Generation of Suspicious Behaviour

Peta Masters[1], **Daniel Gallagher**[2], **Gal A. Kaminka**[3] **Mor Vered**[2]

[1]King's College London, Bush House, Aldwych, London, United Kingdom
[2]Monash University, Exhibition Walk, Clayton, VIC, Australia
[3]Bar Ilan University, Ramat Gan, Israel
peta.masters@kcl.ac.uk, dgal0013@student.monash.edu, galk@cs.biu.ac.il, mor.vered@monash.edu

Abstract

Patterns of behaviour are regarded as suspicious when they deviate from what is normal or expected for a given domain, where the motive is unknown and suggestive of risk. Recognising suspicious behaviour is valuable because it enables effective threat detection and response but in order to train machine learning models and security personnel to identify such behaviour, we must first define it and provide examples from which they can learn. Suspicion is a human construct, a property of the observer rather than of the behaviour itself, therefore we base our approach on goal recognition, which focuses intrinsically on the observer's perspective. We define and demonstrate several types of suspicious behaviour, synthesising plans for *loitering*, *obfuscation*, and *deception*. These reflect the behaviour of an agent that believes it may be being observed, and acts to conceal its intent. We evaluated the generated plans in experiments with human subjects across several domains and found them significantly more suspicious than the baselines, without being obviously adversarial. Our generated Loitering behaviour was deemed suspicious for the longest across all domains.

Introduction

Suspicious behaviour is commonly understood as behaviour that gives rise to uncertainty about an agent's motives and suggests the possibility of hidden, potentially harmful intent (Fein, Hilton, and Miller 1990). The uncertainty resides in the mind of the observer. Suspicion is therefore not a property of behaviour alone but of the observer's cognitive and affective state: the same sequence of actions may appear innocuous to one observer and highly suspicious to another, depending on their expectations, prior knowledge, and tolerance for risk. In this work, we examine suspicious behaviour from the perspective of planning-based goal recognition, where the observer's point of view is intrinsic to the problem formulation: observations are the input; the output is an inferred goal or a probability distribution over possible goals (Masters and Vered 2021).

Suspicious behaviour matters because the stakes of failing to recognise it in time are so high. In an international climate characterised by terrorism, geopolitical instability, and social unrest, concern about encountering "bad actors" in everyday settings, such as airports, hospitals, warehouses and online platforms, is increasing. Human operators and automated systems, including CCTV analytics and autonomous monitoring agents, are deployed to recognise and warn of suspicious behaviour, typically using anomaly detection (Sodemann, Ross, and Borghetti 2012). But for the systematic recognition of suspicious behaviour—whether by humans or AI—it must first be defined in a form amenable to formal modelling and empirical evaluation. Few labelled datasets capture the subtle, often pre-criminal stages of suspicious behaviour. This scarcity motivates the need for methods to generate realistic examples of suspicious and non-suspicious behaviour: (a) to test and compare algorithms for goal and threat recognition; and (b) to train human surveillance personnel and machine learning models.

To address this need, we present a formal framework for the synthesis of suspicious plan generation, grounded in classical planning and model-based goal recognition. Within this framework, we focus on three behavioural types that, drawing on definitions from social psychology, can reasonably be expected to trigger suspicion in a human observer. We label these types: *obfuscation*, *loitering* and *deception*, characterised by: uncertainty and ambiguity—actions compatible with multiple goals, some apparently benign, others harmful; non-rationality—patterns of behaviour that deviate from standard rational planning assumptions, such as inefficient or seemingly aimless activity; and adversarial behaviour—actions chosen to deliberately mislead or obstruct recognition. All three behaviours involve risk, operationalised as activity in the vicinity of goals deemed dangerous by the observer. To actualise deception, we offer the *shoe-tying* algorithm, a novel deceptive plan that uses consecutive detours to mislead the observer.

We evaluated the framework in a controlled user study. Participants observed video clips of agents operating in a range of decision-making domains and were required to judge, in real time, whether and when behaviour became suspicious and for how long it remained so before suspicion gave way to fear or certainty. We found behavioural type to be the primary factor in identifying a plan as suspicious. Behaviours intended to arouse suspicion via obfuscation and deception often succeeded in delaying it, even when agents approached risky goals closely; while loiter-

ing was suspicious for the longest. These results generalise across multiple domains. They provide empirical support for our formalisation of suspicious behaviour and highlight both the promise and the challenges of modelling suspicion as an observer-centred, goal-recognition based planning problem.

Background

No behaviour is inherently suspicious; suspicion resides in *the mind of the observer* (Levi and Loeben 2004). Intuitively, it is unusual, unpredictable or recognisably bad behaviour that an observer believes may, if carried to conclusion, constitute a threat. As such, suspicious behaviour is highly *context-dependent*, subject not only to the environment but to the mental state of the observer and to the observer's impression of the observed (Skolnick 2011).

We therefore approach the problem of generating suspicious behaviour as a process of sequential action selection, guided and constrained by a goal recognition process that considers how selected actions will be interpreted by an observer. We rely on planning-based goal recognition (Ramírez and Geffner 2010; Masters and Vered 2021; Meneguzzi and Pereira 2021), in which a domain model (that specifies which states can be achieved, which actions are available, and the cost of actions) is used with a planner to generate plausible goals given observed actions. The output of the recognition process is either the most likely goal (or set of goals) or, more generally, a probability distribution over all known goals in the domain (Masters and Vered 2021; Meneguzzi and Pereira 2021). In generating actions that form a suspicious plan, the action selection process can take this probability distribution into account, using the same domain model. This allows it, in principle, to control two key aspects of suspicious behaviour: perceived risk and the observer's uncertainty with respect to intent.

Risk For a behaviour to be deemed suspicious, there must be an element of *risk*, a sense that the anticipated outcome is likely to be "malevolent or undesired" (Deutsch 1958, p.267). This notion has been utilised in past research on recognising behaviour that may be risky to the observer (Tambe and Rosenbloom 1995) and, more generally, where expected cost to the observer is significant (Avrahami-Zilberbrand and Kaminka 2007). The solution to a planning problem involves achieving a goal. The solution to a goal recognition problem involves determining which goal—from a set of all known goals—an agent or system is trying to achieve. The solution to a suspicious behaviour generation problem requires recognition and exploitation of *the particular goal or goals that the observer believes may result in negative consequences if achieved* (Avrahami-Zilberbrand and Kaminka 2014). To meet this need, we employ the concept of 'risky' goals, a subset of all known goals: in a given context, ***risky goals*** represent 'bad things that might happen'.

Uncertainty "To be suspicious is to be uncertain, but uncertain in a special way...[about] motivation" (Fein, Hilton, and Miller 1990, p.763). With respect to suspicion, the human observer is attempting goal recognition. It follows that plans which result in uncertainty for a goal recognition system will similarly result in uncertainty for a human observer. Thus, in prescribed domains and in the context of goals perceived to be risky, it is plans with the following characteristics that are most likely to be suspicious.

Ambiguity. If observations are equally suggestive of multiple possible goals—a cook boiling water, for example, which might be preparatory to cooking the pasta or boiling an egg—there may be no way for a goal recognition system to determine which goal applies until the next (or sometimes last) step in the process (Keren, Gal, and Karpas 2014). Where behaviour seems deliberately ambiguous a human may suspect ***obfuscation***.

Non-rationality. Typically, goal recognition systems operate under a 'rationality assumption'—that observed behaviour is more or less rational (Vered, Kaminka, and Biham 2016; Vered and Kaminka 2017; Kaminka, Vered, and Agmon 2018). When behaviour is *non*-rational, goal recognition is only able to flag the problem; it cannot offer a solution (Masters and Sardina 2021). Non-rational behaviour is seemingly random and undirected, In proximity to a risky goal, it suggests ***loitering*** with intent.

Adversarial. Goal recognition has largely focused on 'keyhole recognition', where agents are unaware of being observed (or aware but choosing not to modify their behaviour). This is also true of prior work on suspicious behaviour recognition (Avrahami-Zilberbrand and Kaminka 2014). A second mode of recognition is called 'intended recognition', where agents are aware of being observed and deliberately signal their intent (Carberry 1996). A third type, identified by Carberry as 'adversarial recognition', has always presented a problem. Here, the agent deliberately obstructs the recognition process in order to conceal its intent and, by manipulating what is seen, achieves a goal that the observer was not expecting. The difficulty is that almost any behaviour may turn out, retrospectively, to have been adversarial: successful ***deception*** may only become apparent after the unexpected goal has been achieved, if ever.

Based on the above, we hypothesise that in prescribed domains and in the context of risky goals these three behaviours are likely to be regarded as suspicious by a human observer. We next define them in the context of a suspicious behaviour generation problem (SB) built on goal recognition in a planning domain.

Technical Framework

Building on a combination of a classical planning problem (Russell and Norvig 2015) and model-based goal recognition (Masters and Vered 2021), we assume an environment that is static, discrete, and fully observable in which a planning domain D is a tuple (S, A, δ, c), where: S is a finite set of states; A is a finite set of actions; $\delta : S \times A \to S$ is a deterministic state transition function; and $c : A \to \mathbb{R}$ is the cost function over actions.

The suspicious behaviour generation problem is a tuple $SB = (D, G, P, s_0, \mathcal{G})$, where:

- $D = (S, A, \delta, c)$ is a planning domain;
- $G \subset S$ is a set of known goals;
- $\mathcal{G} \subseteq G$ is a non-empty subset of 'risky' goals;
- P denotes a prior probability distribution across G; and
- s_0 is the initial state.

The solution to SB is a suspicious plan. Generally, a plan $\pi = a_1, ..., a_n$ is a sequence of actions that transform one state to another and $\vec{s} = s_1, s_2, .., s_n$ is a corresponding sequence of achieved states. The cost of a plan is the sum of its action costs. Since we treat suspicion as residing in the mind of the observer, cost may be calculated according to any appropriate heuristic. We therefore use $cost(s, s')$ to denote the *estimated* optimal cost of transforming state s to state s'. Moreover, we refer to the trace of a plan as observations $\vec{o} = o_1, o_2, .., o_n$ where $o_i = (a_i, s_i)$. That is o_i is an action-state pair which represents the complete information available to an observer at each step of a plan in a fully observable domain.

Note that a suspicious plan does not terminate at the risky goal (or at a state that *entails* a risky goal): suspicion pertains only while the goal remain uncertain. A terminating condition for the length of the plan may be set by specifying a maximum cost or maximum plan length. Where costs are uniform (i.e., the cost of each action $= 1$), these values are the same. Thus, the solution to SB is a plan π that starts at s_0 and proceeds for n actions in such a way that π is deemed *suspicious* by a human observer. We hypothesise that plans which conform to the definitions for obfuscating, loitering, and deception given below are three such plans.

Terms and Definitions

Since suspicion depends on the perception of risk, behaviours are deemed suspicious only when they occur 'in proximity' to a risky goal. Our definition of ***proximity*** draws on the notion of a radius of maximum probability (Masters and Sardina 2017b, 2019) being the lower bound for the cost-distance from goal within which that goal is guaranteed to be the most probable, which may be calculated for any goal $g \in G$, as follows.

$$\text{RMP}_g(s_0, G) = \min_{g' \in G \setminus \{g\}} \frac{cost(g, g') + cost(s_0, g) - cost(s_0, g')}{2} \quad (1)$$

The advantage of this formulation is that it is based on cost: no probabilities need to be calculated. Moreover, the approximate cost from initial state to each of the goals is often known or readily determined.[1]

Where needed, a probability distribution across goals is calculated in accordance with the principled approach from Ramírez and Geffner (2010), adapted with reduced complexity for fully observable domains by Masters and Sardina (2017a; 2021).

$$Pr(G|s_n) = \alpha \cdot \frac{1}{e^{(cost(s_n, g) - cost(s_0, g))}} \quad (2)$$

where α is a normalising constant and s_n is the most recently achieved state.

Suspicion is aroused when the observer cannot determine what the target is doing or intends to do. Broadly, there are two behavioural types that give rise to this condition. Either the behaviour seems to be targetted—although the observer is unable to determine exactly what the target is—or untargeted seemingly with no particular intent in mind, which may leave the observer wondering why the behaviour is taking place.

Definition 1. ***Directed behaviour*** *is a plan the trace for which is a sequence of observations $\vec{o} = o_1, o_2, .., o_n$ $o_i \neq o_j$ for all $i, j \in [1, N]$, $i+1 > j < i-1$ where $o_i = (a_i, s_i)$ and a_i is not null. Behaviour that is not directed is* ***undirected****.*

In words, in directed behaviour no state is repeated and action is continuous: it is seemingly purposeful behaviour that appears to be targeting some goal, even if that goal is unknown. Undirected behaviour, on the other hand, may repeat states and may be discontinuous. An agent engaged in undirected behaviour may be stationary or wandering about.

With the above in hand, we now define the three suspicious behaviours with which we are concerned.

Obfuscation is behaviour indicative of an agent that may be deliberately masking its intent to approach a risky goal by seeming, equally, to target some other goal.

Definition 2. *Given a probability distribution across goals $Pr(G|s_n)$ and a risky goal $g^* \in \mathcal{G}$, the most recent state s_n is* ***obfuscatory*** *iff there exists $g \in G \setminus \{g^*\}$ such that $Pr(g^*|s_n) \approx Pr(g|s_n)$.[2] An action a_n that results to an obfuscatory state s_n is* ***obfuscation****.*

In words, obfuscation occurs when the probability of a risky goal is made approximately equal to the probability of some other goal. A deliberately obfuscatory plan includes as much obfuscation as possible.

Definition 3. *Let Π_{sg} be the set of all plans that begin at s_0 and terminate at a real risky goal g^* and let A^*_π be the set of all obfuscating actions in a plan π. A plan $\pi' \in \Pi_{sg}$ is* ***optimally obfuscatory*** *iff $|A^*_{\pi'}| \geq |A^*_\pi|$ for all $\pi \in \Pi_{sg} \setminus \{\pi'\}$*

Loitering is undirected behaviour in proximity to a risky goal. The nearness to goal means that rather than wandering, a loiterer seems to be 'hanging about'. Thus loitering behaviour implies a sequence of states such that the feared goal remains always within reach but never beyond doubt.

Definition 4. ***Loitering*** *is undirected behaviour (as defined at 1) $\pi = a_1, a_2, .., a_n$ where there exists a $g \in \mathcal{G}$ such that a_i gives rise to a state s_i where $cost(s_i, g) \approx \text{RMP}_g$ for $i \in [1, n]$.*

Broadly, a plan may be regarded as deceptive so long as its real goal is unknown to the observer. Expressed in terms of probability, if there is some other goal as probable or more probable than the real goal, then deception is occurring.

[1] If $cost(g, g') = \infty$, it may be approximated (see Discussion).

[2] Clearly, $Pr(g^*|s_n) - Pr(g|s_n)$ should be minimised but is implementation-specific.

Definition 5. *Given a deceiver's real goal $g_r \in G$, the most recent state s_n is **deceptive** iff there exists a goal $g \in G \setminus \{g_r\}$ such that $Pr(g|s_n) \geq Pr(g_r|s_n)$. An action a_n that results in a deceptive state is a **deception**.*

As for obfuscation, a deliberately deceptive plan includes as much deception as possible. Moreover, successful deception requires that the real goal becomes apparent to the observer *as late as possible*, if ever.

Plan Generation

Plans for obfuscation and loitering are synthesised by assembling sequences of actions and—where necessary—using goal recognition to test the resultant plans against Definitions 1 to 4 above (the RMP can be calculated independently of probabilities). For deception, we introduce a novel algorithm that conforms to the above definition but which does not depend on goal recognition for its synthesis.

Shoe-Tying Algorithm. The conundrum of deceptive planning is that the plan must achieve its goal without ever seeming to advance towards it. Numerous authors have grappled with the problem in the context of navigational domains (the archetypes for full observability) notably Dragan, Holladay, and Srinivasa (2014), who defined ambiguous, switching and exaggerated plans, Keren, Gal, and Karpas (2015) who demonstrate environments that support obfuscation, and Kulkarni, Srivastava, and Kambhampati (2019) whose plans cleverly balance obfuscation with legibility.

Algorithm 1 depends on the insight that behaviour can appear to be purposeful even if its goal is unknown provided it conforms to the *directed* behavioural type (as defined at Def. 1). The shoe-tying algorithm is so-called because it simulates behaviour familiar from films and TV where a sleuth tailing a suspect must find an excuse for their presence so stops to look in a shop window or tie their shoelace. That is, at intervals, it arbitrarily selects a non-goal state $r \notin G$ but treats it *as if* it were a goal by adopting directed behaviour towards it such that $Pr(r) \geq Pr(g_r)$ (Def. 5) until the RMP is reached. Thus, our algorithm tackles the two sides of the deceptive planning problem: it seems not to be approaching the real goal (a risky goal in our framework $g_r \in \mathcal{G}$) yet is guaranteed to reach that goal.

Algorithm 1: Shoe-Tying Plan Generator

procedure SHOETIE($D, g_r \in \mathcal{G}, RMP_{g_r}$)
  $\pi \leftarrow [\,]$ ▷ *Initialise plan to empty.*
  $s \leftarrow$ current state
  while COST$(s, g_r) > RMP_{g_r}$ **do**
    $\pi^* \leftarrow$ OPTIMALPLAN(s, g_r)
    $\pi_1 \leftarrow$ first half of π^*
    $\pi \leftarrow \pi + \pi_1$
    $s \leftarrow$ ENDOFPLAN(π_1)
    $r \leftarrow$ RANDOMSTATE(D, G)
    $\pi_2 \leftarrow$ OPTIMALPLAN(s, r)
    $\pi \leftarrow \pi + \pi_2$
    $s \leftarrow$ ENDOFPLAN(π_2)
  return π

Methodology

We recruited 30 participants through Prolific, an online data-collection platform. 33% of the participants identified as female, 63% as male, and 3% as 'Prefer not to say'. Participants were all over the age of 18, and the average age was 44.77 (SD = 13.21). All participants reported high proficiency in English and were recruited from English-speaking countries. Each participant completed a balanced set of 15 tasks (see below). In total, 430 valid instances were recorded. Participants were compensated £3 and the average completion time was 33 minutes (SD = 10.32).

Experimental Task

Participants were tasked with observing a series of agent behaviours across a set of decision-making tasks and determining whether to do nothing, increase alertness or stop the agents. For each task the agent pursued one of several possible goals. Some goals were beneficial while other goals were dangerous or malicious. The agent's true intention—whether helpful or malicious—was not revealed to the participant. Participants were given the opportunity to intervene and/or stop the agent at any time.

During each video, participants could press a button labelled "Suspicious", which would "notify security". After pressing this button, they were also able to press a "Stop" button that would immediately halt the agent's behaviour. At any time, participants could alternatively press a "Safe" button if they were confident that the agent was not acting maliciously, thereby concluding the activity. Each button could be pressed only once per video and could not be reversed.

Each participant was shown three domains in a random order. Within each domain, there were five behaviour types and five distinct problems, resulting in 25 possible video combinations (5 behaviours × 5 problems). For each participant, five videos were shown per domain—one for each behaviour type—such that each behaviour was paired with a different problem. The specific problem–behaviour pairings were randomly selected.

Before each domain was presented, participants were provided with an explanation of the domain rules, a short demonstration video, and an explanation of their task (pressing the buttons) see Figure 1. The duration for which each button was pressed during the 15 video instances was recorded and later mapped to the corresponding agent actions. Following each video, participants were also asked to provide brief, optional, written responses explaining their decisions regarding button presses.

Prior to the experiment, participants completed a consent form and a demographics survey, which collected information on age, gender, English proficiency, level of education, place of residence, and ethnicity. Participants then answered four questions assessing their risk-taking tendencies. Three of these questions were drawn from the Domain-Specific Risk-Taking (DOSPERT) scale for adult populations (Blais and Weber 2006), and the final item was a self-assessment question: "How willing are you to take risks, in general?" Responses were measured on a 7-point Likert scale and averaged to produce a risk-aversion score.

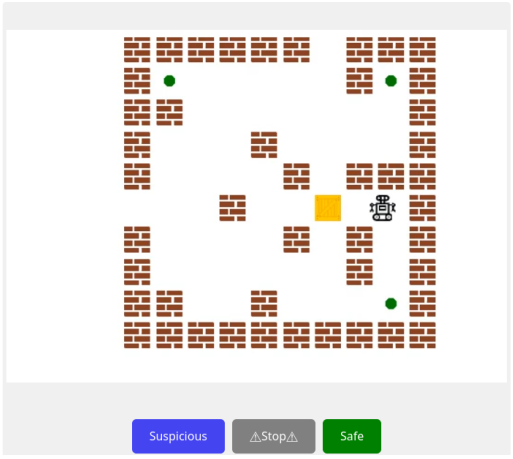


Figure 1: Experiment interface. Pressing *Suspicious* raised an alarm and enabled *Stop*, which immediately halted the agent. Pressing *Safe* ended the video without intervention.

Experimental Behaviours

We implemented three behavioural conditions based on the formal definitions above, along with two optimal baselines for comparison. Each condition was constructed to plausibly evoke suspicion from participants. Their judgments—regarding whether a behaviour was suspicious, when suspicion arose, and how long it persisted—provide an empirical test of both the validity of our formal definitions and the fidelity of their implementation.

Table 1: Generating Uncertainty in the presence of risk.

Behaviour	Characteristic	Observation	Implementation ($g_r \in \mathcal{G}$)
Deception	Adversariality	wrong goal!	$Pr(g\|\pi) \geq Pr(g_r\|\pi)$
Loitering	Non-rationality	what motivation?	$cost(s, g_r) \lesssim \text{RMP}_{g_r}$
Obfuscation	Ambiguity	which goal?	$Pr(g\|\pi) \cong Pr(g_r\|\pi)$

- **Deception:** The agent initially adopts directed behaviour (Def. 1) towards a risky goal but intermittently diverts to a random intermediate goal such that its probability exceeds that of the risky goal (Def. 5).[3] After reaching this temporary goal, it resumes movement toward the risky goal, repeating this pattern of progress and distraction in line with the shoe-tying algorithm.
- **Loitering:** The agent proceeds optimally toward a risky goal until it enters the region corresponding to the RMP (Equation 1). Once within this region, it performs a sequence of random movements for a fixed number of time steps—effectively "loitering" near the goal—before completing the task. This behaviour, conforming to Definition 4, models hesitation or aimless motion in proximity to a sensitive or hazardous area.
- **Obfuscation:** The agent acts to maintain goal ambiguity by dynamically adjusting its plan such that the probability of pursuing the risky goal remains approximately equal to that of at least one safe goal. This sustained uncertainty serves to conceal the agent's true intent for as long as possible, consistent with Definition 2 and prior formulations of goal obfuscation (Masters and Sardina 2017b; Dragan, Holladay, and Srinivasa 2014).
- **Baselines:** Optimal-Dangerous/Optimal-Safe. The agent follows an optimal plan to a randomly selected risky (dangerous) or non-risky (safe) goal. These trajectories provide references for normal goal-directed motion.

[3] In testing, we used the half-way point but alternatives include a random number of steps up to some budget or, in an online context, response to an action by the observer.

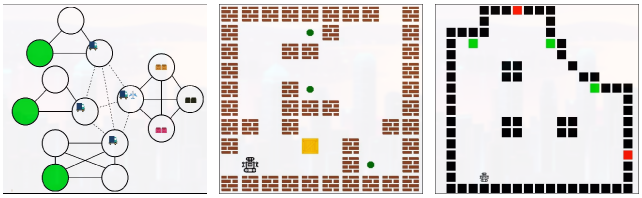
Experimental Domains

The experiment was conducted across three distinct classical planning domains, each instantiated with multiple scenarios, to evaluate the generalisability and robustness of behaviour interpretation. These domains were deliberately selected to extend beyond intuitive, navigation-based conceptions of suspicious behaviour, encompassing actions with both spatial and symbolic significance. This design choice ensured that the evaluation was not constrained to motion planning alone, but also captured higher-level, semantically meaningful behaviours.

- Sokoban (Warehouse domain): In this domain, autonomous robots are operating that can move backwards, forwards, right and left, but not diagonally (Figure 2b). Boxes in the warehouse are heavy. Robots can push them from behind, but cannot pull them. Participants were told that robots were aware that they were being watched; that they should be putting boxes into one of three designated storage areas marked by green dots but that unfortunately some robots may have been manipulated into stealing boxes by pushing them through one of two doorways. Participants were tasked with flagging any suspicious activity and stopping boxes from being stolen.
- Navigation (Hospital domain): In this domain, again, autonomous robots are operating that can move backwards and forwards, right and left, but not diagonally (Figure 2c). The domain avoided narrow corridors, and instead described open rooms connected by doors. This design choice increased the subtlety of loitering and obfuscatory behaviours, which otherwise appeared repetitive or mechanically constrained in tight spaces.

  Green squares mark three charging points, where robots can plug themselves in to recharge as necessary. Red squares mark two MRI (magnetic resonance imaging) machines. Participants were told that robots were aware of being watched; that it would be dangerous for a robot to touch an MRI but that unfortunately, some robots may have been manipulated into sabotaging the hospital facility. Participants were tasked with flagging any suspicious activity and stopping robots from touching an MRI.
- Logistics (Delivery domain): In this domain trucks and planes transport packages of chemicals (Figure 2a). There are three package colours, each representing a different type of chemical. There are three possible destinations, marked green. Packages can be loaded onto trucks or planes. Trucks and planes can carry any number of packages. Trucks travel within cities along solid lines, planes fly between airports along dotted lines. Packages

may be delivered to any green destination. However, because these chemicals can be combined to make explosives, it would be dangerous for packages of all three different colours to be delivered to the same location. Pilots and truck drivers are aware that they may be watched. Participants were tasked with flagging any suspicious activity and preventing a potential explosion. Each task defined seven possible goals: three dangerous combinations involving all chemical types at one location, and four safe delivery configurations.



(a) Logistics Domain (b) Sokoban Domain (c) Navigation Domain

Figure 2: The three domains used in our experiments: Logistics, Sokoban, and Navigation. Positive goals are shown in green, and negative goals are shown in red.

All instances of the Sokoban and Navigation domains were constructed with identical grid sizes (10x10 and 17x17 respectively) to make everything comparable. Despite the Sokoban having smaller maps, the average distance from the initial state to the goals was lower than those in Sokoban (18.08 ± 5.01, range = 9–27, vs 20.76 ± 5.49, range = 12–31). Instances of the Logistics domain had similar numbers of locations (11-16) split across 3-4 cities. Optimal solutions to logistics problems were typically longer, however most of those steps involved the repeated loading and unloading of packages.

Results

The aim of the study was to discover which behaviours participants perceived as suspicious from their interactions with the interface—specifically, whether and when they clicked the "Suspicious" button and for how long they remained in this state prior to clicking the "Safe" or "Stop" buttons.

We analysed the data using a logistic mixed-effects model, chosen because data exhibited a hierarchical structure, with repeated observations nested within participant and problem instances. Random intercepts were included for each participant, problem instance (map), and presentation order to account for within-subject and contextual dependencies. This framework enabled estimation of fixed effects while controlling for individual and task-level variability.

We evaluated *whether* each button ("Suspicious", "Stop" and "Safe") was pressed on each trial, the number of actions *before* each button was pressed (both absolute and proportional to the total number of actions), the number of actions *between* pressing the "Suspicious" and "Stop" buttons and the *distance* to the closest risky goal. We compared performance across the 5 behaviours and evaluated outcomes across the three classical planning domains.

Examination of Behaviours

Button Press Frequency Analysis of variance on the mixed-effects models showed that behaviour was the main factor driving whether participants pressed the "Suspicious", "Stop" or "Safe" buttons, with significant effects observed for all three responses (Suspicious: $F(4, 390.1) = 66.54$, $p < 0.001$; Stop: $F(4, 391.2) = 55.23$, p ¡ 0.001; Safe: $F(4, 394.8) = 29.80$, $p < 0.001$). This effect remained statistically significant when excluding the optimal-to-safe condition. The raw proportions for each behaviour are shown in Table 2.

Table 2: Percentage of trials by button pressed. S*afe behaviour is readily distinguished from potential danger.*

Behaviour	Suspicious	Stop	Safe
Optimal–Safe	23%	12.6%	52.9%
Deception	88.4%	79.1%	11.6%
Obfuscation	84.7%	74.1%	9.4%
Loitering	87.4%	81.6%	9.2%
Optimal–Dangerous	91.7%	81%	6%

Button Press Timing Values in Table 3 represent the mean (±SD) number of steps before each button was pressed, calculated over trials in which the corresponding button was used. Consistent with the binary results, analysis of variance on the mixed-effects models for step counts confirmed a strong main effect of behaviour for both the "Suspicious" and "Stop" responses (Suspicious: $F(4, 255.5) = 20.76$, $p < 0.001$; Stop: $F(4, 217.5) = 32.87$, $p < 0.001$). The low number of "Safe" responses for behaviours other than safe-optimal prevented reliable model estimation for that button. These results remained consistent when accounting for absolute time in the video and when scaling step counts by total plan length.

Table 3: Mean (±SD) number of steps before each button was pressed. *Deception remains unsuspected the longest.*

Behaviour	Suspicious steps	Stop steps
Optimal–Safe	13.6 ± 6.5	12.6 ± 8.1
Deception	30.2 ± 16.7	35.7 ± 16.9
Obfuscation	23.9 ± 10.6	25.8 ± 11.0
Loitering	22.3 ± 11.7	29.4 ± 10.6
Optimal–Dangerous	16.2 ± 6.6	19.0 ± 6.7

Suspicion Duration A linear mixed-effects model analysing the number of actions taken between pressing the "Suspicious" and "Stop" buttons showed a significant main effect of behaviour ($F(4, 226.7) = 9.88$, $p < 0.001$). A corresponding model using the same measure scaled by total plan length confirmed this result ($F(4, 223.3) = 9.14$, $p < 0.001$). Mean differences were small relative to their standard deviations, indicating high variability over when participants chose to stop, after marking a behaviour as suspicious. Specifically, Table 4 indicates that the most suspicious behaviour, i.e. the behaviour which was spent the longest in suspicious mode, was the loitering behaviour.

Table 4: Mean (±SD) number of actions and proportion of total plan length between pressing the "Suspicious" and "Stop" buttons. *Loitering remains suspicious for the longest.*

Behaviour	Stop–Watch actions	Stop–Watch proportion
Optimal–Safe	0.70 ± 0.82	0.018 ± 0.022
Deception	4.38 ± 7.16	0.105 ± 0.155
Obfuscation	2.55 ± 3.11	0.083 ± 0.092
Loitering	8.16 ± 9.54	0.259 ± 0.308
Optimal–Dangerous	2.48 ± 2.68	0.118 ± 0.116

Suspicion Distance Lastly, a linear mixed-effects model predicting the distance to the closest risky goal at the time participants pressed the "Suspicious" button showed a significant main effect of behaviour (F(4, 178.3) = 3.48, p < 0.001). Participants allowed deceptive and obfuscatory agents to approach closer to the risky goal before flagging suspicion, while loitering and optimal behaviours were identified earlier, at greater distances. This is shown in Table 5.

Table 5: Mean (±SD) distance to the closest risky goal when "Suspicious" button was pressed (see Experimental Behaviours). *Obfuscation and deception mask the approach.*

Behaviour	Distance to dangerous goal (steps)
Optimal–Safe	13.8 ± 2.9
Deception	11.6 ± 8.3
Obfuscation	11.3 ± 10.8
Loitering	14.6 ± 7.7
Optimal–Dangerous	14.7 ± 7.8

Examination of Domains

Button Press Frequency Analysis of variance on the mixed-effects models showed that domain was not a significant factor influencing whether participants pressed the "Suspicious"," "Stop" or "Safe" buttons (F(2, 13.67) = 1.59, p = 0.24; F(2, 13.72) = 0.71, p = 0.51; F(2, 387.71) = 1.71, p = 0.18, respectively). Pairwise comparisons confirmed that *no* domains differed significantly from one another on this measure. The overall proportions are presented in Table 6.

Table 6: Percent of trials where each button was pressed.

Domain	Suspicious	Stop	Safe
Sokoban (Warehouse)	73.9%	65.9%	13.8%
Navigation (Hospital)	71.2%	63%	19.2%
Logistics (Delivery)	79.3%	67.6%	20.7%

Button Press Timing Domain had a marginal effect on the number of steps before pressing "Suspicious" (F(2, 11.7) = 3.51, p = 0.064) and a significant effect on stopping behaviour (F(2, 8.0) = 5.06, p < 0.05), but no significant effect on the number of steps before pressing "Safe" (F(2, 387.71) = 0.87, p=0.45) as shown in Table 7.

Suspicion Distance Domain had a significant effect on the distance to the closest risky goal at the time participants pressed the "Suspicious" button (F(2, 11.73) = 8.41, p=0.005). Participants marked agents as suspicious when they were farthest from a risky goal in the Sokoban domain and closest in the Logistics domain, with Navigation between the two. The contrast between Sokoban and Logistics was statistically significant based on post-hoc Tukey pairwise comparisons of the estimated marginal means (p = 0.003). These results are shown in Table 8.

Table 7: Mean (±SD) number of steps before each button was pressed. *Logistics was quickly suspicious.*

Domain	Suspicious steps	Stop steps	Safe steps
Sokoban (Warehouse)	27.8 ± 12.8	33.3 ± 11.5	22.1 ± 6.3
Navigation (Hospital)	23.2 ± 13.8	28.5 ± 14.3	19.7 ± 13.2
Logistics (Delivery)	17.3 ± 9.7	19.7 ± 10.6	20.9 ± 8.9

Table 8: Mean (±SD) distance to the closest risky goal when the "Suspicious" button was pressed (see Experimental Domains). *Sokoban aroused suspicions at a distance.*

Domain	Distance to risky goal (steps)
Sokoban (Warehouse)	19.0 ± 8.8
Navigation (Hospital)	14.3 ± 9.9
Logistics (Delivery)	10.0 ± 4.5

Risk Taking

Four preliminary questions were used to calculate a risk-aversion score, as described in the Experimental Task. Responses were averaged across the four to produce a single metric, with higher values indicating greater risk aversion. This score was a significant predictor of participants pressing the "Safe" button ($F(1, 28.6) = 3.87, p = 0.059$) but did not significantly influence any other response measures.

Discussion

Proximity vs Masking Intent. It is not only proximity to a risky goal that raises suspicion, but also the manner in which it is approached. Certainly, participants generally stopped agents from pursuing risky goals regardless of the specific behaviour exhibited and behaviour directed toward a safe goal resulted in significantly fewer "Suspicious" responses than other behaviours, as confirmed by Tukey-adjusted post-hoc comparisons on the mixed-effects model (all p < 0.001). The same pattern held for the "Stop" and "Safe" buttons, with safe-goal behaviour eliciting markedly fewer presses than any other behaviour (all p < 0.001).

However, analysis of the distance to the nearest risky goal when participants pressed the "Suspicious" button (Suspicion Distance) revealed that deceptive and obfuscatory plans allowed agents to approach significantly closer to the goal before being flagged. A grouped analysis combining these two deceptive behaviours confirmed this, (F(1,176.4) = 14.03, p = 0.0002). This suggests that such behaviours successfully mask intent, delaying recognition *despite* their proximity to risky goals.

Loitering. Participants took substantially more actions between pressing the "Suspicious" and "Stop" buttons when observing loitering behaviour. Post-hoc Tukey tests showed that loitering produced significantly longer delays than deceptive, obfuscatory, and optimal-to-dangerous behaviours (all $p < 0.005$), which did not differ significantly from one another (see Table 4). This pattern aligns with expectations: deceptive and obfuscatory agents act identically to optimal agents once they reach the radius of maximum probability.

Additional confirmation for this is that among the suspicious behaviours, loitering required the fewest actions before being marked as suspicious (see Table 3). This is consistent with its early optimal trajectory toward a risky goal. In contrast, deceptive behaviour, which follows a deliberately indirect path, produced a statistically significant increase in the number of actions before being marked suspicious (vs. loitering, p = 0.0023; vs. obfuscatory, p = 0.058).

It appears that loitering behaviour holds the attention of an observer for longer than other behaviours. This could be relevant in the context of deceptive planning, where it may be useful to employ a 'loitering decoy' to divert attention from a second agent proceeding to the intended goal.

Sokoban is Harder. Although the Navigation and Sokoban domains both featured a single agent moving on a grid, their action dynamics differed markedly. In Sokoban, some actions are irreversible—such as pushing a box into a corner—which can make a goal permanently unreachable or greatly increase the steps required to reach it. By contrast, in Navigation and Logistics, any action can be undone with a single counter-action. This distinction has important implications for generating obfuscatory behaviour: moving toward one goal in Sokoban can render alternatives impossible, quickly exposing the agent's true intention. Consequently, sustained obfuscation was harder to produce in Sokoban. However, the results showed no significant differences in obfuscatory behaviour across the three domains.

The Number of Actions Matters. Another structural difference between the domains was the number of available actions per state. Navigation and Sokoban each provided at most five actions—the four cardinal moves plus a "wait" action—whereas the Logistics domain frequently offered more than fifteen. This disparity is especially relevant for loitering behaviour, where the agent must act randomly while staying near a risky goal. The larger action space in Logistics created far more opportunities to "loiter", as many actions involved irrelevant operations (e.g., moving a truck in another city) that did not progress the agent toward its goal but still kept it active in the environment.

A mixed-effects model with a behaviour–domain interaction showed that loitering behaviour differed significantly across domains ($F(7, 231) = 2.79, p = 0.008$). Tukey follow-up tests indicated that this effect was driven solely by loitering: participants stopped loitering agents earlier in the Logistics domain than in Navigation ($p = 0.0001$) or Sokoban ($p = 0.0027$). No other behaviour type showed meaningful domain differences. Although counter-intuitive—since more available actions might be expected to lengthen Suspicion Duration, the complexity of the task may have produced information overload, prompting a rapid "stop" response under perceived threat.

Goal Design in Logistics. Goal design varied by domain to preserve logical consistency while ensuring intuitive interpretation. In Navigation, goals were simple positional targets: green for safe, red for risky. In Sokoban, goals involved moving boxes onto green targets (safe) or off the map (risky), simulating theft. This mapping provided clear behavioural intent while remaining structurally equivalent for analysis. The Logistics domain required refinement to prevent trivial identification of malicious goals. Early designs defined danger as placing identical chemicals together, which made intent immediately obvious—participants routinely stopped the agent as soon as it picked up two packages of the same type. The final design instead defined danger as delivering one of each package type to a single destination, allowing agents to handle similar packages safely before revealing intent. In addition, participants in preliminary testing often reported difficulty interpreting agent actions, prompting refinements to the task instructions and slower animations. Even after these changes, participants remained substantially more likely to flag optimal-to-safe behaviour as suspicious in Logistics than in Navigation ($OR = 46.15, p = 0.004$) or Sokoban ($OR = 6.46, p = 0.039$).

Mixed-effects model analyses showed that participants pressed "Suspicious" and "Stop" earlier across all behaviours in the Logistics domain, even though plans were longer on average. When adjusting for plan length, domain effects (F(2, 12.5) = 11.72, p= 0.001) outweighed behavioural effects (F(4,274.7) = 7.67, p= 0.687) in predicting the number of steps before "Suspicious" was pressed. By contrast, the interval between "Suspicious" and "Stop" was more consistent across domains (F(2,10.9) = 2.75, p=0.1 for domain; F(4,237.4) = 9.73, p <0.0001 for behaviour). This pattern likely reflects the agent's need to manipulate all package types well before it can complete its deliveries, revealing intent earlier than in other domains.

In summary, loitering appears to be the most useful behaviour to focus on for the generation of suspicious behaviour; the ideal domain should allow ample plan length, be simple enough for observers to fully understand the goals, but complex enough to offer multiple actions. An extension of the Sokoban domain may meet these requirements.

Conclusion

In this paper, we defined three suspicious behaviours grounded in social psychology. We showed how classical planning and model-based goal recognition could be used to synthesise plans for the generation of those behaviours. We introduced a novel deceptive planning algorithm and empirically evaluated all behaviours through a user study conducted across three classical planning domains. We found that, across all domains, loitering was perceived as the most suspicious behaviour. In future work, we plan to extend this framework for suspicious behaviour generation and use it to develop a system for suspicious behaviour recognition.

Acknowledgments

This research was supported by the Monash Statistical Consulting Service, part of Monash eResearch. Peta Masters gratefully acknowledge support from the Government Office for Science and the Royal Academy of Engineering under the UK Intelligence Community Postdoctoral Research Fellowships scheme. Gal.A.Kaminka was supported in part by ISF Grant #1371/24, and thanks K. Ushi. We further acknowledge the contribution to this work of our much missed colleague, Ramon Fraga Pereira.

References

Avrahami-Zilberbrand, D.; and Kaminka, G. A. 2007. Incorporating observer biases in keyhole plan recognition (efficiently!). In *AAAI*, volume 7, 944–949.

Avrahami-Zilberbrand, D.; and Kaminka, G. A. 2014. Keyhole adversarial plan recognition for recognition of suspicious and anomalous behavior. *Plan, activity, and intent recognition*, 87–121.

Blais, A.-R.; and Weber, E. U. 2006. A Domain-Specific Risk-Taking (DOSPERT) scale for adultpopulations. *Judgment and Decision Making*, 1(1): 33–47.

Carberry, S. 1996. Plan recognition: achievements, problems, and prospects. *AA*, 5: 6.

Deutsch, M. 1958. Trust and suspicion. *Journal of conflict resolution*, 2(4): 265–279.

Dragan, A. D.; Holladay, R. M.; and Srinivasa, S. S. 2014. An Analysis of Deceptive Robot Motion. In *Robotics: science and systems*, 10.

Fein, S.; Hilton, J. L.; and Miller, D. T. 1990. Suspicion of ulterior motivation and the correspondence bias. *Journal of personality and social psychology*, 58(5): 753.

Kaminka, G.; Vered, M.; and Agmon, N. 2018. Plan recognition in continuous domains. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Keren, S.; Gal, A.; and Karpas, E. 2014. Goal recognition design. In *Proceedings of the International Conference on Automated Planning and Scheduling*, volume 24, 154–162.

Keren, S.; Gal, A.; and Karpas, E. 2015. Goal recognition design for non-optimal agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29.

Kulkarni, A.; Srivastava, S.; and Kambhampati, S. 2019. A unified framework for planning in adversarial and cooperative environments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 2479–2487.

Levi, B.; and Loeben, G. 2004. Index of suspicion: Feeling not believing. *Theoretical Medicine and Bioethics*, 25: 277–310.

Masters, P.; and Sardina, S. 2017a. Cost-based goal recognition for path-planning. In *Proceedings of the 16th conference on autonomous agents and multiagent systems*, 750–758.

Masters, P.; and Sardina, S. 2017b. Deceptive Path-Planning. In *IJCAI*, 4368–4375.

Masters, P.; and Sardina, S. 2019. Goal recognition for rational and irrational agents. In *Proceedings of the 18th international conference on autonomous agents and multiagent systems*, 440–448.

Masters, P.; and Sardina, S. 2021. Expecting the unexpected: Goal recognition for rational and irrational agents. *Artificial Intelligence*, 297: 103490.

Masters, P.; and Vered, M. 2021. What's the context? implicit and explicit assumptions in model-based goal recognition. In *International Joint Conference on Artificial Intelligence 2021*, 4516–4523. Association for the Advancement of Artificial Intelligence (AAAI).

Meneguzzi, F. R.; and Pereira, R. F. 2021. A survey on goal recognition as planning. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI), 2021, Canadá.*

Ramírez, M.; and Geffner, H. 2010. Probabilistic plan recognition using off-the-shelf classical planners. In *Proceedings of the AAAI conference on artificial intelligence*, 1121–1126.

Russell, S.; and Norvig, P. 2015. Artificial Intelligence: a Modern Approach. *Learning*, 2(3): 4.

Skolnick, J. H. 2011. *Justice without trial: Law enforcement in democratic society*. Quid pro books.

Sodemann, A. A.; Ross, M. P.; and Borghetti, B. J. 2012. A review of anomaly detection in automated surveillance. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(6): 1257–1272.

Tambe, M.; and Rosenbloom, P. S. 1995. RESC: An approach to agent tracking in a real-time, dynamic environment. In *Proceedings of the International Joint Conference on AI*.

Vered, M.; and Kaminka, G. A. 2017. Heuristic online goal recognition in continuous domains. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 4447–4454. An improved version is available as arxiv:1709.09839.

Vered, M.; Kaminka, G. A.; and Biham, S. 2016. Online goal recognition through mirroring: Humans and agents. In *Annual Conference on Advances in Cognitive Systems 2016*. Cognitive Systems Foundation.