \( 0 \leq \gamma < 1 \). The model receives a scalar reward at each time step, \( r(s_t, a_t) \). The optimal strategy is the one that maximizes the expected long-term reward:

\[
E[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)],
\]

given the constraint defined above.

### 3.2.1 Belief update

The state of the environment (i.e., the target position) is not directly known to the model. For this reason, the model maintains a belief \( b(t) \) about the state given the information observed so far. Specifically, at each time step \( t \), the agent takes an action \( a(t) \) and observes \( o(t) \), the belief of the current state is updated by integrating the previous belief at \( t - 1 \), \( b(t - 1) \), and the new observation at time step \( t \), \( o(t) \), using Bayes rule (a Kalman filter) \([11, 40, 50]\).

The belief is initially assumed to be a uniform distribution over all possible states. That is, without any evidence, the model believes that the target could be anywhere on the display. After taking an action (i.e., fixating at one location), the model receives a noisy estimate of where the target is (see Observation function above). Specifically, after the first time step \( (t = 1) \), the belief, \( b(t) \), and uncertainty, \( \sigma_b^2(t) \), of the target position is solely based on the first observation, \( o(1) \). The belief update from \( b(t - 1) \) to \( b(t) \) given the new observation \( o(t) \) is shown in Equation (1) below.

\[
\begin{align*}
b(t) &= b(t - 1) + K_t \times [o(t) - b(t - 1)] \\
\sigma_b^2(t) &= \sigma_b^2(t - 1) - K_t \times \sigma_b^2(t - 1)
\end{align*}
\]

where,

\[
K_t = \frac{\sigma_b(t - 1)^2}{\sigma_b^2(t - 1) + \sigma_o^2(t)}
\]

### 3.3 Implementation

#### 3.3.1 Workflow

Some of the parameters were determined from the psychological literature. Specifically, the perceptual noise parameter was set to a value from \([42]\) \((\rho_{\text{spatial}} = 0.09)\). In contrast, the ocular motor noise and jitter parameters were fitted to human behaviour using grid search.

Other parameters were set using the workflow in Figure 3. To do so, we first generated a set of agent parameter values (ocular motor
noise and perceptual noise) using grid search. In parallel we generated artificial training tasks given the environment parameters (left hand side of Figure 3). Details of how tasks were generated is given in the state description of Section 3.2. An important feature of the workflow is that we did not train on a large database of human data. The parameterized agent is trained on the sampled tasks (“Train the agent” in Figure 3). The trained agent is tested using a separate set of tasks generated by the state. Subsequently, the behaviour of the trained agent is compared to summative human data and the results recorded. The summative data consisted of the number of saccades for a single width/distance combination. The process was repeated until an acceptable fit to the summative human data was found.

A key property of this workflow – that distinguishes our work from much ML work on human behaviour – is that the agent is not trained on human data. Rather, it is trained on tasks sampled from an artificial environment.

### 3.3.2 Datasets.

We compare the model’s behaviour to human data collected in empirical studies of gaze-based selection published at CHI [56, 72]. Due to limited space, selective details about the experiments are given below.

In [56] participants were asked to make eye movements from a starting location to a circular target. Starting locations were presented pseudo-randomly in nine of one positions. This circular target was presented at a 5° or 10° away from the starting location at a random angle on a semicircle oriented towards the center of the screen (see Figure 2 in [56]). There were targets with seven different sizes. The diameters of target circles W are 0.5°, 1°, 1.5°, 2°, 3°, 4°, 5°. Trials were presented in blocks of 28 (7 target sizes × 2 saccade amplitudes × 2 repetitions). The gaze input device was an Eyelink 1000 plus (SR Research, Ottawa, ON, Canada) video-based eye tracker at sampling rate of 1000 Hz. Calibration accuracy in [56] ranged from 0.5° to 1.30° degrees (mean=0.50, median=0.47). Therefore, we did not model the condition with target size of 0.5°.

In [72] Experiment 1, each trial starts with fixating at a start button for 450ms; the start button randomly appears at one of the predefined positions. Then, the button group, including five circular buttons arranged as a ‘+’ shape, appeared at a given distance. The participants were asked to look at the center area of the desired target, which was arranged at the center of the group, as quickly as possible (Figure 2 in [72]). The target was 6.13° away from the start button. The diameters of target circles were: 1.23°, 1.72°, 2.15°, 2.64°, 3.07°. The gaze input device was the eye tracker of EyeLink II. The pupil-only tracking mode at the sampling rate of 250 Hz was used.

### 3.3.3 Hyperparameters.

We built the gaze-based selection task as an OpenAI gym like environment and used the Stable Baselines3’s implementation of the deep RL algorithm, Proximal Policy Optimization to train the model.1 Hyperparameters were as follows: Horizon=500, Clipping=0.15, Gamma=0.99. Other hyperparameters for RL model are defaults as in Stable Baselines implementation. The hyperparameters for the human bounds used in the model is provided in Table 1.

---

1https://github.com/DLR-RM/stable-baselines3

---

### Table 1: The hyper-parameters used in the model. These parameters were chosen based on fitting to one of the D/W combinations in each of the two data sets [56, 72].

<table>
<thead>
<tr>
<th></th>
<th>$\rho_{\text{spatial}}$</th>
<th>$\rho_{\text{ocular}}$</th>
<th>$\rho_{\text{drift}}$</th>
<th>Latency (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zhang 2010 [72]</td>
<td>0.09</td>
<td>0.08</td>
<td>0.33</td>
<td>260ms</td>
</tr>
<tr>
<td>Schuetz 2019 [56]</td>
<td>0.09</td>
<td>0.08</td>
<td>0.33</td>
<td>150ms</td>
</tr>
</tbody>
</table>

---

**Figure 4:** Number of saccades required to reach the target (y-axis) against parameter value for ocular motor noise ($\rho_{\text{ocular}}$, each colored line) and spatial uncertainty in target location estimate ($\rho_{\text{spatial}}$, x-axis). Higher values indicate more noise. These two parameters are defined as follows. (1) **Ocular motor noise** $\rho_{\text{ocular}}$: Given the aim fixation $a$, the actual fixated position is sampled from $N(a, \sigma_{\text{ocular}})$, where $\sigma_{\text{ocular}}$ is amplitude dependent, $\sigma_{\text{ocular}} = \text{Ocular} \times \text{Amplitude}$. (2) **Visual spatial noise** $\rho_{\text{spatial}}$: Given the true target position $s$, the estimate of the target location is: $\sigma_{t} \sim N(s, \sigma_{o})$, where $\sigma_{o}$ is eccentricity dependent, $\sigma_{o} = \rho_{\text{spatial}} \times \text{Eccentricity}$.

### 4 RESULTS

We compare our model results quantitatively to human data previously published at CHI [56, 72]. Before comparison with human data, we first show how the number of saccades required to reach the target is influenced by changes in visual spatial noise ($\rho_{\text{spatial}}$) and ocular motor noise ($\rho_{\text{ocular}}$). As shown in Figure 4, the number of saccades increases as the spatial noise and ocular motor noise increase.

#### 4.1 The effect of target size on selection time

The selection time in gaze-based selection tasks is usually composed of the time to search for the target (Eye Movement Time), and the
Figure 5: Eye movement time (EMT) and selection time plotted against target size. EMT is the time between when the target appears and when the gaze first enters the target region. Selection time is EMT plus the Dwell time. Data (the bars) are from Figure 3 in [72]. The model prediction of EMT (black triangles) and selection time (red circles) are plotted over human EMT (yellow strip bar) and human selection time (blue bars).

dwell time to commit the selection (Dwell time). The dwell time must be at least as long as the dwell criterion. Next, we compared the model’s predicted selection time with the data reported in Zhang 2010 [72]. In Zhang’s experiment, participants were asked to make eye movements from a starting location to a circular target as fast as accurate as possible. There were five different target sizes (Figure 5, x-axis). A dwell time of 800 ms was used to commit the target selection. That is, a continuous 800 ms within the target region is required to make the target selection. The dwell time was recalculated when the gaze fell out and re-enter the target region.

As shown in Figure 5, there is an exceptional correspondence between the model predictions and the data (Root mean squared error (RMSE) between the model and the data for the Eye movement time (EMT) is 29.35 ms. For the selection time, RMSE is 54.50 ms). The EMT is the duration between when the target appears and when the cursor first enters the target region. This is achieved by the primary saccade and possibly one or two corrective saccades. For the model, the number of the saccade required is adaptive to the visual spatial uncertainty ($\rho_{\text{spatial}}$), the ocular motor noise ($\rho_{\text{ocular}}$), and the task properties (size W and distance D. Figure 4). Specifically, parameter $\rho_{\text{spatial}}$ was chosen based on previous literature ($\rho_{\text{spatial}} = 0.09$ [42]). Parameter $\rho_{\text{ocular}}$ were calibrated to EMT of only one of the target size and distance combination ($D = 6.13, W = 3.07$), see Table 1). The predicted EMTs were then calculated for all other target sizes. The selection time increased with smaller target sizes. This shows the model was able to capture the extra selection time required for greater precision.

Figure 6: Number of saccade plotted against target size. The data (black circles) are aggregated from a 10° distance condition and a 5° distance condition as reported in [56]. The models performance is separated by target distance (5° orange squares), and (10° blue diamonds), and the average of these two (black triangles).

4.2 The effect of target size on dwell time

In Figure 5, the numbers on the bars denote the difference between EMT and the selection time in the human data (i.e., dwell time). For the smallest target (1.25°), human participants took 1075 ms to make the selection after the first entry in the target region which is a full 275 ms longer than the dwell criterion (800 ms). The drift parameter $\rho_{\text{drift}}$ was calibrated to the smallest target size (Table 1). The dwell times of other target sizes were then calculated. The actual dwell time was decreasing and gradually close to the given dwell time of 800 ms with the increase of the target size.

4.3 The effect of target size on the number of saccades

Next, we compared the model’s predicted performance time with the data reported in Schuetz et al., 2019 [56]. In Schuetz’s experiment, participants were asked to make eye movements from a starting location to a circular target. This circular target was presented at a 5° or 10° away from the starting location at a random angle. There were targets with six different sizes.

As shown in Figure 6, the number of fixations made by participants [56] increased as target size decreased. The predictions of the model are plotted for $D = 5°$ (orange squares) and $D = 10°$ (blue diamonds) in the figure, along with the average data across two conditions (dashed line). The RMSE between the model (black dashed line) and data (black solid line) is 0.13 saccade. Schuetz [56] did not report the number of saccades for each condition separately. Figure 6 also shows that the model over predicts the number of fixation required for smaller targets. Human participants exhibited fewer fixations especially at target size 1°. We suspect that this is due to the fact that the calibration accuracy of the eye-tracker
Figure 7: *Left panel:* Selection time data from [56]. The selection times (black circles) are plotted against the target size (aggregated across a 5° distance condition and a 10° distance condition. This is how it was reported in [56]). The colored lines show the saccade end times for the first, second, and third saccade (see figure legend). The human data is from Figure 3 in [56]). Error bars show +/- 1 SE. *Right panel:* Model prediction of the selection time. Color lines are the saccade end times for the first, second, and third+ saccade. Worth noting that, the third and plus saccade was rare, only happened in a very small proportion of the total trials (as shown in Figure 8 below).

Figure 8: The proportion of trials complete with one, two and three saccade(s) respectively (see the legend), with different target sizes and distance ranged from 0.25-1.30 [56], which suggests that it might be inadequate to capture the small corrective saccades that are likely to be even smaller than the target diameter (1°) (see more on this in the discussion).

4.4 The effect of target size on the saccade duration of main saccades

We also compared the model’s predicted selection time with data from [56]. The model again does a good job in matching this human selection time. Moreover, as shown in Figure 7, while the selection times (black circles) increases with smaller target sizes, the individual saccade time is independent of the target size (the colored lines in Figure 7). In other words, the increase in selection time is due to the higher frequency of needing secondary corrective saccades in the smaller target sizes (as shown in Figure 8). It is more likely to make the secondary saccade with smaller and further targets. In general, dwells with more than one secondary saccade are reported to be rare [70]. Both the model and human data (Figure 7) only did the third saccade for target sizes of and below 2°, and this was only true for a very small proportion of the total (Figure 8).

4.5 The effect of Index of Difficulty on the number of saccades

Studies have shown that saccadic sequences increase in length with Index of Difficulty ($ID = \log_2(2D/W)$), in agreement with Fitts’ law [43, 56, 68–70]. Our model showed that this is mainly due to the use of error-correcting secondary saccades that have a frequency which increases with target distance and decreases with target size (Figure 8).

In contrast, to Fitts’ Law, which predicts movement time being a linear function of ID, our model predicts that when ID is below a certain value (around 2 bits), the selection time is independent of ID. This is due to the fact that for these easier selections, the secondary saccade is no longer needed, and the model often uses only a single saccade to reach the target. Figure 9 shows that there is an inflection point in both model and data at about ID=2bits. Similar findings can be found in [33, 68].

Figure 9: Data in the right panel (recreated from [56]): Number of saccades as a function of Index of Difficulty (ID) ($ID = \log_2(2D/W)$). Error bars show +/- 1 SE.

5 DISCUSSION

In summary, we have reported a computational model that predicts not only aggregate-level data like selection time distributions but step-by-step gaze trajectories in gaze-based selection. The key assumption is that the control policy is optimised for utility (selection time) bounded by limits imposed by the visual and motor system. Formulated like this, the resulting POMDP problem can be solved using modern reinforcement learning methods, such as Proximal...
Policy Optimization. The elegance of this approach is in the fact that eye movements emerge from the assumption of boundedly optimal control. Unlike in some previous models, no hand-written rules are needed to prescribe how eye movements are placed.

The model produces human-like performance. It accurately reproduces five key gaze-based selection effects, including the effect of target size and distance on the number of fixations and selection times. The direction and shape of the predicted curves are generally accurate across a range of parameters. As a consequence, we believe that the model is the first to offer a deep account of the underlying processes by which humans perform gaze-based selection. Importantly, the model does not have to be fitted to the whole data set, rather it can be calibrated to just a single data point and then the others can be predicted.

One feature of the model is its strong scientific basis. This is evident in the empirically grounded parameter space: All three primary variables – spatial noise, ocular motor noise, and jitter – are directly observable aspects of human behaviour. Mathematically, it might be possible to use one, instead of two separate parameters (spatial noise and ocular motor noise), and doing so might provide the same range of model behaviours and equally good model fits but with a more parsimonious model. However, there are strong scientific reasons for keeping two parameters as each parameter models a separate, and independently observable, physical property of the user. An advantage therefore of a scientific model is the capacity to verify each parameter by observation. Using separate parameters also prepares the ground for the exploration of further theoretical ideas. For example, motor noise is known to have larger end-point error in the direction of movement [19]. Therefore, an alternative noise profile could be applied, and we would be able to investigate the adaptive behaviour to this new assumption.

One extension of the model might be to make jitter adaptive. As we have said, when holding a fixation on small targets, our eyes do not stay perfectly still, but instead keep moving around the focal area, therefore possibly leaving and reentering the target area several times. Fixed non-adaptive jitter is a feature of the model reported above. Interestingly, however, [72] found that the range of jitter is adaptive to the target size. That is, when the target is big, the participants allow their eyes to ‘loosely wander’ within the target region while waiting for the minimum dwell duration to be reached. However, when the target is small (though not too small), the participants are able to adjust the jitter range accordingly. When the target size is close to the bounds of the eye’s minimum jitter limit, the selection time is often increased significantly because the eyes are not able to hold on the target for long enough to reach the selection criterion.

Another idea worth further exploration concerns location uncertainty in human peripheral vision, which appears to be biased toward the current fixation [65]. When asked to estimate a target position in the periphery, instead of Gaussian noise around the true target position, the estimate tends to be skewed toward the current fixation. This skewed estimation is sometimes considered as the reason why the primary movement towards the target tends to undershoot [65]. However, these undershoots are sometimes attributed, by others, to the fact the there is bigger motor cost for the motor system to make a correction to the primary movement with a a movement that involves a reversal of direction [13]. Excitingly, the adaptive reinforcement learning (RL) framing of our model provides a platform for the investigation of these effects. Doing so would provide a better understanding of user behaviour during gaze-based interaction.

Importantly, generative models of eye movements, like ours, may inform the design of better selection techniques for gaze-based interaction. Predictive models in HCI have traditionally had a role in guiding the evaluation of prototypes. The proposed model can be used to assess the effect of changes in spatial aspects of a layout (e.g., distances, sizes) as well as the qualities of the eye tracker (e.g., latency). HCI models have also been used to derive design guidelines. However, our model does not make a simplistic design recommendation like ‘make all targets this big’. Rather, it allows systematic study of how the involved factors jointly shape optimal design. Finally, models have been used for optimization. Heuristic optimization methods could be used to identify optimal combinations of design parameters for given conditions [47]. An important application area is ability-based optimization [55], where the goal is to computationally improve designs for users with impairments. The proposed model could be used to simulate, for example, impairments of the vision system, by changes in parameters like noise or drift. Modern machine learning methods like Approximate Bayesian Computation (ABC) could be used to fit selected parameters to gaze data obtained from an individual [31].

As mentioned, the reported model can be tuned using a single target size and distance combination. Theoretically, it does not matter which combination to use. However, in practice, due to the limited precision of the eye trackers and the way that saccades are determined, the number of saccades identified in human data tends to be less accurate for cases with really small target size. In general, saccades are determined based on a velocity and/or a acceleration criteria. A small travel from a point A that is just outside the target region to a point B that is very close to point A but within the target region, is likely to be ignored in human data analyse. In contrast, for the model presented in this paper, without a constraint on the precision, it is possible to measure vanishingly small eye travel and count it as yet more saccades. We suspect that this is the reason that the model prediction on the number of saccades tends to be greater than data in the smallest target size (1°) in Figure 6 and ID+4 bits in Figure 9. A better fit between the model and data might be achieved if the precision of the model is carefully calibrated to the precision of eye trackers used and specific methods used to determine the saccade.

6 CONCLUSION
In conclusion, we have reported a model of gaze-based selection that adapts eye movement strategies to target size and width, as well as to noise in the motor and vision systems. The model predicts a number of the key empirical findings in the literature and provides a useful basis for guiding further research.

ACKNOWLEDGMENTS
We gratefully acknowledge the Academy of Finland projects BAD and Human Automata.
REFERENCES


An Adaptive Model of Gaze-based Selection


