

Defining the Saliency of Agent Perceptions in CATALINA^{*}

Massimo Cossentino¹, Giovanni Pilato¹ and Guido Aversa¹

¹National Research Council of Italy (CNR), Via U. La Malfa, 153, Palermo, 90146, Italy

Abstract

The CATALINA cognitive agent architecture extends the classical Belief-Desire-Intention model with a consciousness model inspired by Baars' theory of the Global Workspace; it also features a sophisticated attention mechanism in which every stimulus carries a numerical *saliency* value. That value is the single quantity that determines whether a perception is admitted to the agent's conscious memory and, through it, whether the agent reorients its behaviour. In the current implementation of the architecture, however, stimulus saliency is hand-coded by the analyst at design time, agent-independent, and constant throughout the agent's lifetime. This paper proposes a novel definition of stimulus saliency based on the notion of Bayesian surprise [1], formalised as the Kullback-Leibler divergence between the agent's posterior and prior beliefs after a perception. We show that the proposal requires a minor architectural change, equipping each belief with a confidence value in $[0, 1]$ rather than a Boolean truth value. An example of the autonomous-vehicle case study of the original architecture illustrates the qualitative behaviour of the proposed mechanism, including the automatic suppression of uninformative perceptions and a limited form of open-world compatibility for stimuli that can be mapped to existing beliefs.

Keywords

BDI agents, Bayesian surprise, perception saliency, attention selection, cognitive agent architectures, CATALINA, probabilistic beliefs

1. Introduction

Autonomous agents that operate in real-world environments are exposed to a continuous, high-bandwidth stream of perceptions, only a small fraction of which deserve to influence the agent's current behaviour. Distinguishing the perceptions that are worth attention from those that are not is a long-standing problem of artificial intelligence and a defining concern of the Belief-Desire-Intention paradigm [2], in which limited deliberative resources must be allocated across a typically large set of potentially relevant beliefs and competing desires. The classical BDI model offers no explicit mechanism to address this allocation problem: all relevant beliefs are assumed to be uniformly accessible to the reasoner, and the reasoner is expected to select intentions through a largely sequential, centralised deliberation loop. As a consequence, BDI agents tend to be either excessively reactive, constantly reorienting their behaviour in response to minor stimuli, or excessively rigid, locked into an ongoing plan and blind to events that would justify a re-prioritisation.

The CATALINA cognitive agent architecture [3, 4, 5] addresses this gap by integrating the BDI core of Bratman et al. [2] with the Global Workspace Theory of consciousness [6, 7] and with the Buehler's executive system [8, 9, 10]. In CATALINA, perceptions, beliefs, desires and intentions live in a shared short-term memory (the Global Workspace) and a set of concurrent executive functions regulates which information is admitted to that memory, how attention is focused on a particular intention, and how desires are promoted to intentions and back.

The pivot of this regulation is a numerical *saliency* value attached to every stimulus and to every desire: a comparison between the saliency of a stimulus and one of two thresholds maintained in the workspace decides whether the stimulus reaches the agent's conscious memory at all, and a comparison between a desire's saliency and the corresponding threshold decides whether the desire becomes a candidate for promotion to intention.

WOA 2026, the 27th Workshop From Objects to Agents, Salerno, June 15-17, 2026

*Corresponding author: M. Cossentino

✉ massimo.cossentino@icar.cnr.it (M. Cossentino); giovanni.pilato@icar.cnr.it (G. Pilato); guido.averna@icar.cnr.it (G. Aversa)

ORCID: 0000-0003-1258-9767 (M. Cossentino); 0000-0002-6254-2249 (G. Pilato); 0000-0002-0935-5866 (G. Aversa)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

In the case of desires, saliency reflects the analyst’s judgement of how relevant or urgent the desired state of the world is to the agent’s mission, and it is sensible for that judgement to be expressed once, at design time. In the case of stimuli, however, the same answer is much less satisfactory. Stimuli arise from perceptions, and perceptions are the agent’s interface with an environment that is, by definition, not under the analyst’s control. The current implementation of CATALINA acknowledges this problem but does not solve it: it specifies only the *interface* of the saliency-assignment step, a stimulus carries a saliency value, assigned by the Information Extraction module of the Perception Executive Function, and leaves the actual computation of that value as a domain-specific responsibility of the programmer. The saliency of any specific perception is therefore hand-coded, agent-independent, and constant for the agent’s lifetime.

This paper proposes a definition of stimulus saliency that replaces the hand-coded value with a quantity the agent can compute from its own state. The proposal takes its theoretical foundation from Itti and Baldi [1], who introduced *Bayesian surprise*, the Kullback-Leibler divergence between an observer’s posterior and prior beliefs after a new datum, as a principled, observer-relative measure of how much a perception changes the observer’s view of the world, and showed empirically that this measure outperforms classical saliency-map metrics at predicting where human gaze lands in natural video. We adapt this notion to the symbolic belief base of a BDI agent and define the saliency of a stimulus as the Kullback-Leibler divergence between the agent’s posterior and prior over the affected beliefs. The adaptation requires a single architectural change: equipping each belief with a confidence value in $[0, 1]$, in line with the existing probabilistic-BDI literature [11, 12], and a new computation within the Information Extraction module of the Perception Executive Function. The downstream architecture, the Stimulus Filtering algorithm, the Saliency and Attention Thresholds, the inhibition regions, the Desire Promotion mechanism, the Reasoner, continues to operate exactly as in the current implementation: the proposal modifies only the *provenance* of the saliency value, not its *role*.

The contribution of this paper, rather than designing a new attention mechanism from scratch, is to improve the existing architecture by replacing one hand-coded quantity with one computed quantity, justified by an established line of empirical and theoretical work in perception science. The single-factor information-theoretic formulation we adopt is deliberately minimal: it does not yet incorporate goal- and value-driven modulation of saliency, which the modern visual-search literature [13] and the active-inference framework [14, 15] both argue is necessary for a complete account of attention. We acknowledge this limitation and discuss how it can be addressed in future work without disrupting the proposal of this paper.

The remainder of the paper is organised as follows. Section 2 reviews the relevant literature on saliency, surprise and attention, including the bottom-up tradition originating in computational vision, the multi-factor consensus emerging from modern visual-search studies, the active-inference framework, the treatment of attention in cognitive architectures based on Global Workspace Theory and the existing probabilistic extensions of the BDI model. Section 3 provides a self-contained recap of CATALINA and characterises how saliency is computed and used in the current implementation. Section 4 develops the proposal: it motivates the choice of Bayesian surprise, gives the formal definition adapted to the BDI belief base, describes the architectural integration in CATALINA, illustrates the resulting behaviour on a worked example from the autonomous-vehicle case study and discusses the numerical considerations that arise in practice. Section 5 concludes with a summary of the contribution, a list of its current limitations and a roadmap for the extensions that we deem most promising.

2. State of the Art

The notion of *saliency* sits at the intersection of perception science, information theory, neuroscience and cognitive architectures. A perception is salient when it warrants the allocation of limited cognitive resources to it, and a long line of research, spanning more than three decades, has tried to make precise what that means and how saliency can be computed from the perception itself, from the observer’s internal state, or from both. This section reviews the strands of literature that are most relevant to the

present work: bottom-up saliency models from computational vision; the empirical critiques that ruled them out as a complete account of attention; the modern multi-factor consensus in the visual-search literature; the normative account of saliency offered by the active-inference framework; the treatment of attention in cognitive architectures based on Global Workspace Theory; and the small body of work that introduces probabilistic and uncertainty-aware extensions to the BDI model. We discuss each strand in turn, paying attention to the formal definitions it offers (or fails to offer), to the role it assigns to the observer, and to its compatibility with a symbolic, BDI-style agent architecture such as CATALINA.

2.1. Bottom-up saliency: from feature maps to Bayesian surprise

The first influential computational account of saliency [16] treated it as a property of the stimulus alone. A visual scene was decomposed into a set of feature maps (intensity, colour, orientation), each of which was processed by centre-surround filters in the spirit of early visual cortex; a single *saliency map* was then obtained as a weighted combination of the feature maps, and the most active location of the saliency map was selected as the next focus of attention. This pipeline is purely data-driven, it ignores the internal state of the observer entirely, and it produces a saliency value that is reproducible and observer-independent. Itti and Koch [17] consolidated this line of work into a broader framework that became the dominant operational definition of saliency in computational vision for the following decade.

Two limitations of the saliency-map account were already evident at the time. First, the saliency value of a stimulus is a property of the *image*, not of any cognitive process; in particular, two observers facing the same scene receive identical saliency assignments even when they have radically different goals or expertise. Second, the model relies on a fixed, hand-tuned weighting of the feature maps, so the saliency value at a location is essentially predetermined by the choice of features, rather than emerging from anything resembling a learning or updating process.

Both limitations were addressed, though only partially, by the seminal work of Itti and Baldi [1], which redefined surprise (and, implicitly, saliency) in Bayesian terms. The observer holds a prior distribution $P(M | D)$ over a model space M given previously observed data D . After observing a new datum D' , the posterior becomes $P(M | D, D')$. *Bayesian surprise* is then the Kullback-Leibler divergence between posterior and prior:

$$S(D', D) = \text{KL}(P(M | D, D') \| P(M | D)) = \int_M P(M | D, D') \log \frac{P(M | D, D')}{P(M | D)} dM. \quad (1)$$

Two properties of this definition are important for the present work. First, surprise is observer-relative: the same datum produces different surprise values for observers that hold different priors, which reflects the intuitive idea that what is surprising depends on what one already believes [1]. Second, surprise is *not* the same as Shannon information [18]. Itti and Baldi [1] are explicit about this: an unstructured stimulus such as television static carries near-maximal Shannon information, because the symbol stream is unpredictable, but it produces near-zero Bayesian surprise, since it leaves the observer’s beliefs unchanged. Conversely, a single observation that disambiguates a hypothesis can carry low Shannon information yet enormous Bayesian surprise. The Bayesian-surprise account therefore separates “unpredictability” from “belief change”, and only the latter is treated as a driver of attention.

Itti and Baldi [1] validated their proposal empirically by comparing several computational metrics of saliency against human eye-tracking data on natural video. Bayesian surprise outperformed standard saliency-map metrics, motion energy, intensity variance and entropy at predicting where human gaze would land. This is a strong argument that *change in belief*, rather than *intensity of stimulation*, is the cognitively meaningful quantity.

2.2. The empirical case against pure bottom-up saliency

Even an observer-relative, change-driven saliency value remains, in the end, a property of the stimulus-prior pair, it does not reference the agent’s task or current goal. Tatler et al. [19] provide an extensive

empirical critique of saliency-based explanations of eye guidance in natural vision, showing that they “are poor at accounting for many aspects of picture viewing and can fail dramatically in the context of natural task performance”. A canonical example, often quoted from this line of work, is ball-catching: the human observer’s gaze lands on the future landing point of the ball, where image salience is, by construction, low (the ball is not yet there). The authors propose that any complete account of attention must include systematic biases in how eyes move, prior knowledge of scene structure, and, crucially for our purposes, the value of potential fixation locations to the current task. Saliency, in this view, is *one* contributor among several, not the dominant one.

For an agent architecture, the Tatler critique translates directly: saliency cannot be defined in isolation from the agent’s goals. However, the same critique also says that defining saliency from the stimulus-belief pair is a necessary first step, even if it is not the final step. This positions Bayesian surprise as the right *building block* for a saliency theory of agent perception, with the understanding that further enrichment will be needed before the saliency value can drive task-relevant behaviour optimally.

2.3. The modern multi-factor consensus

A more recent synthesis is offered by Wolfe and Horowitz [13], who identify five factors that, in the light of two decades of visual-search research, are jointly necessary to explain attention guidance: (i) bottom-up salience; (ii) top-down feature guidance, i.e. the searcher’s knowledge of what they are looking for; (iii) scene structure and meaning; (iv) the previous history of search, ranging in timescale from milliseconds to years; and (v) the relative value of targets and distractors. The first factor maps onto the bottom-up saliency literature reviewed in Section 2.1.

The relevance of this five-factor account to the present work is twofold. First, it makes explicit that a saliency mechanism for a cognitive agent should, in the long run, fold in goal- and value-based modulation alongside the bottom-up component; in our roadmap this amounts to weighting the surprise value by the relevance of the affected beliefs to the agent’s current desires, an extension we discuss in Section 5. Second, it tells us where the purely surprise-based proposal of this paper sits within the broader framework: it provides a principled, observer-relative replacement for factor (i), while leaving factors (ii)-(v) for subsequent work.

2.4. Surprise in the active-inference framework

A largely independent thread of research, originating in theoretical neuroscience, treats attention, perception and action selection as facets of a single normative principle: the minimisation of expected free energy. The free-energy principle [20] posits that biological agents act to minimise long-run surprise, where “surprise” is the negative log-evidence of sensory observations under a generative model of the world. The active-inference process theory [14] operationalises this principle for discrete-time decision making and decomposes the expected free energy of a candidate observation into two terms: an *extrinsic value*, quantifying how much the observation would move the agent toward its prior preferences (its goals), and an *epistemic value*, quantifying how much it would reduce uncertainty about the world. Attentional saliency, in this framework, is therefore intrinsically two-component: a goal-driven term and an information-driven term, combined under one formal umbrella [15, 21, 22].

This framework is the closest formal alternative to the line of work proposed in this paper. There are, however, two reasons not to adopt it directly. First, active inference is a global commitment: it prescribes that the agent’s entire cognitive cycle, perception, attention, deliberation, and action selection, be cast as a single optimisation over generative models and probabilistic policies. Retrofitting it onto an existing symbolic architecture with explicit beliefs, desires and intentions is non-trivial and risks rebuilding the architecture from scratch under a new formalism. Second, the empirical validation of the epistemic-value term in active inference is largely indirect, mediated by the broader free-energy framework; in contrast, Bayesian surprise as defined by Itti and Baldi [1] has been validated directly against human gaze data and, as a localised quantity, can be computed point-wise from the agent’s current beliefs without committing to any particular generative model of the world. The position taken

in this paper is therefore that a Bayesian-surprise account is a lighter-weight, more readily integrable proposal that captures the information-theoretic core of active-inference saliency without requiring the host architecture to be rebuilt around it.

2.5. Saliency in cognitive architectures

Among cognitive architectures with an explicit conscious-agent ambition, the closest theoretical neighbour of CATALINA is LIDA, the Learning Intelligent Distribution Agent of Franklin and Patterson [23], Franklin et al. [24], which, like CATALINA, is built on Baars’ Global Workspace Theory [6, 7]. A LIDA agent runs a fast cognitive cycle in which *attention codelets* group portions of the current situational model into *coalitions* that compete for entry into the global workspace; the most highly activated coalition wins the competition and is broadcast to the unconscious modules of the architecture. The architecture implements a multi-factor saliency mechanism in which both bottom-up and top-down considerations are present, although the specific weighting of the factors is procedural and codelet-dependent.

LIDA’s mechanism is informative for the present work since it shows that a Global-Workspace architecture can support a multi-factor saliency assignment without abandoning its symbolic backbone. The contribution of this paper is therefore narrower and more specific than LIDA’s overall mechanism: rather than designing an end-to-end attention architecture, we propose a principled, KL-derived saliency value for a single stimulus in an existing symbolic architecture, and we show how that value can be computed from quantities the architecture already maintains.

2.6. Probabilistic and uncertainty-aware BDI agents

The classical BDI model [2] treats beliefs as boolean propositions: an agent either holds a belief or it does not. Under this assumption, a perception either confirms a belief, falsifies it, or adds a new one, and there is no graded notion of belief change. Any serious attempt to define saliency as the magnitude of a belief update therefore requires beliefs that carry a confidence value in $[0, 1]$ rather than just a truth value.

The probabilistic-BDI literature has already taken this step, although it has done so for reasons unrelated to saliency. Álvaro Carrera and Iglesias [11] introduce a Bayesian-BDI agent in which beliefs are represented in a multiply-sectioned Bayesian network and updated causally as new evidence is received. More recent work [12] addresses the modelling and verification of BDI agents in which beliefs are subject to uncertainty, and discusses the consequences for plan selection and reasoning. None of these proposals, to the best of our knowledge, defines saliency as a function of the magnitude of a belief update, but together they establish that the architectural step required to do so, equipping beliefs with a confidence value in $[0, 1]$, is already present in the literature and can be imported into CATALINA without re-inventing it.

2.7. Positioning of this work

The literature reviewed above suggests three points of consensus and one open opportunity. First, a plausible saliency value for a perception is observer-relative and tied to the change the perception induces in the observer’s beliefs [1]; pure stimulus-only definitions of saliency [16, 17] have been shown to be empirically insufficient [19]. Second, a complete account of attention also requires goal- and value-driven modulation [13], although how to combine the bottom-up and the top-down component remains an open issue. Third, the active-inference framework [14, 15] provides a unified normative account of saliency that combines the two components but at the cost of a deep architectural commitment which is not appropriate for a symbolic BDI architecture such as CATALINA.

The opportunity is that no existing work, to the best of our knowledge, applies the Bayesian-surprise definition of Itti and Baldi [1] to the belief base of a BDI agent in order to replace a hand-coded saliency value with a quantity that the architecture can compute internally, transparently and on demand. Section 4 details this proposal for the CATALINA architecture.

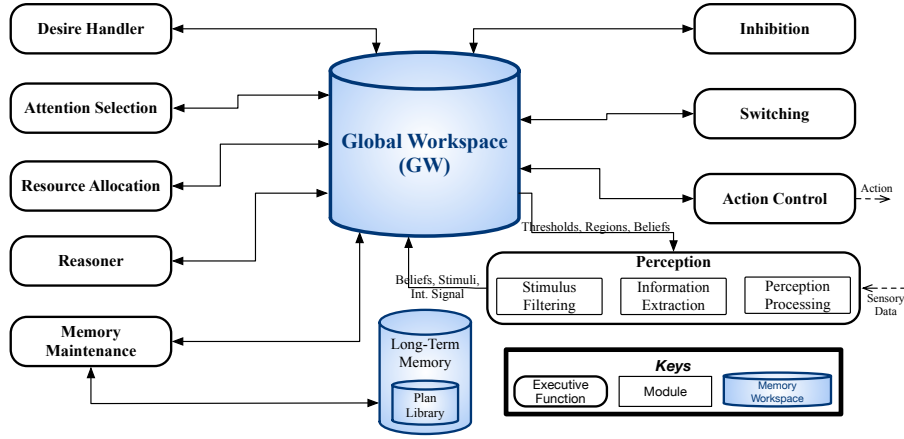


Figure 1: The CATALINA architecture. Because of space constraints, this figure only details the modules and data flow of the Perception Executive Function.

3. Saliency in CATALINA

This section grounds the proposal of Section 4 in the existing CATALINA architecture (see Figure 1). It first recaps the theoretical commitments of CATALINA and the components that implement them; we focus on the parts of the architecture that the proposal of this paper directly touches and refer the reader to [5, 4] for the full description of the architecture, including the components that are not affected by the present work. The second subsection then concentrates on saliency itself: where it enters the architecture, how it is currently computed, and how it propagates through the attention and deliberation machinery.

3.1. The CATALINA Architecture

CATALINA is a cognitive agent architecture, currently in version 0.3.1, that integrates three theoretical strands. Its practical-reasoning backbone is the Belief-Desire-Intention model of Bratman et al. [2]; its account of conscious access to information follows Baars’ Global Workspace Theory [6, 7]; and its top-down regulation of behaviour is modelled on Buehler’s account of the executive system [8, 9, 10]. The architecture is implemented as a distributed, microservices-oriented system in which nine concurrent *executive functions* interact through a single shared structure, the Global Workspace (GW), by means of a publish-subscribe pattern. The motivation for the integration is the observation that classical BDI architectures lack mechanisms for adaptive attention regulation and meta-level executive control, and therefore tend to be either excessively reactive or excessively rigid in dynamic, information-rich environments.

Beliefs, Desires, Intentions, and Stimuli. A *belief* in CATALINA is a logical predicate equipped with a truth value. Predicates follow the canonical $\langle \text{subject}, \text{relationship}, \text{complement} \rangle$ form, so that propositions such as [“TrafficlightX is green”, FALSE] or [“RouteY is closed”, TRUE] are first-class elements of the agent’s belief base. *Desires* fall into four categories. *Practical* desires specify a target state of the external world (e.g. *Visit Madrid by tomorrow*) and are expressed in metric interval temporal logic; *epistemic* desires specify a target state of the internal belief base (e.g. *Find the nearest gas station*) and aim to acquire knowledge rather than to change the world; *quality* desires (e.g. *Drive safely*) impose soft constraints on other desires without a crisp satisfaction condition; *green* desires (e.g. *Limit Noise Pollution*) encode environmental policies that the agent must respect and that, unlike quality desires, cannot be relaxed. Practical and epistemic desires are jointly referred to as *attentional desires* because their consideration for pursuit depends on the attentional state of the agent.

Every desire in CATALINA carries a numerical *saliency* value, typically in $[0, 1]$. The saliency of a desire is fixed by the analyst at design time and reflects how relevant or how urgent the desire is, in the

analyst's judgement, for the agent's overall mission.

A *stimulus* is defined in CATALINA as a special kind of belief, bound to a specific perception. Stimuli inherit the predicate-truth-value structure of beliefs and add a saliency value of their own, intended to represent the relevance or urgency of the percept. The two flavours of saliency - desire saliency and stimulus saliency - are kept distinct in the architecture but they play structurally similar roles: both are scalar values that determine what passes through the attention mechanism described below. The present paper is concerned exclusively with the second flavour, the saliency of stimuli.

The Global Workspace and the executive functions. The GW is implemented as a publish-subscribe dashboard that acts as the agent's short-term shared memory. It stores the beliefs, stimuli, desires, and intentions that are currently relevant to the agent's tasks, along with the saliency and attention thresholds described below. Executive functions do not communicate with each other directly: each function publishes updates to the GW and subscribes to the events, *internal signals* in CATALINA terminology, that are relevant to its task. Of the nine executive functions in version 0.3.1, four are central to the present work and deserve a brief individual mention.

The *Perception Executive Function* interfaces the agent with the external environment. It is composed of three modules: *Perception Processing* controls the sensors and performs raw data acquisition and enhancement; *Information Extraction* extracts a semantic representation from the raw data, generating a stimulus and *assigning the corresponding saliency value*; *Stimulus Filtering* compares the saliency of each stimulus with the current thresholds and the inhibition regions, and posts to the GW only those stimuli that pass both checks. Information Extraction is where saliency assignment takes place, and is the focus of this paper. Stimulus Filtering, by contrast, is generic and remains untouched by our proposal.

The *Attention Selection Executive Function* maintains the two thresholds that govern attention. Its *Agent Focus* module raises the Saliency and Attention Thresholds whenever the agent commits to a new intention, with values that depend on the saliency of that intention; its *Agent Unfocus* module restores the default low values whenever the agent ceases to pursue any intention.

The *Desire Handler Executive Function* promotes standing desires to active desires (and demotes them back) on the basis of their precondition and their saliency, comparing the latter against the current Saliency Threshold (when the desire is not inhibited) or against the Attention Threshold (when the desire is inhibited). Standing exogenous desires are activated by an incoming stimulus whose saliency, in conjunction with the desire's precondition, is sufficient to overcome the threshold relevant to the agent's current attentional state.

The *Inhibition Executive Function* maintains, whenever the agent is focused, a set of *inhibition regions* over the perception space and the corresponding sets of inhibited beliefs and inhibited desires. The inhibition regions identify which portions of the incoming stimulus stream are not relevant to the current intention and should therefore be filtered out unless their saliency is high enough to break through.

Desire lifecycle and the two thresholds. Each desire in CATALINA progresses through a three-stage lifecycle. A *standing desire* is one that the analyst has injected into the agent, but that has not been activated yet because it is either not salient enough or the preconditions are not met. An *active desire* is a standing desire whose precondition is verified and whose saliency is greater than the relevant threshold; once active, the desire becomes a candidate for promotion to intention by the Reasoner. A *satisfied desire* is one that has been fulfilled: the option used to fulfil it is stored in the plan library for future reuse, and the desire is moved to long-term memory.

The progression from standing to active is governed by the comparison between a desire's saliency and one of two thresholds, both maintained in the GW. The *Saliency Threshold* is the value a desire's saliency must exceed for the desire to be activated when the agent is not focused (or when the desire is not inhibited). The *Attention Threshold* is the higher value that applies when the agent is focused on an intention and the desire under consideration is inhibited; in this case, the desire is activated only if its saliency is sufficient to overcome the agent's deliberate effort to remain on task. The presence of two

thresholds enables CATALINA to balance task-driven persistence with responsiveness to unexpected events: the agent maintains focus when nothing relevant occurs, but salient stimuli can still break through and reorient behaviour.

The same logic applies, with the obvious adaptations, to the activation of standing exogenous desires from incoming stimuli. The saliency of a stimulus is compared against the Saliency Threshold (when the stimulus comes from an uninhibited region of the perception space, or when the agent is not focused) or against the Attention Threshold (when the stimulus comes from an inhibited region while the agent is focused on an intention). Whichever path is taken, the decision to admit the stimulus to the GW is made on the basis of a single scalar value: the stimulus saliency.

Endogenous and exogenous attention. CATALINA adopts the standard psychological distinction between *endogenous* attention, driven by the agent's own goals and intentions, and *exogenous* attention, captured by the inherent features of external stimuli [25, 26]. Endogenous attention manifests in the architecture as the activation of standing endogenous desires whose saliency overcomes the relevant threshold; exogenous attention manifests as the activation of standing exogenous desires triggered by an incoming stimulus whose saliency overcomes the corresponding threshold. The two mechanisms share a similar algorithm, the comparison of saliency against a threshold, but differ in the source of the saliency value being compared: a desire's saliency in the endogenous case, a stimulus's saliency in the exogenous case. Because both mechanisms reduce to a similar comparison, any improvement in how stimulus saliency is computed propagates uniformly through the entire attention pipeline, requiring very limited changes.

3.2. Saliency in CATALINA Today: Definition, Computation, Use

The treatment of saliency in CATALINA is asymmetric. Downstream of the Information Extraction module, the architecture is rich: the Stimulus Filtering module implements a precise, threshold-based admission rule, the two thresholds maintained in the GW support a clean distinction between focused and unfocused states, the inhibition regions allow context-dependent masking of irrelevant perceptions, and the publish-subscribe propagation of stimuli arrivals to the Desire Handler ensures that exogenous attention reaches the deliberation mechanism without bottlenecks. Upstream of the same module, however, the architecture is less refined: the saliency value of a stimulus is whatever the analyst decided perception should be worth, and the architecture treats it as a primitive datum. This is the part of the architecture we aim to improve in this paper.

Where the saliency value comes from. The responsibility of the Information Extraction module consists in assigning the appropriate salience to the perception and the generated belief, thereby forming a stimulus. For instance, this module will assign low saliency to the perception of an accident in the opposite lane, while assigning high saliency to the perception of brake lights from the vehicle ahead. Indeed, the current version of CATALINA only defines the method skeletons for this module, leaving it to the programmer to properly define them according to the problem at hand. In other words, the current implementation specifies the *interface* of the saliency-assignment step - a stimulus carries a saliency value - but not its *contents*: the saliency value of any specific perception is supplied by the analyst, embedded in domain-specific code, and remains constant throughout the agent's lifetime. The programmer is asked to author, by inspection, a fixed mapping from perception types to saliency values, and that mapping is the only source of the numbers that drive the entire downstream attention pipeline.

How saliency propagates through the architecture. Once a stimulus has been emitted by Information Extraction, its saliency value participates in three architectural mechanisms. First, in Stimulus Filtering, it is compared against the Saliency Threshold or the Attention Threshold (depending on the agent's focus state and on whether the stimulus comes from an inhibited region) and either admitted to the GW or discarded. Second, once the stimulus has been admitted, it can act as the trigger for activating a standing exogenous desire associated with that stimulus, again by comparison against the

relevant threshold; the activated desire then enters the Reasoner’s pool of candidates for promotion to intention. Third, the saliency of an admitted stimulus indirectly contributes to the agent’s attention dynamics: a sufficiently salient stimulus can trigger a desire that, once promoted to intention, causes the Agent Focus module to recompute the thresholds and may force the demotion of previously active desires. Stimulus saliency, therefore, carries the entire weight of the architecture’s reactive responses to the environment: it is the only quantity that decides whether a perception is broadcast to the rest of the agent and, ultimately, whether the agent reorients its behaviour in response to it.

An illustration on the AVA case study. In [4] we introduced the case study of a simulated autonomous vehicle agent (AVA) with several practical desires (*Visit Paris*, *Visit Frankfurt*, *Visit Rome*, *Refuel*), an epistemic desire (*Inquiry Estimated Delay*, with saliency 0.95) and some quality and green desires. The agent operates in a perception space that includes events such as the vehicle ahead’s brake lights, the state of traffic lights, road closures along the planned route, and updates to the fuel level. In the current implementation, every one of these perceptions arrives at Information Extraction with a saliency that has been hand-coded by the analyst: the analyst inspects each perception type, decides what the saliency “ought to be”, writes it into the module, and from that moment on every brake-light perception (regardless of whether the agent expected it or not, regardless of whether the lane ahead was already known to be congested) carries the same numerical value.

This produces qualitatively wrong behaviour in two recurring cases. The first case is the *redundant perception*: the agent already believes that the lane ahead is clear with high confidence, and a new sensor reading confirms exactly this. The hand-coded saliency of the “lane status” perception is non-zero, so the perception passes through the architecture, competes against other candidates for admission to the GW, and possibly displaces stimuli that would actually have changed the agent’s view of the world. The architecture spends attention on a perception that, in information-theoretic terms, brings nothing new. The second case is the *context-blind saliency*: the same brake-light perception is assigned the same fixed saliency regardless of whether the agent already knows that the vehicle ahead has been braking for some time, in which case the perception is largely confirmatory, or whether it follows a long sequence of cruise-speed observations, in which case the perception is genuinely surprising. The hand-coded value collapses these two qualitatively different situations into one and prevents the architecture from giving them the differential treatment they deserve.

The subsequent sections of this paper take that statement at face value and propose a principled, architecture-internal definition of stimulus saliency that replaces the hand-coded value with a quantity the agent can compute from its own state.

The replacement is, by design, surgically local. Nothing in the downstream architecture (the Stimulus Filtering algorithm, the Saliency and Attention Thresholds, the inhibition regions, the Desire Promotion mechanism, the Reasoner) needs to be modified. The only architectural change is a single new computation inside the Information Extraction module, together with a small but significant extension of the belief representation that we discuss in Section 4.2.

4. Improving Saliency Definition

As already discussed, the saliency value of a stimulus in CATALINA is the sole factor that determines whether that stimulus will compete for the global workspace, and this saliency carries the full weight of the architecture’s attention dynamics. Yet that value is fixed a priori by the system analyst; it does not depend on what the agent already believes nor on whether the perception adds anything to the agent’s state of knowledge. As argued in Section 2.1, this is precisely the kind of stimulus-only saliency that Itti and Baldi [1] showed to be empirically inferior to a belief-relative measure. The aim of this section is to replace the hand-coded saliency value with one derived from Bayesian surprise, while preserving every other aspect of CATALINA’s existing pipeline.

4.1. Why Bayesian surprise?

Three properties of Bayesian surprise make it a natural candidate for the saliency value of a stimulus in a BDI architecture. First, it is *observer-relative*: the same incoming perception receives different saliency values for two agents whose belief bases differ, which is the desired behaviour for any agent that operates with its own world model rather than as a passive sensor. Second, it is *computable from quantities the architecture already maintains*; the only ingredient required is a representation of the agent’s beliefs that supports a notion of belief change, which we discuss in Subsection 4.2. Third, it is *principled and non-arbitrary*: the value is not a tunable parameter but a divergence between two distributions, with a clear information-theoretic meaning.

A fourth property, which becomes important in Subsection 4.3, is that Bayesian surprise is *additive across independent belief updates*. If a single perception simultaneously affects several beliefs, the total surprise is the sum of the per-belief contributions, provided the affected beliefs are conditionally independent. This makes the resulting saliency value structurally compatible with the existing CATALINA stimulus, which already supports the case in which a perception updates more than one belief in a single cognitive cycle.

By contrast, alternative information-theoretic quantities are less suitable. Shannon information [18] measures the unpredictability of a symbol stream, not the change it induces in the observer’s beliefs, and it would assign maximal saliency to unstructured noise, which is not properly aligned with the behaviour required by an attention mechanism, as Itti and Baldi [1] argued. Cross-entropy or likelihood-based scores require a generative model of the world that CATALINA does not maintain. The Kullback-Leibler divergence [27] between the posterior and the prior over the agent’s belief base is, in this sense, the minimal information-theoretic quantity that captures “how much the agent’s view of the world has changed”.

4.2. Formal definition adapted to the BDI belief base

The proposal requires one preliminary change to the way beliefs are represented in CATALINA. In the classical BDI model, a belief is either held or not held by the agent; in symbols, its truth value is in $\{0, 1\}$. To make a notion of *change in belief* usable as a quantitative saliency, we attach to each belief b_i a *confidence value* $c_i \in [0, 1]$, interpreted as the agent’s degree of belief that the proposition is true. The intuitive reading is the familiar one: $c_i = 1$ means the agent is certain that b_i holds, $c_i = 0$ means the agent is certain that b_i does not hold, $c_i = 0.5$ means complete uncertainty, and intermediate values express graded confidence in either direction. The classical Boolean case is the special case $c_i \in \{0, 1\}$, so the existing belief base, the inhibition regions, and every belief-handling component of the architecture continue to work without modification on legacy beliefs.

Saliency from a single-belief update. Consider a perception p that affects exactly one belief b_i , moving its confidence from a prior value c_i to a posterior value c'_i . The saliency contribution of that single update, expressed in nats, is

$$S_i = c'_i \log \frac{c'_i}{c_i} + (1 - c'_i) \log \frac{1 - c'_i}{1 - c_i}. \quad (2)$$

The formula has a simple structure: each of the two possible outcomes of the belief, “true” and “false”, contributes one term. The contribution of the “true” outcome is the new probability of that outcome (c'_i) multiplied by the logarithm of how much the probability of that outcome changed (c'_i/c_i); the contribution of the “false” outcome is the analogous quantity for the false side. The two contributions are then added. Mathematically, Eq. (2) is the Kullback–Leibler divergence [27] between two Bernoulli distributions of parameters c'_i and c_i , a standard information-theoretic quantity that measures how different the posterior coin-flip is from the prior coin-flip in the units, nats, of natural logarithms.

Suppose the agent believes that the brake lights of the vehicle ahead are off with confidence $c_i = 0.85$, and the new perception “brake lights on” updates this to $c'_i = 0.05$. Plugging in:

$$S_i = 0.05 \log_{0.85} \frac{0.05}{0.85} + 0.95 \log_{0.15} \frac{0.95}{0.15} \approx -0.14 + 1.75 \approx 1.61 \text{ nats.}$$

The number is large because the perception caused a substantial swing in the agent’s view of the world: from being almost sure the brake lights were off to being almost sure they are on.

Saliency from a multi-belief update. A perception may simultaneously affect more than one belief, for example, the report “road closure ahead” might revise both the “route is clear” belief and a “time to destination” belief. The proposed saliency of such a perception is simply the sum of the per-belief contributions:

$$S(p) = \sum_{i \in \mathcal{A}(p)} S_i, \quad (3)$$

where $\mathcal{A}(p)$ is the set of beliefs affected by p . If only one belief is affected, the sum collapses to a single term and $S(p) = S_i$.

The additive form of Eq. (3) rests on a standard *independence assumption* on the agent’s belief base: the agent does not represent correlations between the truth values of different beliefs, and treats them as independent. Formally, if we let P_B denote the agent’s joint distribution over belief configurations before the perception and P'_B the corresponding distribution after, the assumption is that both factorise as products of independent Bernoullis; the KL divergence between two such product distributions equals the sum of the per-component KLs, which is exactly Eq. (3). The assumption is reasonable in the present setting because CATALINA does not currently maintain explicit correlations between beliefs; the proposal therefore inherits an existing simplification rather than introducing a new one. Future work that equips CATALINA with correlated beliefs would also need to revisit this step, but the core of the proposal, a saliency value derived from belief change, is unaffected.

Two corner cases. Two limiting behaviours of Eq. (2) are worth noting because they are precisely the behaviours we want from a saliency value. First, if the perception leaves a belief untouched ($c'_i = c_i$), both terms of the formula vanish, and that belief contributes exactly zero to the sum. A perception that confirms what the agent already believes, therefore costs zero saliency, which is the expected outcome. Second, if a belief is pushed from a mid-range confidence to a value very close to 0 or 1, the per-belief contribution grows rapidly because the logarithm of the ratio diverges as the denominator approaches the boundary. This captures the intuition that *committing to certainty* about a previously uncertain belief is a much larger update than reshuffling confidence in the middle of the unit interval.

4.3. Architectural integration in CATALINA

The replacement of the hand-coded saliency by the Bayesian-surprise saliency of Eq. (3) is local to the Information Extraction module of the Perception Executive Function described in Section 3. The other two modules of the Perception function, Perception Processing and Stimulus Filtering, are not affected. In particular, the Saliency Threshold and the Attention Threshold, which decide whether a stimulus is admitted and whether the agent is in a focused or unfocused state, continue to operate exactly as in the current implementation; only the meaning and the provenance of the value they compare against change.

Concretely, Information Extraction is now responsible for three operations on each incoming perception. (i) It identifies which beliefs in the current belief base B are affected by the perception. In the simplest implementation, this is a lookup against the agent’s predicate ontology, reusing the same $\langle \text{subject}, \text{relationship}, \text{complement} \rangle$ predicate vocabulary that CATALINA already adopts. (ii) For each affected belief b_i , it estimates the posterior confidence c'_i that follows from the perception. The computation of c'_i is not part of the saliency definition itself, but of the belief-update policy implemented by the Information Extraction module. In the simplest case, when the perception directly observes

the truth value of a predicate, c'_i is obtained by assigning a high or low confidence according to the perceived value and the reliability of the corresponding sensor. For example, if a sensor with reliability r reports that predicate b_i is true, the posterior confidence can be set to $c'_i = r$; if it reports that b_i is false, it can be set to $c'_i = 1 - r$. More generally, when a perception is noisy or only indirectly related to a belief, c'_i can be obtained by a standard Bayesian update:

$$c'_i = P(b_i | p) = \frac{P(p | b_i)c_i}{P(p | b_i)c_i + P(p | \neg b_i)(1 - c_i)}. \quad (4)$$

The likelihood terms $P(p | b_i)$ and $P(p | \neg b_i)$ encode the reliability of the perceptual channel or the domain model used by the agent. Thus, the proposed saliency mechanism does not require a single universal update rule: it requires only that Information Extraction expose, for each affected belief, the prior–posterior pair (c_i, c'_i) . The KL computation then measures the magnitude of that update independently of how the posterior was produced. (iii) It computes the per-belief contributions of Eq. (2) and emits a stimulus $\langle \text{belief}, S(p) \rangle$ to the Stimulus Filtering module, in which the belief carries the standard CATALINA predicate format and $S(p)$ is the saliency value computed by Eq. (3).

A useful side-effect of this integration is that the resulting saliency value is *automatically zero* for perceptions that do not change any belief, irrespective of the surface form of the perception. A redundant sensor reading that confirms what the agent already believes does not enter the global workspace, even if its hand-coded saliency in the original implementation would have made it competitive: the architecture’s attention is, by construction, spent only on perceptions that carry new information about the agent’s state of the world.

4.4. Worked example on the Autonomous Vehicle Agent case study

To make the proposal concrete, consider five perceptions from a simulated case study involving an Autonomous Vehicle Agent (hereafter, AVA) with several competing desires (visiting several cities, respecting time constraints, pollution limits, and so on). Let us suppose the AVA may have the following perceptions:

- p_1 “The vehicle ahead has its brake lights on.”
- p_2 “The traffic light just turned red.”
- p_3 “Road closure ahead on the planned route.”
- p_4 “Fuel level updated: 47%.”
- p_5 “A child’s ball just rolled into the lane 30 m ahead.”

The agent maintains five beliefs with prior confidences $(c_1, \dots, c_5) = (0.85, 0.70, 0.90, 0.95, 0.90)$, corresponding respectively to the propositions “brake lights of the vehicle ahead are off”, “the traffic light is green”, “the planned route is clear”, “the fuel level is sufficient” and “the lane ahead is clear”. After applying the conditional update sketched in Subsection 4.3, the affected belief in each case is updated to a posterior confidence of approximately 0.05, 0.05, 0.02, 0.95, and 0.05 respectively (the fuel reading is above the sufficiency threshold, so the belief remains unchanged). Plugging these values into Eq. (2) gives the saliency values of each perception as shown in Table 1.

Three observations are worth noting. First, the saliency ranking ($p_3 > p_5 > p_1 > p_2 > p_4$) reflects the relative magnitude of the belief change rather than any analyst-supplied ordering: the perception that overturns the strongest prior, the road closure overturning a confidence of 0.90 in a clear route, is automatically the most salient. Second, the uninformative fuel update receives saliency exactly zero: the agent’s belief about fuel sufficiency is unchanged by an observation within the sufficiency band, so the perception consumes no attention. Third, the open-world perception p_5 is handled by the same mechanism without any special treatment: as long as the perception is mapped to the existing belief about lane clearance, the saliency value follows from the same KL formula. This last point illustrates a limited form of open-world readiness: the mechanism can assign saliency to perceptions that were

Table 1

Saliency values produced by the proposed mechanism on five perceptions of the AVA case study. All values are in nats.

Perception	Affected belief	c	c'	S_i	$S(p)$
p_1 Brake lights on	brake lights off	0.85	0.05	1.612	1.612
p_2 Red light	traffic light green	0.70	0.05	0.963	0.963
p_3 Road closure ahead	route clear	0.90	0.02	2.161	2.161
p_4 Fuel update (47%)	fuel sufficient	0.95	0.95	0.000	0.000
p_5 Child’s ball in lane	lane clear	0.90	0.05	1.994	1.994

not explicitly enumerated by the analyst, provided that Information Extraction can map them onto one or more beliefs already represented in the agent’s belief base. This mapping step is non-trivial and remains a responsibility of the perception and semantic-interpretation layer; the present proposal should therefore be understood as making the saliency computation open-world-compatible, rather than solving open-world perception in general.

4.5. Numerical considerations

A few numerical remarks are necessary before the proposed mechanism can be deployed in practice. First, the per-belief Bernoulli KL of Eq. (2) diverges if either c or c' is exactly 0 or 1. A standard remedy is to clamp confidences to a small interval $[\varepsilon, 1 - \varepsilon]$ with ε on the order of 10^{-6} ; the effect on the resulting saliency value is negligible for any realistic update, and the clamp prevents the belief base from acquiring unrevisable certainties that the architecture cannot subsequently update with new evidence. Second, when several beliefs are simultaneously affected by a perception, the additive decomposition of Eq. (3) can produce saliency values substantially larger than the per-belief maximum; if the downstream Saliency Threshold has been tuned for the current single-belief implementation, it will need to be recalibrated. Third, the saliency values produced by the proposal are in nats and are therefore unbounded above, whereas the existing CATALINA convention treats saliency as a scalar in $[0, 1]$. For backward compatibility, a monotone squashing function such as $\sigma(S) = 1 - e^{-S/\tau}$ (with τ a hyperparameter) can be applied to the raw KL value before exposing it to the rest of the architecture; this preserves the ordering of stimuli and bounds the saliency value while letting the analyst recover the unbounded version when needed.

4.6. Discussion

The proposal discussed in the previous subsections introduces limited changes to the architecture, a confidence value attached to each belief, and one new computation in the Information Extraction module, the Bernoulli KL of Eq. (2), in exchange for which the agent obtains an observer-relative, context-aware saliency mechanism that is principled and free of hand-coded numbers. Every other aspect of CATALINA, the GW selection rule, the Saliency and Attention Thresholds, the inhibition regions, and the deliberation cycle continue to operate unchanged. This is, in our view, an advantage of the proposed solution to the previous limitations in Saliency assignment for CATALINA: it produces an immediately implementable change with a clear theoretical justification, and it leaves room for further extensions without committing the architecture to any particular direction.

Two such extensions are worth flagging here. The first is the introduction of a goal-driven term that weights the surprise value by the relevance of the affected beliefs to the agent’s currently held desires, in line with factor 5 of Wolfe and Horowitz [13] and the multiplicative gain mechanism of Reynolds and Heeger [28]; this extension lifts the proposal from a one-factor saliency model to a two-factor one and brings the formalism closer to the structure of expected free energy in active inference [14], without requiring the architecture to be recast in active-inference terms.

The second extension is to enrich the per-belief update step beyond the conditional assignment used in the worked example, for instance, by exploiting a learned or LLM-based belief-change estimator

constrained to the agent’s predicate ontology. In that case, however, the posterior-estimation policy itself should remain inspectable: explanation techniques such as SHAP and LIME, which have already been used to analyse and assess predictive models in application domains such as smart agriculture [29], could help make explicit which perceptual features contribute to a given posterior confidence assignment. Both extensions are deferred to future work and discussed briefly in Section 5.

5. Conclusion and Future Work

This paper has proposed a principled, observer-relative definition of stimulus saliency for the CATALINA cognitive agent architecture. The hand-coded, agent-independent saliency value attached to a stimulus by the Information Extraction module of the Perception Executive Function is replaced by the Kullback-Leibler divergence between the agent’s posterior and prior beliefs after the perception, following the Bayesian-surprise formulation of Itti and Baldi [1]. The replacement requires a single architectural change, equipping each belief with a confidence value in $[0, 1]$ rather than just a Boolean truth value, an extension that the probabilistic-BDI literature has already developed, and one new computation inside Information Extraction. Every other aspect of CATALINA (Stimulus Filtering, the Saliency and Attention Thresholds, the inhibition regions, the Desire Promotion mechanism, the Reasoner) continues to operate exactly as in the current implementation. The proposal modifies the *provenance* of the saliency value, not its *role*, and is therefore a minimal, surgical retrofit of the existing architecture.

Three behavioural changes follow directly from the proposal, each addressing one of the recurring failure modes of the hand-coded scheme discussed in Section 3.2. First, perceptions that confirm what the agent already believes receive a saliency value of exactly zero and are therefore filtered out before they reach the Global Workspace; the architecture’s attention is spent only on perceptions that carry new information about the agent’s state of the world, irrespective of how the analyst would have rated their importance. Second, the same perception receives different saliency values depending on the agent’s prior beliefs, so that a brake-light event that follows a long sequence of cruise-speed observations is correctly recognised as more salient than the same event following an established braking pattern; saliency becomes context-aware without any change to the perception ontology or to the analyst’s design effort. Third, the proposed strategy can assign saliency to perceptions that lie outside the analyst’s original perception set, provided that the Information Extraction module can map them onto existing beliefs in the agent’s belief base. This does not solve the general problem of open-world perception: the semantic mapping from novel perceptions to internal predicates remains a non-trivial task. Rather, the contribution is that once such a mapping is available, saliency no longer needs to be hand-coded for every possible perception type.

The proposal captures only the information-theoretic component of saliency, how much the perception changes the agent’s view of the world, while leaving out the goal- and value-driven modulation that the modern visual-search literature [13] and the active-inference framework [14, 15] both identify as necessary for a complete account of attention. This limitation is important in practical BDI settings. A perception may produce a large belief update while being irrelevant to the agent’s current desires, and conversely a small belief update may concern a belief that is crucial for the currently pursued intention. Therefore, Bayesian surprise should not be interpreted as a complete model of attentional priority, but as a principled bottom-up component that can later be combined with goal relevance, desire saliency, or intention-dependent modulation. As a consequence, the current proposal does not distinguish between two perceptions that yield equal Bayesian surprise but bear on beliefs of very different relevance to the agent’s current desires: the formula treats them symmetrically, whereas a more sophisticated mechanism would amplify the saliency of the perception that bears on desire-relevant beliefs. The most immediate and promising extension of this work is a *two-factor* saliency formula in which the Bayesian-surprise term is modulated by a goal-driven term that weights the perception by the relevance of the affected beliefs to the agent’s currently held desires.

A further limitation of the present work is that the proposal has been evaluated only analytically and through a worked example. A systematic simulation-based comparison with the current hand-coded

saliency mechanism and alternative information-based saliency measures is left for future work. Such an evaluation should measure, for example, how often irrelevant perceptions enter the Global Workspace, how quickly genuinely belief-changing perceptions break through attention thresholds, and how often the agent inappropriately reorients its intentions.

Finally, a longer-term direction is to carefully position the proposal in relation to active inference [14, 15, 22]. The two-factor extension brings the formalism close enough to the active-inference decomposition that a principled discussion becomes worthwhile, both to clarify what CATALINA loses by not committing to a unified normative framework and to make explicit what it gains, in terms of compatibility with the existing symbolic BDI vocabulary and with the publish-subscribe Global Workspace.

Acknowledgments

We would like to acknowledge the great contribution that the discussions we had with John Mylopoulos and Myrto Mylopoulos about CATALINA, its possible evolutions, advantages, and limitations, had to the conception of this paper. They are co-authors of previous papers on CATALINA and played a fundamental role in defining this architecture.

This work has been partially supported by the EU Project SYNAPSE, Horizon Europe, under Grant Agreement No 101120853.

Declaration on Generative AI

During the preparation of this work, the authors used Grammarly and Claude to paraphrase and reword, perform grammar and spelling checks. After using these tools/services, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

- [1] L. Itti, P. Baldi, Bayesian surprise attracts human attention, *Vision Research* 49 (2009) 1295–1306.
- [2] M. E. Bratman, D. J. Israel, M. E. Pollack, Plans and resource-bounded practical reasoning, *Computational Intelligence* 4 (1988) 349–355.
- [3] M. Cossentino, G. Pilato, G. Aversa, M. Mylopoulos, J. Mylopoulos, Practical reasoning with attention mechanisms, in: *Proc. of the IEEE SIMPAR Int. Conf. on Simulation, Modeling, and Programming for Autonomous Robots*, IEEE Press, 2025.
- [4] M. Cossentino, G. Aversa, G. Pilato, M. Mylopoulos, J. Mylopoulos, The agent working cycle in CATALINA, in: *Proc. of WOA 2025 – 26th Workshop From Objects to Agents*, CEUR Workshop Proceedings, 2025.
- [5] M. Cossentino, G. Pilato, G. Aversa, M. Mylopoulos, J. Mylopoulos, The agent metamodel in CATALINA (cognitive agent practical reasoning architecture), in: *Proc. of the 22nd European Conference on Multi-Agent Systems (EUMAS 2025)*, 2025.
- [6] B. J. Baars, In the theatre of consciousness. Global workspace theory, a rigorous scientific theory of consciousness, *Journal of Consciousness Studies* 4 (1997) 292–309.
- [7] B. J. Baars, Global Workspace Theory of consciousness: toward a cognitive neuroscience of human experience, *Progress in Brain Research* 150 (2005) 45–53.
- [8] D. Buehler, Psychological Agency – Guidance of Visual Attention, Ph.D. thesis, UCLA, 2014.
- [9] D. Buehler, The central executive system, *Synthese* 195 (2018) 1969–1991.
- [10] D. Buehler, Agential capacities: a capacity to guide, *Philosophical Studies* 179 (2022) 21–47.
- [11] Álvaro Carrera, C. Iglesias, B2DI: A Bayesian BDI agent model with causal belief updating based on MSBN, in: *Proceedings of the 4th International Conference on Agents and Artificial Intelligence (ICAART)*, 2017.

- [12] B. Archibald, M. Sevegnani, M. Xu, Modelling and verifying BDI agents under uncertainty, *Science of Computer Programming* 242 (2024).
- [13] J. M. Wolfe, T. S. Horowitz, Five factors that guide attention in visual search, *Nature Human Behaviour* 1 (2017) 0058.
- [14] K. Friston, T. FitzGerald, F. Rigoli, P. Schwartenbeck, G. Pezzulo, Active inference: A process theory, *Neural Computation* 29 (2017) 1–49.
- [15] T. Parr, K. J. Friston, Generalised free energy and active inference, *Biological Cybernetics* 113 (2019) 495–513.
- [16] L. Itti, C. Koch, E. Niebur, A model of saliency-based visual attention for rapid scene analysis, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20 (1998) 1254–1259.
- [17] L. Itti, C. Koch, Computational modelling of visual attention, *Nature Reviews Neuroscience* 2 (2001) 194–203.
- [18] C. E. Shannon, A mathematical theory of communication, *Bell System Technical Journal* 27 (1948) 379–423.
- [19] B. W. Tatler, M. M. Hayhoe, M. F. Land, D. H. Ballard, Eye guidance in natural vision: Reinterpreting salience, *Journal of Vision* 11 (2011) 5.
- [20] K. Friston, The free-energy principle: A unified brain theory?, *Nature Reviews Neuroscience* 11 (2010) 127–138.
- [21] P. Schwartenbeck, T. FitzGerald, R. J. Dolan, K. Friston, Exploration, novelty, surprise, and free energy minimization, *Frontiers in Psychology* 4 (2013) 710.
- [22] G. Pezzulo, E. Cartoni, F. Rigoli, L. Pio-Lopez, K. Friston, Active inference, epistemic value, and vicarious trial and error, *Learning & Memory* 25 (2018) 309–319.
- [23] S. Franklin, F. G. J. Patterson, The LIDA architecture: Adding new modes of learning to an intelligent, autonomous, software agent, *Integrated Design and Process Technology, IDPT-2006 Proceedings* (2006).
- [24] S. Franklin, T. Madl, S. D’Mello, J. Snider, LIDA: A systems-level architecture for cognition, emotion, and learning, *IEEE Transactions on Autonomous Mental Development* 6 (2014) 19–41.
- [25] J. Theeuwes, Top-down and bottom-up control of visual selection, *Acta Psychologica* 135 (2010) 77–99.
- [26] F. De Brigard, J. Prinz, Attention and consciousness, *Wiley Interdisciplinary Reviews: Cognitive Science* 1 (2010) 51–59.
- [27] S. Kullback, R. A. Leibler, On information and sufficiency, *The annals of mathematical statistics* 22 (1951) 79–86.
- [28] J. H. Reynolds, D. J. Heeger, The normalization model of attention, *Neuron* 61 (2009) 168–185.
- [29] A. Cartolano, A. Cuzzocrea, G. Pilato, Analyzing and assessing explainable ai models for smart agriculture environments, *Multimedia Tools and Applications* 83 (2024) 37225–37246.