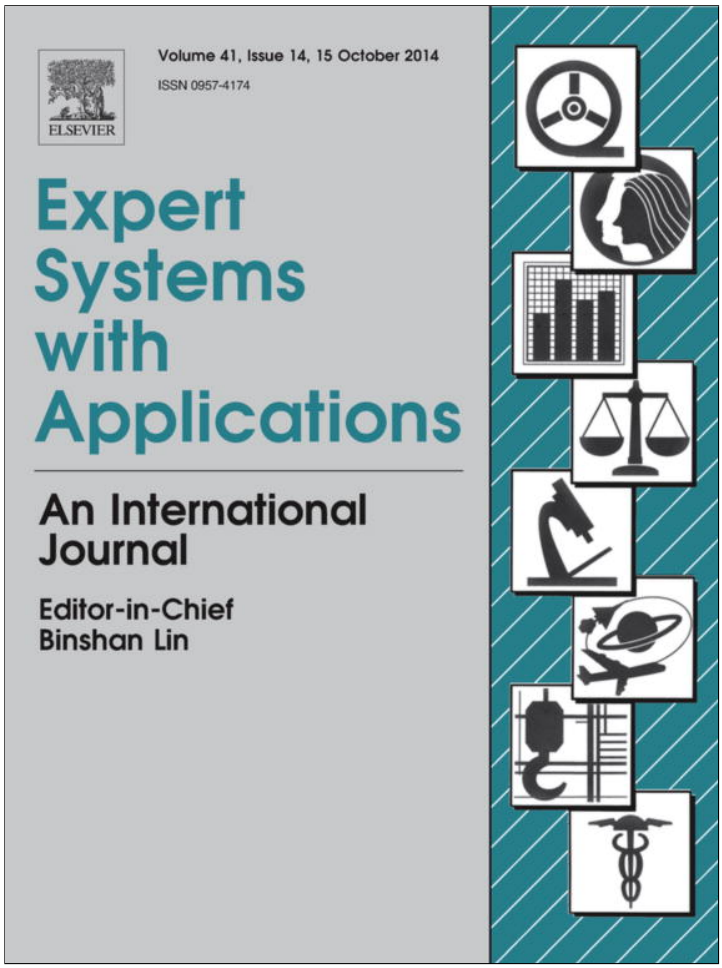

Expert Systems with Applications 41 (2014) 6327–6345

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Behavior monitoring under uncertainty using Bayesian surprise and optimal action selection

Luis Avila, Ernesto Martínez *

INGAR (CONICET-UTN), Avellaneda 3657, Santa Fe S3002 GJC, Argentina

ARTICLE INFO

Keywords:
Artificial pancreas
Bayesian surprise
Behavior monitoring
Kullback–Leibler divergence
Optimal action selection

ABSTRACT

The increasing trend towards delegating tasks to autonomous artificial agents in safety–critical socio-technical systems makes monitoring an action selection policy of paramount importance. Agent behavior monitoring may profit from a stochastic specification of an optimal policy under uncertainty. A probabilistic monitoring approach is proposed to assess if an agent behavior (or policy) respects its specification. The desired policy is modeled by a prior distribution for state transitions in an optimally-controlled stochastic process. Bayesian surprise is defined as the Kullback–Leibler divergence between the state transition distribution for the observed behavior and the distribution for optimal action selection. To provide a sensitive on-line estimation of Bayesian surprise with small samples twin Gaussian processes are used. Timely detection of a deviant behavior or anomaly in an artificial pancreas highlights the sensitivity of Bayesian surprise to a meaningful discrepancy regarding the stochastic optimal policy when there exist excessive glycemic variability, sensor errors, controller ill-tuning and infusion pump malfunctioning. To reject outliers and leave out redundant information, on-line sparsification of data streams is proposed.

1. Introduction

Safety–critical systems integrate software, hardware and humans in an increasing number of applications that are prone to failures, errors and malfunctioning which could result in loss of life, significant property damage, or damage to the environment. Moreover, the increasing trend towards delegating tasks to autonomous artificial agents in safety–critical socio-technical systems makes monitoring an action selection policy of paramount importance. As an example, consider the case of collision avoidance in driving systems (Broggi, Medici, Zani, Coati, & Panciroli, 2012) where the monitoring task involves a number of autonomous vehicles interacting with each other in a high-speed highway. Any monitoring system aimed to warn or prevent collisions and dangerous circumstances must contemplate the expected behavior of nearby cars to detect quickly a collision scenario. However, monitoring tasks are often formulated around the idea of isolated agents with perfect rationality (Thimbleby, 2009). For on-line traffic monitoring, the key aspect is to characterize the uncertain environment the autonomous car is in and its desired optimal behavior. Similarly, there is an increased requirement for condition monitoring of nuclear power plants to ensure they are still able to operate safely, yet efficiently (West, McArthur, & Towle, 2012). Decision support to detect anomalies is limited by the availability of expert knowledge and the variability of the plant conditions. Proper control and on-line monitoring of the interaction between the operators and the plant would be helpful to prevent catastrophic accidents (Salge & Milling, 2006). In a different field, researchers have evaluated different solutions to automate the task of gazing at a monitor to find suspicious behaviors in video surveillance (Fernández-Caballero, Castillo, & Rodríguez-Sánchez, 2012). Detecting dangerous objects and intruders is essential for safety in crowded environments, but monitoring human behaviors and reporting about anomalies is a complex task for any computing system.

Current automated systems function well in environments they are designed for, that is, around their nominal operating conditions or expected scenarios. They also perform well in environments with "predictable" uncertainties as treated, for example, in the advanced adaptive and robust control frameworks. Nevertheless, control systems of today require substantial human intervention when faced with novel and unanticipated situations, i.e. situations that have not been considered at the design stage. Such situations can arise from discrete changes in the environment, extreme disturbances, structural changes in the system (for example, as a result of damage), and the like. More specifically, biological control systems such as the artificial pancreas (AP) must face significant

* Corresponding author. Tel.: +54 (342) 4534451; fax: +54 (342) 4553439.
E-mail address: ecmarti@santafe-conicet.gov.ar (E. Martínez).

http://dx.doi.org/10.1016/j.eswa.2014.04.031
0957-4174/

Nomenclature

Symbols for the glycemic model

λ	drift parameter
σ	variability parameter
BG	blood glucose level
G_{in}	systemic appearance of glucose via glucose absorption from the gut
$NHGB$	net hepatic glucose balance
G_{out}	overall rate of peripheral and insulin dependent glucose utilization
G_{ren}	excretion of glucose
V_G	volume of distribution of glucose
Sh	hepatic sensitivity
Sp	insulin sensitivity
I	insulin infusion level
IG	interstitial glucose
τ	sensor time-lag parameter
ξ	sensor calibration parameter
k_C	PID proportional gain
τ_I	PID integral time
τ_D	PID derivative time

Symbols for the control algorithm

m	mean function
cov	covariance function
$p(\bullet\|\bullet,\bullet)$	controlled transition probability
$\hat{\mathbf{x}}$	state estimation
Δx	state change
u	control action
$\pi(\bullet)$	policy
$V(\bullet)$	state function value
$Q(\bullet,\bullet)$	state-action function value
$\ell(\bullet,\bullet)$	immediate cost
$h(\bullet\|\bullet)$	passive dynamics
γ	discount factor
r	reward function
GP	Gaussian process
ϑ	cardinality
ρ	exploration parameter
β	exploitation parameter
$\mathfrak{x}$	state space
U	action space

Symbols for monitoring task

$P(\bullet)$	prior probability
$P(\bullet\|\bullet)$	posterior probability
$KL(\bullet\|\|\bullet)$	Kullback–Leibler distance
P_{KL}	pointwise Bayesian surprise
T_{KL}	robust Bayesian surprise
$\mathrm{k}(\bullet,\bullet)$	kernel
δ	stop threshold
η	threshold for the level of sparsity
$N_{\max}$	size of the training set used to model the Gaussian process
$\mathfrak{D}$	dictionary

levels of variability. When an action is executed by an agent, the perceived result of the action depends on the environmental response, including other agents, noisy measurements, hidden states and the quality of the sensory data. In most cases, the agent has only an approximate knowledge of these effects, but it must nevertheless choose a nearly-optimal course of action to accomplish the desired control task (Sanger, 2011). Under uncertainty, a probabilistic characterization of the desired behavior is needed to assess if a given agent behavior respects its specification. Such a specification is an essential element of using Bayesian inference to detect deviations from an optimal control policy.

Almost all of the existing literature about system monitoring, is concerned with the task to make certain controlled variables track given set-points or set-point trajectories, while assuring closed-loop stability. However, the purpose of autonomy (and that of automation as a whole), is not primarily to keep the controlled variables at their set-points as well as possible or to nicely track dynamic set-point changes. For example, a feasible controller for glycemic regulation based on model predictive control has been designed to control to a zone instead of a set-point, which may prevent unnecessary and dangerous overcorrection (Grosman, Dassau, Zisser, Jovanoviĉ, & Doyle, 2010). An important issue is that the agent decision-making policy is mainly focused on the net return which should be maximized in the presence of disturbances and different sources of variability, while exploiting the available noisy and scarce measurements. Thus, behavior monitoring under uncertainty has to be built upon a stochastic process specification of the desired optimal policy.

The novelty and relevance of information contained in new data, can be measured by the effect such data has on the observer (monitor) (Hasanbelliu, Kampa, Principe, & Cobb, 2012). Fundamentally, this effect is to transform the observer's prior beliefs into posterior beliefs, according to the Bayes theorem. The amount of information can be measured in a natural way by the Kullback–Leibler (KL) distance -also called relative entropy- between the prior and posterior distributions in the observer, regarding the available space of hypotheses about the state of a controlled system. This facet of information, termed "surprise," is important in behavior monitoring where beliefs change over time, in particular when malfunctioning causes a deviant behavior. Surprise is a subjective information measure that quantifies how much information a new observation contains, in relation to the current knowledge of the system being monitored (Baldi & Itti, 2010; Itti & Baldi, 2005b). Surprise can exist only in the presence of uncertainty, and it is related to beliefs about the dynamics of state transitions, where the same data convey different amount of surprise to different observers or to the same observer at different times. To quantify the surprise factor of an observation, in this work the novelty of information in a data stream regarding deviations from the specified behavior is measured using twin Gaussian processes (Bo & Sminchisescu, 2010).

Behavior specification under uncertainty is formalized here as a controlled stochastic process that makes the agent policy as close as possible to the desired one by describing both the policy and the state transition dynamics in probabilistic terms. To this aim, *optimal choice of actions under uncertainty* is a fundamental problem to be addressed in order to characterize the desired behavior of an intelligent agent. The abstract setting for the latter can be framed as an agent choosing actions over time, an uncertain dynamical system whose state is affected by those actions, and a performance criterion that the agent seeks to optimize (Todorov, 2009). The agent has the power to reshape the system dynamics in any way it wishes. However, it pays a price for too much reshaping (Dvijotham & Todorov, 2012). The key question for on-line behavior monitoring is how the "distance" from optimal reshaping can be measured using small samples from realizations of a

stochastic process bearing in mind optimal decision-making under uncertainty. In this contribution, a probabilistic approach for on-line behavior monitoring is proposed and a Bayesian surprise metric is defined to measure deviations from a specification based on samples from a given realization of the stochastic process for observed state transitions in the agent's environment.

This article is structured as follows. Section 2 briefly introduces the principles of optimal action selection along with a policy learning algorithm that integrates reinforcement learning and Gaussian processes to characterize the expected optimal behavior of the agent. In Section 3, Bayesian surprise is quantified as the Kullback–Leibler divergence between the prior and posterior distributions in order to pinpoint any deviation from the expected performance. Also in this section, sparsification of the arriving data stream is proposed in order to include only relevant data in the training set used to model the state transition dynamics. In Section 4, an artificial pancreas is used as case study and different scenarios which endanger safe and optimal operation of an AP are considered. Finally, in Section 5, final remarks and future works are discussed.

2. Optimal action selection

2.1. Performance loss

Optimal actions under uncertainty can be compactly defined by the stochastic dynamics $p(\hat{\mathbf{x}}|\mathbf{x},u^p)$. It specifies the optimal transition probability from state $\mathbf{x}$ to state $\hat{\mathbf{x}}$, under a scalar control action u^p, given by an optimal control policy $\pi^p(\mathbf{x})$. As soon as the agent behavior gives rise to a different state transition dynamics $g(\hat{\mathbf{x}}|\mathbf{x},u^g)$ -due to a suboptimal control policy $\pi^g(\mathbf{x})$ which deviates from optimal action selection- it does not conform to its specification. Note that the next state $\hat{\mathbf{x}}$ corresponds to a stochastic variable at instant $k+1$ determined by the state-action pair ($\mathbf{x}$, u) at time k. Since both transition probabilities are computed for the same state $\mathbf{x}$, what makes them different are the control actions resulting from their respective control policies. The rationality behind $p(\hat{\mathbf{x}}|\mathbf{x},u^p)$ is the optimal cost-to-go function $V^p(\mathbf{x})$, defined as the expected cumulative cost for starting at state $\mathbf{x}$ and acting optimally thereafter. It compresses all relevant information about the future and thus enables greedy computation of optimal actions. The value function $V^p(\mathbf{x})$ equals the minimum of the immediate cost plus the expected cost-to-go $\mathbf{E}[V(\hat{\mathbf{x}})]$ from the next state $\hat{\mathbf{x}}$

$$V^p(\mathbf{x}) = \min_u \left\{ \ell^p(\mathbf{x},u^p) + \gamma \cdot \mathbf{E}_{\hat{\mathbf{x}}\sim p(\hat{\mathbf{x}}|\mathbf{x},u^p)}[V^p(\hat{\mathbf{x}})] \right\} \tag{1}$$

where the subscript indicates that the expectation is taken with respect to the transition probability distribution $p(\hat{\mathbf{x}}|\mathbf{x},u^p)$, induced by following the policy that generates the control action u^p. The term $\ell^p(\mathbf{x},u^p)$ represents the cost for being in state x and taking action u^p and γ is the discount factor. Eq. (1) is fundamental to the optimal control theory and is called the Bellman fundamental equation. The cost of reshaping the dynamics of a system can be measured with reference to the passive dynamics characterizing the behavior of the system in the absence of controls (Todorov, 2009). The passive dynamics will usually be defined as a random walk in discrete domains and as a diffusion process in continuous domains. Note that the notion of passive dynamics is common in continuous domains but is rarely used in discrete domains. The passive system dynamics is denoted as $h(\hat{\mathbf{x}}|\mathbf{x})$ in probabilistic terms. The agent can influence this dynamics in any way it wishes. However, it pays a price for reshaping the passive dynamics beyond what is strictly necessary for optimal control. The minimum immediate cost for optimal reshaping can be estimated as follows (Dvijotham & Todorov, 2012).

$$\begin{aligned}\ell^p(\mathbf{x},u^p) &= q(\mathbf{x}) + \mathbf{E}_{\hat{\mathbf{x}}\sim p(\hat{\mathbf{x}}|\mathbf{x},u^p)} \log\left[\frac{p(\hat{\mathbf{x}}|\mathbf{x},u^p)}{h(\hat{\mathbf{x}}|\mathbf{x})}\right] \\ &= q(\mathbf{x}) + KL(p(\hat{\mathbf{x}}|\mathbf{x},u^p)\|h(\hat{\mathbf{x}}|\mathbf{x}))\end{aligned} \tag{2}$$

The state cost $q(\mathbf{x})$ is an arbitrary function encoding how (un)desirable different states are, and KL is the Kullback–Leibler divergence (Kullback & Leibler, 1951) that measures the difference between the optimally-controlled dynamics with respect to the passive one. Eq. (2) can be written as well for any arbitrary shaping associated with $g(\hat{\mathbf{x}}|\mathbf{x},u^g)$ which allows modeling the performance loss for suboptimal reshaping of the passive dynamics $h(\hat{\mathbf{x}}|\mathbf{x})$ as follows

$$\ell^g(\mathbf{x},u^g) - \ell^p(\mathbf{x},u^p) \cong KL(g(\hat{\mathbf{x}}|\mathbf{x},u^g)||p(\hat{\mathbf{x}}|\mathbf{x},u^p)) \tag{3}$$

The KL divergence, or relative entropy, should be understood as a measure of the difficulty of discriminating between two distributions. Entropy quantifies any probability distribution with properties that agree with the intuitive notion of information content. In stochastic optimal control entropy can be directly related to the approximate solution of the Fokker–Planck–Kolmogorov equation (Günel, 2010; Plastino, Miller, & Plastino, 1997). Also, Majda, Kleeman, and Cai (2002) have recently demonstrated that the relative entropy provides a measure of the information content of a prediction ensemble. More specifically, entropy and predictability are intertwined concepts relevant for performance monitoring of control loops (Ghraizi, Martinez, & De Prada, 2009).

2.2. Optimal behavior specification

The key issue to be addressed in behavior monitoring is how optimal action selection can be characterized in the face of uncertainty to readily detect discrepancies between optimal and degraded performance. To this aim, we focus on the novel methodology that combines policy iteration with functional approximations presented by De Paula and Martinez (2012) and Deisenroth, Rasmussen, and Peters (2009). This policy learning methodology (GPRL) integrates Reinforcement Learning (Sutton & Barto, 1998) with Gaussian processes (GP) (Rasmussen & Williams, 2006) to obtain an optimal control policy π^p under uncertainty which guarantees safe operation, and is capable of achieving near-optimal behavior. The optimal control policy is compactly represented using the hyper-parameters of a GP over a range of operating conditions using support states. The optimally-controlled dynamics $p(\hat{\mathbf{x}}|\mathbf{x},u^p)$ with $u^p = \pi^p(\mathbf{x})$ can be used as reference behavior to detect any deviation caused by a suboptimal control policy π^g. This latter, gives rise to a closed-loop dynamics $g(\hat{\mathbf{x}}|\mathbf{x},u^g)$ where action selection are defined by $u^g = \pi^g(\mathbf{x})$. We interpret the optimal controls $u^p = \pi^p(\mathbf{x})$ returned by the GPRL algorithm as noisy measurements of optimal control policy under uncertainty, i.e. a probabilistic map from states to actions.

Suppose that the optimally-controlled stochastic process GP^p, has been trained to describe the state transition dynamics through expectations on state changes when the optimal control policy π^p is followed. Given a state vector $\mathbf{x}$, a separate GP model is trained for each state dimension x in such a way the effect of uncertainty about its change due to a control action is modeled statistically as

$$\Delta x_k^p = (x_{k+1} - x_k) \sim GP^p(m^p, cov^p) \tag{4}$$

where m^p is the mean function and cov^p is the covariance function.

For policy iteration, GPRL describes the state-value function $V(\mathbf{x})$ and the control value function $Q(\mathbf{x},u)$ directly in function space by representing them using fully probabilistic GP models. Moreover, GPs provide information about confidence intervals for value function predictions and optimal actions. To generalize an optimal, continuous-valued control policy we have to model it based on a

finite number of evaluations in a support set $\mathfrak{x}$. GP models of the transition dynamics for feasible states are built on the fly using data gathered from interactions with a model of the system. The reinforcement learning problem consists in learning iteratively to achieve a goal, or to accomplish a control task, from ongoing interactions with a real or simulated system. During learning, the agent interacts with the target system by taking an action u_k causing the system to evolve from the state $\mathbf{x}_k$ to $\mathbf{x}_{k+1}$. Later the agent receives a numerical cost (given by the function ℓ^p) which provides a measure of how good (or bad) is the action taken at x_k in terms of the observed state transition. To this aim, the learning agent executes actions to minimize the cost of state transitions when starting from any initial state $\mathbf{x}_0$ and acting optimally thereafter. Under any given policy π, let's assume the expected cumulative cost $V^\pi(\mathbf{x}_0)$ or value function over a certain time interval is a function of $\mathbf{x}^\pi$. Here, $\mathbf{x}^\pi = \{\mathbf{x}_k\}_{k=1}^N$ are the corresponding states and $u^\pi = \{u_k\}_{k=1}^N$ defines the policy-specific sequence of control actions. The sequence of state transitions gives rise to costs $\{r_k\}_{k=1}^N$ which are used to define a discounted value function

$$V^\pi(\mathbf{x}_0) := \left[\gamma^N \cdot r_N + \sum_{k=1}^{N} \gamma^k \cdot r_k\right] \tag{5}$$

where $\gamma \in (0,1]$ is the discount factor which weights future rewards. Therefore, an optimal policy π^p over a rolling horizon N for the agent is one that minimizes Eq. (5) for any initial state $\mathbf{x}_0$. The associated state-value function satisfies the Bellman's equation in Eq. (1). The control value function Q^p is defined by

$$Q^p(\mathbf{x}_k, u_k) = \ell^p(\mathbf{x}_k, u_k) + \gamma \mathbf{E}[V_{k-1}(\hat{\mathbf{x}})|\mathbf{x}_k, u_k] \tag{6}$$

such that $V^p(\mathbf{x}_k) = \arg\min_u Q^p(\mathbf{x}_k, u_k)$ for all $\mathbf{x}_k$. Once Q^p is known through interactions, then the optimal control policy is obtained directly from

$$\pi^p(\mathbf{x}_k) = \arg\min_u Q^p(\mathbf{x}_k, u_k) \tag{7}$$

It is worth noting that by using the Q-function, the specification for optimal action selection is model-free. As a result, checking for trace equivalence assessment with the observed implementation can be made without the need of resorting to a generative model for which state transition probabilities should be known. For a Gaussian model GP^p describing the state transition dynamics $p(\hat{\mathbf{x}}|\mathbf{x}, u^p)$, a discrete and invariant set of possible actions $\mathfrak{u}$ and a finite set of next states $\mathfrak{x}$ are available. Policy iteration increasingly approximates the optimal controls $\pi^p(x)$ based on an arbitrary initialization of the value function $V^p(x)$. The first iteration is based on an initial policy $\pi^0(x)$ (e.g., a random policy) and the estimation of optimal state values $V^p_{s=1}(x)$ based on the cost function $\ell^p(\mathbf{x}, u)$. In the sth iteration, the optimal policy is re-estimated through $\pi_x(x)$ using an improved value function $V^p_s(x)$ with costs generated with the control policy $\pi^p_{s-1}(x)$ from the previous iteration. To generalize the policy iteration algorithm to unseen states a function approximation technique is required. GP are introduced to handle continuous state and action spaces using GP regression to describe both value functions and the control policy. For policy iteration, the GPRL algorithm describes the value functions $V(\mathbf{x})$ and $Q(\mathbf{x},u)$ directly in function space by representing them by fully probabilistic GP models. The advantage of modeling the state-value function in such a way, is that the GP provides a predictive distribution of function value for any state $\mathbf{x}_k$.

A *pseudocode* of GPRL using the transition dynamics GP^p and Bayesian active learning is given in Algorithm 1. GPRL starts from a small set of input locations $\mathfrak{x}_0$ and U_0 generated by applying a random policy π^0 to the model or system. Using Bayesian active learning (line 10) new set of states $\bar{\mathfrak{x}}$ and actions $\overline{U}$ are chosen (Deisenroth et al., 2009); both sets with cardinality ϑ, are added to the current set $\mathfrak{x}_k$ and U_k at stage k. Through this mechanism only a relevant part of the state space is explored. Parameters ρ and β are used as a tradeoff between exploration and exploitation in policy learning. To apply Bayesian active learning, first the set $\mathfrak{x}_k$ has to be obtained; this set accounts for the predicted means of the successor states when starting from $\mathfrak{x}_{k-1}$ and applying all actions from the finite and invariant set $\mathfrak{u}$ such that a set of candidate states for the support set are defined through $\bar{\mathfrak{x}}_k = \mathbf{E}[GP^p(\mathfrak{x}_{k-1}, \mathfrak{u})]$.

Each set $\mathfrak{x}_k$ and U_k obtained, provides training input locations for both the transition dynamics GP^p and the value function GP^V. At each iteration k of the Algorithm 1, the controlled dynamics GP^p is updated (line 12) in order to incorporate the most recent information from simulated state transitions. Furthermore, the GP models of the value functions Q^p and V^p are updated. After each control action $u_k \in \mathfrak{u}$ is taken, a cost function is called to reward the observed state transition.

Algorithm 1. *GPRL*

1: **Input**: $\mathfrak{x}^0, \mathfrak{u}, M, m, \rho, \beta, \vartheta, \omega$
2: $s = 1$
3: $\varepsilon = \infty$
4: **Until** $\varepsilon < \delta$ **do**
5: Simulate M trajectories applying $\pi^0 \to \mathfrak{x}^0, U^0$
6: Train GP^p around $\mathfrak{x}^0, U^0$
7: $V^0(\mathfrak{x}^0) = \ell^p(\mathfrak{x}^0, U^0)$
8: Train GP^p around $\mathfrak{x}^0, U^0$
9: **For** $k = 1$ **to** m
10: $\bar{\mathfrak{x}}, \overline{U} \leftarrow$ Bayesian active learning (ρ, β, ϑ)
11: $\mathfrak{x}_k := \{\mathfrak{x}_{k-1} \cup \bar{\mathfrak{x}}\}, U_k := \{U_{k-1} \cup \overline{U}\}$
12: Update GP^p around $\mathfrak{x}_k, U_k$
13: **For all** $\mathbf{x}_i \in \mathfrak{x}_k$ **do**
14: **For all** $u_j \in \mathfrak{u}$ **do**
15: $Q(\mathbf{x}_i, u_j) = \ell^p(\mathbf{x}_i, u_j) + \gamma E[V_{k-1}(\hat{\mathbf{x}})|\mathbf{x}_i, u_j, GP^p]$
16: **End for**
17: $Q^p(\mathbf{x}_i, u) \sim GP^Q$
18: $\pi^p(\mathbf{x}_i) \in \arg\min_u Q(\mathbf{x}_i, u)$
19: $V^p(x_i) = Q(x_i, \pi^p(x_i))$
20: End for
21: $V^p(\mathbf{x}) \sim GP^V$
22: **End for**
23: $\mathfrak{x} = \mathfrak{x}_m$
24: Approximate $\pi^p \sim GP^{\pi^p}$ with and $\pi^p(\mathfrak{x})$
25: **If** $s \geqslant 2$ **then** verify $\varepsilon \leqslant \omega$
26: **End If**
27: $s = s + 1$
28: **End loop**
29: **Return** $\mathfrak{x}, U = \pi^p(\mathfrak{x}), GP^p$

A GP model is used to approximate the optimal policy as $\pi^p \sim GP^{\pi^p}(m^{\pi^p}, cov^{\pi^p})$ in line 24, consequently optimal action selection for state $\mathbf{x}_k$ corresponds to $u^p_k = \pi^p(\mathbf{x}_k) := m^{\pi^p}$. Since control policies in successive stages are also modeled using GPs, policy iteration can be stopped when the sum of the Kullback–Leibler distance between two successive policy distributions over the support set is lower than a small tolerance ω given by

$$\varepsilon = \frac{\mathrm{Div}(\pi_{s-1}) - \mathrm{Div}(\pi_s)}{\mathrm{Div}(\pi_s)} \leqslant \omega \tag{8}$$

where previous and current values are defined as

$$\begin{aligned}\mathrm{Div}(\pi_{s-1}) &= \sum_{\mathbf{x}\in x_s} KL\left(GP^{\pi^p_{s-2}} \| GP^{\pi^p_{s-1}}\right) \Big/ \|x_{s-1}\| \\ \mathrm{Div}(\pi_s) &= \sum_{\mathbf{x}\in x_s} KL\left(GP^{\pi^p_{s-1}} \| GP^{\pi^p_{s}}\right) \Big/ \|x_{s}\|\end{aligned} \tag{9}$$

2.3. Case study: artificial pancreas

Insulin-Dependent Diabetes Mellitus (IDDM) is a chronic disease characterized by the inability of the pancreas to produce the hormone insulin, which leads to high blood glucose (BG) levels. Poorly controlled diabetes mellitus is known to cause serious health problems, including heart disease and stroke, hypertension, retinopathy, nephropathy, and neuropathy (Franco, Steyerberg, Hu, Mackenbach, & Nusselder, 2007; Gerich, 2005). Also, abnormal glycemic variability contributes to oxidative stress, which has been linked to the pathogenesis of diabetes (Marling, Shubrook, Vernier, Wiley, & Schwartz, 2011; Siegelaar, Holleman, Hoekstra, & DeVries, 2010). Compensating for this deficiency in endogenous insulin production requires 4–6 insulin injections to be taken daily; the aim of this diabetes therapy is to maintain normoglycemia – i.e., a blood glucose level between 3 and 10 mmol/l. In defining the amount and timing of these injections, poor predictability of BG dynamics is a key issue that both patients and doctors must deal with (Bremer & Gough, 1999). Manual control of BG often results in high glycemic variability and the risk of a life-threatening hypoglycemic event is at stake. Hypoglycemia – i.e., low blood glucose levels – may lead to brain damage, coma and eventually death.

Glycemic variability can be simulated using a stochastic processes superimposed on an otherwise deterministic model of the glucose–insulin dynamics. To this aim, The Lehmann and Deutsch model (Lehmann & Deutsch, 1992) parameterized as described in Acikgoz and Diwekar (2010) is used as the basis to describe the deterministic glucose-insulin dynamics. The glucoregulatory model considers two subsystems (compartments) whose interactions describe the glucose–insulin dynamics. There exists not only uncertainty in the estimation of model parameters -that prevents describing variability among daily values of glucose in patients-, but also other sources of structural errors that give rise to model-patient mismatch. Also, there are measurement errors in glucose sensors which prevent uniquely determining the true glycemic state in a diabetic patient. Thus, there is uncertainty in both estimating system states and predicting the outcome of control actions. All these sources of time-dependent uncertainties could be represented by adding a stochastic process to a deterministic model of the glucose dynamics. Considering the simplest generalization of an Ito's stochastic process given by

$$dx = \lambda\, dt + \sigma dz \quad \text{with} \quad dz = \varepsilon_t \sqrt{dt} \tag{10}$$

where λ is called the drift parameter, σ the variance parameter and ε a random number sampled from a normal distribution $\varepsilon \sim \mathcal{N}(0,1)$. The glycemic variability in a cohort of diabetic patients is modeled as follows

$$\frac{dBG}{dt} = \frac{G_{in} + NHGB - G_{out} - G_{ren}}{V_G} + \frac{\sigma\varepsilon}{\sqrt{dt}} \tag{11}$$

where BG represents the plasma glucose concentration; G_{in} is the systemic appearance of glucose via glucose absorption from the gut; $NHGB$ is net hepatic glucose balance; G_{out} is the overall rate of peripheral and insulin dependent glucose utilization; G_{ren} is the renal excretion of glucose and V_G is the volume of distribution of glucose. G_{out} assumes a relationship between glucose utilization and plasma glucose concentration because of two patient dependent parameters, namely hepatic sensitivity Sh and insulin sensitivity Sp. The full-model is detailed in Appendix A. It is worth noting that most glucose–insulin models are based on "average patients". So, these models are only able to simulate the glucose dynamics in a standard population but intra-individual variability is not described. Introducing a stochastic process ensures a cohort of *in silico* subjects that accounts sufficiently well for the observed patient-to-patient variability. Simulating inter-subject variability is indeed crucial in the design of optimal controllers under uncertainty which can be reliably used as the specification for behavior monitoring.

Let's consider a glucose 24-h profile for an *in silico* Type 1 diabetic patient, with a sampling time of 6 min. The state variables used to define the state of the system at time k are the measured BG concentration at k and the insulin flow rate in the previous step $I_{k\text{-}1}$. Thus, a perception of the system state is defined as $\mathbf{x}_k = (BG_k, I_{k\text{-}1})$. The control action u_k is the change in the insulin infusion rate. The proper addition of an Ito's stochastic process to a deterministic model using the variance parameter σ allows simulating a cohort of patients with a wide range of physiological characteristics. In Table 1 the meal routine used is detailed in terms of the carbohydrate intake profile.

As we consider behavior monitoring of an AP system, we should focus on sensor readings rather than actual blood or plasma glucose. As continuous glucose monitor systems (CGMS) are placed in the subcutaneous tissue, they determine interstitial fluid (IG) rather than blood or plasma concentration. Hence, under dynamic conditions, IG and BG values are necessarily different because of a time-lag between both concentrations. The magnitude of the time lag may be no more than 5 min in normal conditions (G. M. Steil et al., 2005). In the simulation model, each IG value is obtained through integrating a BG-IG model, where ρ is the static gain of the IG dynamics (considered equal to 1) and τ is the time-lag constant

$$dIG(k) = -\frac{1}{\tau} IG(k) + \frac{\rho}{\tau} BG(k) \tag{12}$$

In order to simulate a noisy CGMS time series, the IG profile is multiplied by a random time-varying calibration error $\xi(k)$ and then corrupted by an additive noise sequence sampled from a zero mean white Gaussian noise process $v(k)$ giving

$$\begin{aligned} CGM(k) &= (1 + \xi(k))IG(k) + v(k) \\ \xi(k+1) &= 3\xi(k) - 3\xi(k-1) + \xi(k-2) + w(k)\end{aligned} \tag{13}$$

Clearly, setting $\xi = 0$, corresponds to an optimally calibrated CGMS. The calibration error $\xi(k)$ has been created using a triple integrator of a zero mean white noise $w(k)$. We talk about BG instead of IG, but recall it refers hereafter to glucose levels determined through the continuous monitor. By default, we use a time-lag $\tau = 5$ min and a calibration error $\xi = 2\%$.

In Fig. 1, the key concept of using an optimal policy for controlling glycemic variability in the face of uncertainty is shown. The optimal policy obtained using the GPRL algorithm was trained using a simulation model with variability $\sigma = 0.10$ and insulin sensitivity $Sp = 0.5$. The optimal policy is evaluated for controlling simulated patients exhibiting a larger level of variability than the one used in the training phase. The inter- and intra-variability in diabetic patients are modeled by significantly increasing the parameter σ while 125 independent simulations of the controlled Ito's process were made to model the glucose–insulin dynamics. Observe that despite the effects of control actions are quite

Table 1
Carbohydrate intake schedule.

Carbohydrate content (g)	47	16	63	31	63	31
Meal times (h)	8.00	10.00	12.30	16.00	19.30	22.00

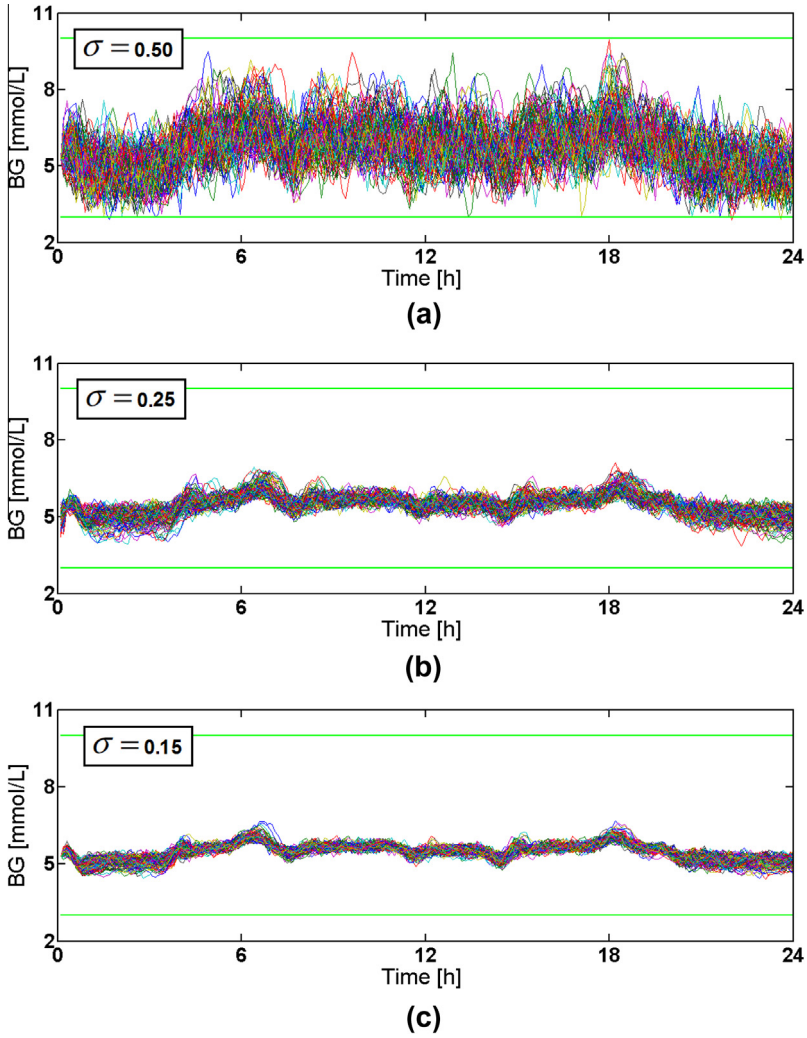

Fig. 1. Glycemic variability using stochastic optimal control. Sample paths are generated using different values of the Ito's parameter σ.

uncertain, the optimal policy is able to achieve effective glycemic control by maintaining levels within the normoglycemic range.

3. Bayesian surprise

3.1. KL-divergence between priors and posteriors

A quantitative definition of the relevance a new observation has about the state or condition of a system being monitored is required to measure the information available in new data based on the current system knowledge. In information theory, information measures the uncertainty or probability of occurrence of an outcome (Cover & Thomas, 2012). This classical information measure is based solely upon objective probabilities and cannot discriminate between outcomes which may be relevant or not for system monitoring (Hasanbelliu et al., 2012). Surprise, no matter how one defines it, is obviously related to the Shannon's information concept. Any rare event carries a great deal of Shannon information since it is not expected, but this condition does not make it a surprising event (Weaver, 1966). Surprise measures the information that is contained in data in an observer-dependent way and related to his changes in expectation (Baldi, 2002). The *a priori* probabilities considered are essentially subjective probabilities about the expected system behavior which serves as a reference or benchmark (Jaynes, 1988). Thus, the main function of a feeling of surprise is to make us reconsider the validity of our previous assumptions (Good, 1988). We tend to be surprised when new data reinforces an alternative hypothesis (Bayarri & Berger, 1997), i.e. when the result of an observation has much greater probability *a posteriori* under an alternative hypothesis to the one used *a priori* (Good, 1956; Weaver, 1966).

It has been argued that a suitable measure of information ought to remain probabilistic in nature but to depend on the observer (monitor) and its prior beliefs or expectations (Itti & Baldi, 2005a). The fundamental impact data has on a monitor is thus captured by Bayes' theorem for computing the posterior probability of a hypothesis or model M given the data

$$P(M|D) = \frac{P(D|M)}{P(D)} P(M) \tag{14}$$

Because surprise exists only in the presence of uncertainty its essence must be probabilistic in such a way distributions are used to capture monitor's beliefs over the current space of plausible hypotheses or models M (Itti & Baldi, 2005b). In this view, the information contained in a data stream is not its entropy but rather what changes the monitor's belief in M from its prior hypothesis to its posterior. It is worth noting that any measure of surprise lacks any sense without such hypothesis since no expectations can be made. In this manner, a new observation D carries no surprise if it leaves beliefs unaffected, that is, if the posterior is almost

identical to its prior; conversely, D is surprising if the posterior distribution resulting from observing D significantly differs from the prior distribution, namely the observer expectations are not fulfilled. Surprise is then measured using the distance between the posterior and prior distributions, based on the Kullback–Leibler divergence

$$S(D,M) = KL(P(M)||P(M|D)) = \int_M P(M)\log\frac{P(M)}{P(M|D)}dM \quad (15)$$

According to the frequency definition of probability a low-probability event is most informative. Conversely, for the surprise definition this is not always true (Weaver, 1966). Consider for instance the case where we have three competing internal models or hypotheses about new data $d_i = \{d_1, d_2, d_3...\}$ from a diabetic patient acquired by a sensor, as depicted in Fig. 2. The first model 'Hypo' accounts for BG levels less than 3 mmol/L, the second 'Normal' according to BG between 3 and 10 mmol/L and the last 'Hyper' according to levels higher than 10 mmol/L. Notice that, between 3 and 10 mmol/L, there exist as many BG values as allowed by the sensor resolution. Thus, the probability of a given BG value occurring in this range is very small. However, all these different values are equally likely to occur; and one of them is absolutely certain to occur every time a sensor reading is available. But, although any of the BG readings in the Normal range is a low-probability event, hence carries a lot of information, any such readings cannot be aptly named a surprising event. There is no basis for being surprising since its probability of occurrence is the same for any other particular value in the normality range (Weaver, 1966). It all depends whether the monitor had previously set up some hypothesis about what to be expected next. To continue, consider the monitor have its own hypothesis about the range in which the incoming glycemic values may fall. Assume that the initial agent's prior belief (or probability) for incoming readings falling in the Normal range is quite larger than the other two hypothesis (left side). If certainly BG readings d_i fall between 3 and 10 mmol/L, we have a reinforcement signal of agent's prior hypothesis that Normal is a highly probable hypothesis. In such a case the glucose data stream does not promote a change of the monitor's prior beliefs, hence the set d_i is not surprising. In contrast, imagine the situation in which d_i falls in the Hyper range. Now, observing the incoming data the agent's will update its beliefs favoring now the hypothesis Hyper, and therefore it can be said that d_i carries now a high level of surprise.

While the definition of surprise above may seem overly simplistic, note that the integral in (15) is over the model space M, i.e., takes into account all possible models for the measurement D. This definition, in its current form, is too conceptual to be useful for computational purposes. In practice, we translate the model space into a parametric distribution family captured as a mixture of Gaussians distributions (Hasanbelliu et al., 2012; Ranganathan & Dellaert, 2009). To estimate the prior and posterior distributions, a non-parametric distribution family captured as Gaussian processes for the state transition dynamics for both $p(\hat{\mathbf{x}}|\mathbf{x}, u^p)$ and $g(\hat{\mathbf{x}}|\mathbf{x}, u^g)$ are used. The Model GP^p describes the prior beliefs a monitor has about the expected agent behavior, namely an optimally-controlled dynamics with well-performing sensor and actuator devices. Note that as the arriving data is changing the hyper-parameters of the GP^g used to model the agent implementation some short of Bayesian updating is taking place. On the other hand, GP^g describes the observed dynamics of the controlled system, governed by a suboptimal policy and possibly underperforming due to ill-functioning of sensors and actuators mechanisms. The amount of surprise in the data stream can be measured by looking at the changes that take place in going from GP^p to GP^g. The only restriction in the general theory of Bayesian surprise, is that the prior and the posterior must have the same functional form (Baldi & Itti, 2010). An agent whose GP^g reflects that is operating far away from its specified behavior, will produce larger amounts of surprise compared to a situation where GP^g is rather close to GP^p. The Kullback–Leibler divergence between both Gaussian processes emphasizes the fact that similar states should produce similar estimates on both covariates and responses. As a result, Bayesian surprise describes how distant a suboptimal behavior is from its specification.

As the transition probabilities $p(\hat{\mathbf{x}}|\mathbf{x}, u^p)$ and $g(\hat{\mathbf{x}}|\mathbf{x}, u^g)$ that describe the changes for both system dynamics are modeled as GPs as follows

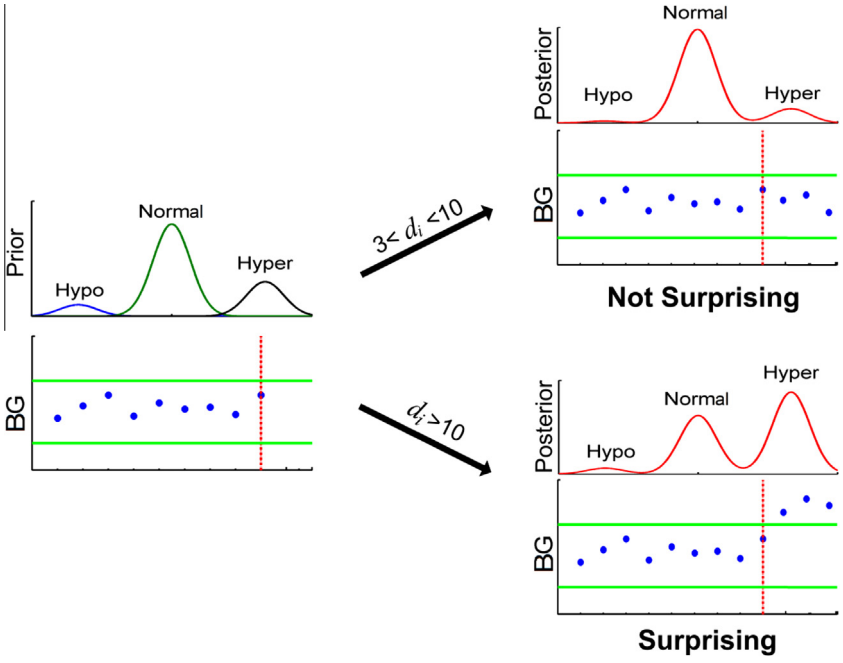

Fig. 2. Simple description of how surprise may be computed at a high level of abstraction for an observer who has prior beliefs about incoming glycemic data in an artificial pancreas system.

$$\Delta x^g \sim GP^g(m^g, cov^g)$$
$$\Delta x^p \sim GP^p(m^p, cov^p) \quad (16)$$

these GPs are then used to infer states changes due to each policy which are obtained separately as in Eq. (4) for each dimension x of the vector state $\mathbf{x}$. An advantage of modeling probability distributions for state changes as GPs is that Bayesian surprise can be readily computed based on the divergence between Gaussian distributions using the *KL* distance

$$S(g\|p) = KL(\Delta x^g \| \Delta x^p) \quad (17)$$

where Δx^g and Δx^p can be used to compute stepwise surprise. Alternatively, vectors containing the last W estimations $\{\Delta x_k^g\}_{k-W}^k$ and $\{\Delta x_k^p\}_{k-W}^k$ can be used to obtain a more robust metric of Bayesian surprise.

3.2. Stepwise Bayesian surprise

In order to compute the distance between the agent behavior and the observed state transition dynamics, we should compute general terms of the form

$$F(GP^p, GP^g) = \int GP^p \log(GP^g)\, d\Delta x \quad (18)$$

being $GP^p(m^p, cov^p)$ and $GP^g(m^g, cov^g)$ two Gaussian models with the same functional form. The stepwise Bayesian surprise P_{KL} is then given by

$$P_{KL}(GP^g \| GP^p) = F(GP^p, GP^p) - F(GP^p, GP^g) \quad (19)$$

where GP^p is the reference and GP^g is the current density of interest. After some algebra, the first term is given by

$$F(GP^p, GP^p) = \int_{-\infty}^{+\infty} GP^p \log(GP^p) d\Delta x = \frac{1}{2}\log(2\pi e(cov^p)^2) \quad (20)$$

and the cross term is given by

$$F(GP^p, GP^g) = \int_{-\infty}^{+\infty} GP^p log(GP^g) d\Delta x = -\frac{1}{2}\log\left(2\pi(cov^g)^2\right) - \frac{(cov^p)^2 + (m^p - m^g)^2}{2(cov^g)^2} \quad (21)$$

hence, surprise is obtained as

$$P_{KL}(GP^g \| GP^p) = -\frac{1}{2} + \log\left(\frac{cov^g}{cov^p}\right) + \frac{(cov^p)^2 + (m^p - m^g)^2}{2(cov^g)^2} \quad (22)$$

It is worth noting that in P_{KL} mean and covariance differences are computed from current prediction of state changes without considering previous evaluations which gives rise to a stepwise procedure. If both variance parameters cov^p and cov^g in Eq. (22) remain mainly constant when computing $P_{KL}(GP^g||GP^p)$, then $\log(cov^g/cov^p)$ becomes a steady-state value for the Bayesian surprise metric reflecting the intrinsic variability of the agent policy with respect to the reference distribution. The last term in Eq. (22) is proportional to the square difference between both means m^p and m^g. This difference gives rise to most of the variability in the surprise measure. Thus, each change in the stepwise surprise value is mostly proportional to the deviation of the observed mean m^g from the reference mean m^p.

3.3. Artificial pancreas (cont'd)

The specified behavior represents the optimal conditions the closed-loop is expected to work in. This represents the prior hypothesis of an optimal control policy acting over a regularly working artificial pancreas (sensor, pump, control algorithm and glucose metabolism). This close-loop dynamics is represented by GP^p in Eq. (22). Otherwise, GP^g describes the current behavior of the glycemic control loop, considering the effects of measurement errors, infusion pump malfunctioning, algorithm ill-tuning and glycemic variability. Stepwise surprise may detect deviations of glucose control from the expected optimal policy under uncertainty. The effect of a deviation from the specified behavior is revealed by P_{KL} as it is shown in Fig. 3. Here, surprise is computed step-by-step using observed mean and covariances obtained through Gaussian processes describing both close-loop dynamics GP^g and GP^p.

The suboptimal dynamics GP^g is accomplished by a PID and a PID-Fuzzy schemes. PID algorithm in Eq. (23) for the glucose–insulin model is fully described in Farmer, Edgar, and Peppas (2009), where u_b is the basal insulin, k_C is the proportional action, τ_I is the integral time and τ_D is the derivative time. To maintain basal conditions, setpoint is set to be the basal glucose concentration $G_r = 5.5$ (mmol/L).

$$I_{PID}(t) = u_b + k_C\left[(G(t) - G_r) + \frac{1}{\tau_I}\int_0^t (G(t) - G_r)dt + \tau_D\frac{d(G(t) - G_r)}{dt}\right] \quad (23)$$

Parameters for the PID algorithm in Eq. (23) are $u_b = 16.667$ mU/min, $k_C = 12$ (mU/min)(mmol/L)$^{-1}$, $\tau_I = 3300$ min and $\tau_D = 40$ min.

The design of the expert Fuzzy-PID controller is based on the work of Susanto-Lee, Fernando, and Sreeram (2008). For these controllers, a sliding scale method is used which determines insulin infusion rates based on a lookup scale containing a set of BGL ranges. The fuzzy controller is based on a Mamdani-type fuzzy inference system, with one input variable (BG level) and one output variable (Insulin dose) resulting from the sum of the three P, I & D actions. The defuzzification for the output is calculated using the centroid method, which essentially determines the center of mass of the set of fuzzy outputs. For the proportional action, the scale has a continuous blood glucose level partitioned into zones with a linear increment of insulin rate according to the BG level. If the current sliding scale does not provide enough insulin to lower BG level, an integral controller provides the increment of insulin. This amount depends on a normalized weighted average of the BG levels, calculated over the past 2 h to give more weight to recent glycemic values. The derivative action is performed through a bolus which provides a mechanism to boost the insulin delivery during a rapid increase of BGL. A least square regression technique is used to calculate a projected insulin level, over a 30-min window that contains the five most recent BG readings. If IG falls below the 3 mmol/L, a protection scheme shut-off insulin injection and terminate any active bolus. Expert control is implemented by a set of rules which are triggered when BG crossed a few predefined boundaries. These rules effectively are protection schemes to prevent hypoglycemia.

Both dynamics are then characterized using a sample of 30 state-control pairs and the corresponding state transitions. Changes in the controlled dynamics GP^g are assessed for three increasing degrees of variability $\sigma = \{0.15; 0.25; 0.50\}$ respect to the optimal specification GP^p which uses $\sigma = 0.10$. At the 12th hour, the parameter σ is significantly increased to simulate a sudden rise in glycemic variability. As observed, stepwise Bayesian surprise P_{KL} is sensitive to any deviation from the expected behavior of the AP. Note that as glycemic variability is increased, so does the Bayesian surprise which vividly reflects the level of degradation in the AP performance. Even though the KL-divergence should be equal to 0 for two identical distributions, the surprise measure P_{KL} is not strictly equal to 0 over the interval [0,12] hours as the control system experiments an inherent variability. Changes in the surprise value are proportional to the deviations of the estimated mean

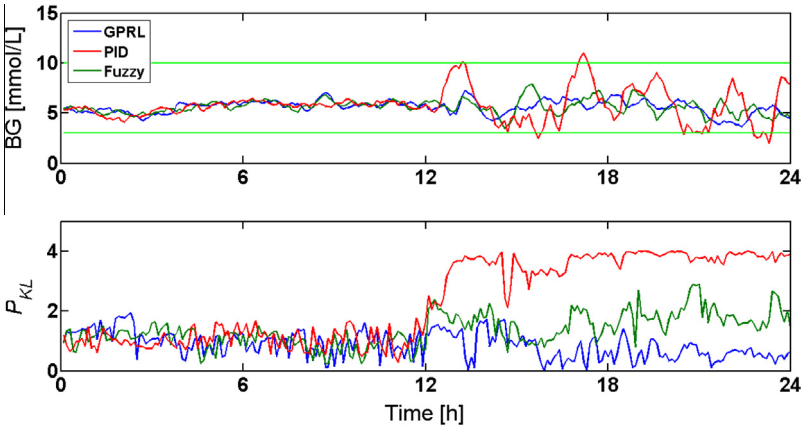

Fig. 3. Stepwise Bayesian Surprise P_{KL} for different values of the glycemic variability parameter σ.

m^g from the mean m^p corresponding to the optimally-controlled AP. It is worth noting that, as surprise is proportional to the immediate difference between successive estimations, it presents high fluctuations when computed step-by-step; rendering it less informative for pinpointing a deviant behavior. Clearly, Fig. 3 shows in the left side how changes in the control policy, that affect the state transition distribution, are revealed by Bayesian surprise. This way surprise allows us to detect deviations of the implemented dynamics from optimal behavior.

3.4. Robust Bayesian surprise

Despite step-by-step estimation of Bayesian surprise may help detecting deviations from the specified close-loop dynamics, it clearly does not characterize a progressive degradation in the behavior of a controller. As P_{KL} is computed stepwise its value may result too noisy due to the inherent variability in the controlled dynamics. To overcome this drawback, a robust Bayesian surprise metric is proposed using the approach of Bo and Sminchisescu (2010) for structured prediction based in twin Gaussian processes.

A Gaussian process is a collection of random variables, any finite number of which has a joint Gaussian distribution. For Gaussian process regression (GPR), those random variables represent the value of the function $f(\mathbf{x})$ for inputs $\mathbf{x}$. GPR assumes $f(\mathbf{x})$ is a zero mean stationary Gaussian process with covariance function $\mathrm{k}(\mathbf{x}_i, \mathbf{x}_j)$, encoding correlations between pairs of random variables

$$\mathrm{cov}(f(\mathbf{x}_i, \mathbf{x}_j)) = \mathrm{k}(\mathbf{x}_i, \mathbf{x}_j) \tag{24}$$

One covariance function particularly used is the Gaussian

$$\mathrm{k}(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma_r \|\mathbf{x}_i - \mathbf{x}_j\|^2) + \lambda_r \delta_{ij} \tag{25}$$

with $\gamma_r \geqslant 0$ the kernel width parameter, $\lambda_r \geqslant 0$ the noise variance and δ_{ij} the Kronecker delta function, which is 1 if $i = j$, and 0 otherwise. This prior for the kernel function constrains input samples that are nearby to have highly correlated outputs.

Suppose a particular set containing a sequence of the last W state differences $\boldsymbol{X}^g = \{\Delta x_k^g\}_{k-W}^{k}$ for the stochastic process that results from applying an arbitrary control policy π^g. Then, the joint distribution of observed state differences can be modeled using a zero mean multivariate Gaussian distribution as

$$(\boldsymbol{X}^g)^T \sim \mathcal{N}^g\left(0, \begin{bmatrix} \boldsymbol{R}_{ij} & r_i \\ r_i^T & r \end{bmatrix}\right) \tag{26}$$

whose covariance $\mathbf{K}^g$ is given by the kernel matrix $\boldsymbol{R}_{ij} = \mathrm{k}(\Delta x_i^g, \Delta x_j^g)$, the kernel vector $r_i = \mathrm{k}(\Delta x_i^g, \Delta x_k^g)$ and the kernel value $r = \mathrm{k}(\Delta x_k^g, \Delta x_k^g)$.

For the given Gaussian dynamics $\mathcal{N}^g(0, \boldsymbol{K}^g)$ of a sampled sequence of state transitions, the offset or distance to a specified distribution $\mathcal{N}^p(0, \boldsymbol{K}^p)$ is key to calculate a robust measure of Bayesian surprise T_{KL}. This is achieved by computing the Kullback–Leibler divergence between Gaussian processes as

$$T_{KL}(\mathcal{N}^g \| \mathcal{N}^p) = -\frac{N}{2} - \frac{1}{2}\log|\boldsymbol{K}^g| + \frac{1}{2}\mathrm{Tr}\left\{\boldsymbol{K}^g(\boldsymbol{K}^p)^{-1}\right\} + \frac{1}{2}\log|\boldsymbol{K}^p| \tag{27}$$

The Kullback–Leibler divergence is therefore non-negative and zero if and only if the two multivariate Gaussian distributions have the same covariance. In the latter case, the sequence of state transitions caused by an agent policy has no surprise regarding its specification.

Gaussian process regression is an efficient method to model non-linear input–output dependencies, but it only focuses on the prediction of a single output at a time. Although generalizations to multiple outputs can be derived by training independent models for each output, this procedure fails to leverage information about correlations among output components in the predictor. A major advantage of using twin Gaussian processes to compute Bayesian surprise is that the proposed approach may be easily extended to MIMO systems.

3.5. Artificial pancreas (cont'd)

Robust Bayesian surprise T_{KL} is computed using a moving window strategy employing a sequence containing the last W estimations of state transitions when the optimal control policy is applied. Simulated state transition are assumed *i.i.d.* multivariate normal distributions as stated in the left side of Eq. (26). In Fig. 4, the window size varies from $W = \{10; 20; 30\}$ points. Note that as the size of the window is increased the corresponding surprise curve becomes smoother, i.e., less sensitive. Certainly, the sample size W results from a tradeoff between the speed of detection of any event or disturbance causing a deviant behavior and the proper characterization of a progressive shift in the agent behavior. As can be seen in Fig. 3, for small samples the surprise metric T_{KL} tends to be similar to the robust surprise metric P_{KL}. Theoretically, both metrics should be quite similar when $W \to 1$.

In Fig. 5, T_{KL} is computed for the evaluated dynamics GP^g with respect to the specification GP^p. Window size has been limited to $W = 20$ points. At 12th h, the Ito's parameter σ is increased from 0.1 to 0.15, 0.25 and 0.50, to simulate three different degrees of variability. It is quite clear that an increase of parameter σ certainly change the loop dynamics (see the right half of Fig. 3). As it can be observed, the amount of surprise in data from the AP is

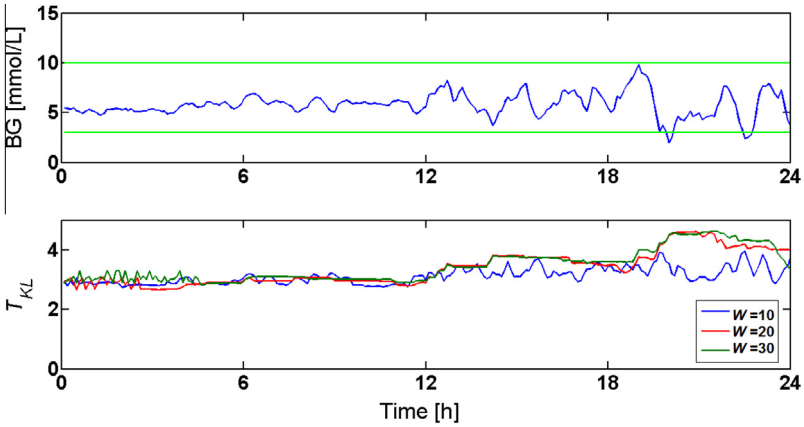

Fig. 4. Robust Bayesian Surprise T_{KL} computed using different window sizes.

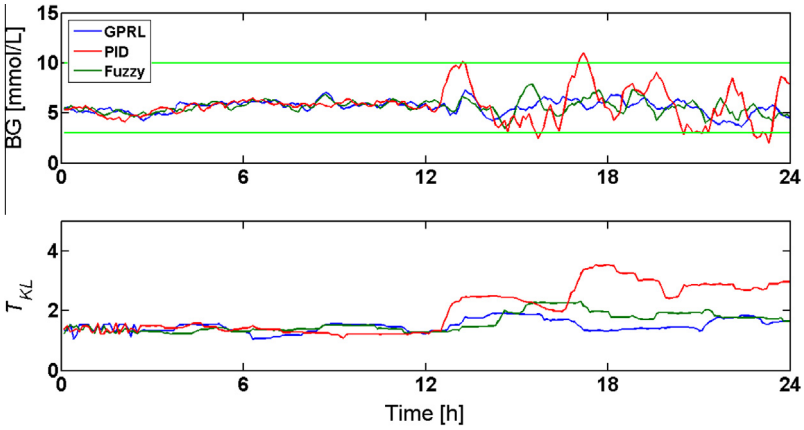

Fig. 5. Robust Bayesian Surprise T_{KL} for different values of the glycemic variability parameter σ.

seen in T_{KL} now as a rather smooth curve, unlike P_{KL} in Fig. 3. Glycemic variability in the simulation model gives rise to values for T_{KL} that are not strictly equal to 0 even when the same level of variability is used in the reference behavior (see the left side of Fig. 3). However, Bayesian surprise exhibits a nearly constant value before the parameter σ was shifted. Despite an increase of variability in glucose dynamics, even for a small change in the Ito's parameter σ, e.g. from 0.10 to 0.15, a clear increase in the level of surprise can be observed.

3.6. Incremental on-line sparsification

As arriving data may indicate changes in an agent behavior, models used in monitoring systems are required to be updated on-line to accommodate new pieces of information which would potentially enhance the available knowledge for the monitor and improve its performance. However, if this is not handled with care, models can become corrupted and hinder a realistic computation of surprise with a relevant information content; this is the so called stability-plasticity dilemma (Hasanbelliu et al., 2012). Hence, it is necessary to assess the information available in a new sample and its effect on the capability of the monitoring system to detect deviations from the optimal behavior.

Most of the data the monitoring system will observe will have little or no new information, i.e., it is not surprising. Allowing no informative samples to update the GP model is not only wasteful and time-consuming but also will over-train it making the model too rigid to adapt to new, interesting samples. A good sample instead, would provide information beneficial to the detection of a sub-optimal policy. Also, it would carry reasonable amount of new information that would sufficiently but not excessively change the state transition model. An outlier, on the other hand, would provide information detrimental to the monitoring task. The monitoring system would prevent outliers from incorrectly modifying the GP model but allow interesting observations to improve the accuracy of model predictions. The key question in behavior monitoring is preventing outliers and corrupted data to affect the reference behavior for the controlled loop dynamics.

From the preceding discussion, it results evident the importance of describing any change in the loop dynamics using a reduced and relevant set of updated information. It is therefore necessary to establish criteria in order to insert or replace data points in the training set. Following to the work of Engel, Mannor, and Meir (2004) and Nguyen-Tuong and Peters (2011), a method for incremental on-line sparsification is used here. Sparsification can be integrated with many existing on-line regression methods to enable fast real-time model learning.

The suggested sparsification method is implemented here by means of incremental kernel regression, which uses a test of linear independence to select a sparse subset of the training data points often called the *dictionary* (Nguyen-Tuong & Peters, 2011). At all

times, the algorithm maintains a dictionary $\mathfrak{D} = \{\mathbf{d}_i\}_{i=1}^{m}$, where m denotes the current number of dictionary points d_i. To test whether a new point d_{m+1} should be inserted into the dictionary, it is necessary to ensure that it cannot be approximated already in the feature space spanned by the current dictionary. This test can be performed using a measure δ defined as

$$\delta = \left\| \sum_{i=1}^{m} a_i \phi(\mathbf{d}_i) - \phi(\mathbf{d}_{m+1}) \right\|^2 \tag{28}$$

where a_i denote the coefficients of linear dependence and $\phi(\mathbf{d})$ is a feature vector which maps $\mathbf{d}$ into a higher dimensional space. Parameter δ can be understood as the distance of a new point $\mathbf{d}_{m+1}$ to the linear plane spanned by the dictionary set $\mathfrak{D}$ in the feature space. Thus, the value of δ is considered an independence measure, indicating how well a new data point $\mathbf{d}_{m+1}$ can be approximated in the feature space by other points in the data set. The larger the value of δ, the more independent $\mathbf{d}_{m+1}$ from the support dictionary set $\mathfrak{D}$, hence more informative for computing Bayesian surprise.

The coefficients a_i can be determined by minimizing δ using a mathematical program formulated in matrix form as the following optimization problem (Nguyen-Tuong & Peters, 2011).

$$a = \min_{\mathbf{a}} [\mathbf{a}^T \mathbf{Q}_{ij} \mathbf{a} - 2\mathbf{a}^T \boldsymbol{q}_i + q] \tag{29}$$

where $Q_{ij} = \mathrm{k}(\mathbf{d}_i, \mathbf{d}_j)$ represents a kernel matrix, $q_i = \mathrm{k}(\mathbf{d}_i, \mathbf{d}_{m+1})$ is a kernel vector and $q = \mathrm{k}(\mathbf{d}_{m+1}, \mathbf{d}_{m+1})$ denotes a kernel value. The employed kernel here is the Gaussian kernel introduced by Eq. (25).

In addition to the spatial allocation of the dictionary state space through δ, temporal allocation is taken into account by introducing a time-variant forgetting factor for every dictionary point. This parameter accounts for the time when each dictionary point has been inserted into a sparse set. For deleting a point from the dictionary, the deletion score is the independence value δ weighted by the corresponding forgetting value. The deletion score results in a trade-off between temporal relevance and novelty. The sparsification method is given in **Algorithm 2** below.

Algorithm 2. *Sparsification*

1: **Input**: new point $\mathbf{d}_{m+1}$, threshold η, $N_{\max}$
2: Compute $\boldsymbol{a} = Q_{ij}^{-1} q_i$
3: Compute $\delta = q - q_i^T \boldsymbol{a}$
4: **If** $\delta > \eta$ **then**
5: **If** number of dictionary points $m < N_{\max}$ **then**
6: Insert $\mathbf{d}_{m+1}$ into $\mathfrak{D}$ and update the dictionary
7: **Else**
8: Insert $\mathbf{d}_{m+1}$ into $\mathfrak{D}$ and update the dictionary by replacing the less relevant point from $\mathfrak{D}$
9: **End if**
10: $\mathfrak{D} = \mathfrak{D} \cup \mathbf{d}_{m+1}$
11: **End if**

Sparsification should protect relevant information from being replaced using new, yet trivial data. Parameters δ and $N_{\max}$ will define how fast the GP model adapts on-line to new environments as well as when learned information is not useful anymore. If the independence value δ is small the more likely the corresponding dictionary point is going to be replaced. The idea is to delete points that are more dependent on other dictionary points, i.e., where the corresponding independence value δ is small. In line 4, δ is used as a criterion for selecting new data points through the threshold parameter η that implicitly controls the level of sparsity.

Parameter $N_{\max}$ defines the size of the training set used to model the Gaussian process dynamics. This is a decisive factor for a monitoring task because it will define how fast the GP for state transitions model adapts to new environments or scenarios. As this parameter is reduced, each piece of information becomes more significant in the training set and more informative for model adaptation to a novel environment. For systems that change fast in time it is important to be able to quickly adapt to these changes, hence it would be better to use small training sets.

3.7. Artificial pancreas (cont'd)

As the glucose–insulin dynamics is parsimonious due to homeostasis, a dictionary containing $N_{\max} = 30$ points with a sampling time of 6 min (3 h of data), provides enough data to be useful for behavior monitoring of the AP. The effect of increasing the parameter η, describing the selection of states to be incorporated in the dictionary, can be observed in Fig. 6. Two variables that made up the perception of the patient state $\mathbf{x}_k = (G_k, I_{k-1})$, are plotted against each other to highlight the boundaries of the feature space. From an initial training set with $l = 200$ states (blue points), only the $N_{\max} = 30$ most relevant states are labeled (red points), for $\eta = 0.1$ and $\eta = 0.9$. Larger values of η give rise to enlarging the area where the preferred states are located. This is because only those points with a high value for the independence parameter δ are eligible to be included into the dictionary.

In Fig. 7, a scenario in which the control fails to maintain normoglycemia when the variability parameter is increased from $\sigma = 0.1$ to $\sigma = 0.5$ at the 12th hour is shown. Surprise is computed

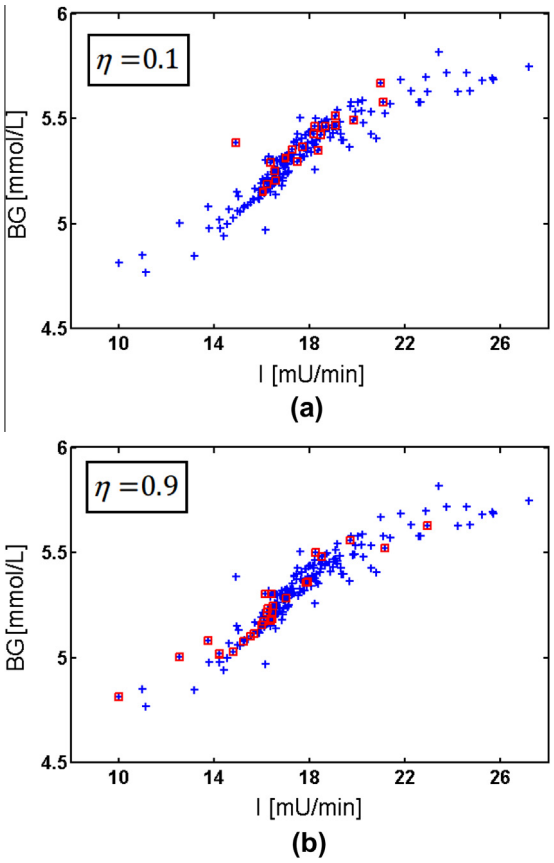

Fig. 6. Dictionary points selection through the threshold parameter η that implicitly controls the level of sparsity.

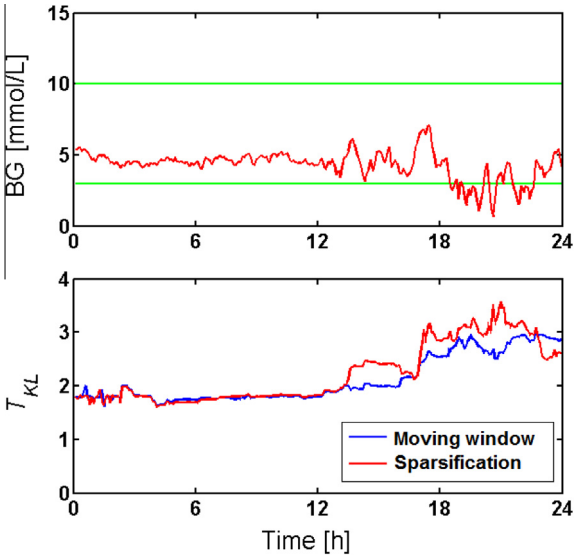

Fig. 7. Moving window strategy vs. sparsification of the state space.

upon a window moving strategy containing the last 30 states as well as using sparcification. The maximum number of points in the dictionary and the threshold parameter are set to $N_{\max} = 30$ and $\eta = 0.5$, respectively. A perception of the physiological state of a diabetic patient is conveniently defined through $\mathbf{x}_k = (G_k, I_{k-1})$. The control action u_k, is the change made to the insulin infusion rate. Each point $\mathbf{d}_k \in \mathfrak{D}$ is built by the current state-action pair and the corresponding state difference $\mathbf{d}_k = (\mathbf{x}_k, u_k, \Delta\mathbf{x}_k)$. An advantage of perceiving the patient state in this way is the fact that it only involves readily known variables, yet they are informative enough to describe its physiological condition for successfully controlling glycemic variability. It is noticeable how sparsification speeds up the detection of a deviant behavior in the AP. This is achieved by incorporating into the dictionary only states that are relevant to describe current changes in the loop dynamics. Surprise quickly rises when variability is augmented, reaching its maximum value at the hypoglycemic event. Moreover, Bayesian surprise also quickly decreases at the end of the curve, when glycemic values return to normality.

3.8. On-line monitoring

Algorithm 3 sums up the approach for on-line behavior monitoring. The training inputs to the Gaussian model GP^g are state-action pairs and the targets are the differences between the successor state and the state in which the action is applied, as given in Eq. (4). In line 2, the suboptimal control policy π^g interacts with a system model to build an initial training dictionary $\mathfrak{D}^0$; a total of l states are collected. Each point $\mathbf{d}_k \in \mathfrak{D}$ is formed as $\mathbf{d}_k = (\mathbf{x}_k, u_k, \Delta\mathbf{x}_k)$, where u_k describe the current action taken at state $\mathbf{x}_k$ and causing a change $\Delta\mathbf{x}_k$ in the stochastic process for a controlled variable. From the initial set $\mathfrak{D}^0$ only a set of $N_{\max}$ states are used to model GP^g in line 3. The most $N_{\max}$ relevant states are extracted from the initial set using sparsification to define the support dictionary $\mathfrak{D}^r$. As new states are acquired through interaction of the control policy and the GP model, whereas sparsification is used to update the support dictionary. For the current state-action pair the successor state is estimated for both the controlled and the specified model dynamics. Note that both the stochastic process GP^p modeled around the training set $\mathfrak{x} = \{\mathbf{x}_k, u_k^p), (\Delta\mathbf{x}_{k+1})\}$ and its optimal control policy $\pi^p(x)$ are both inputs to the procedure. Bayesian surprise is used to compute the performance of the controlled dynamics GP^g against the corresponding specified dynamics GP^p for optimal action selection.

Algorithm 3. *On-line monitoring*

1: **Input**: l, m, η, $N_{\max}$, GP^p, π^p, W, $\mathbf{x}_0$
2: Interact with the controlled system during l samples applying π^g to setup $\mathfrak{D}^0$
3: $\mathfrak{D}^r \leftarrow$ sparcification $(\mathfrak{D}^0, \eta, N_{\max})$
4: Train GP^g around $\mathfrak{D}^r$
5: **For** $k = 1$ to m
6: Compute $u_k^g = \pi^g(\mathbf{x}_{k-1})$ and $u_k^p = \pi^p(\mathbf{x}_{k-1})$
7: Compute $\mathbf{x}_k$ applying u_k^g to the model/system
8: Estimate $\Delta x_k^g \sim GP^g(\mathbf{x}_k, u_k^g)$ and $\Delta x_k^p \sim GP^p(\mathbf{x}_k, u_k^p)$
9: $\boldsymbol{X}^g = \{\Delta x_k^g\}_{k-W}^{k}; \boldsymbol{X}^p = \{\Delta x_k^p\}_{k-W}^{k}$
10: $P_{KL} \leftarrow$ Stepwise surprise $(\Delta x_k^g \| \Delta x_k^p)$
11: $T_{KL} \leftarrow$ Robust surprise $(\boldsymbol{X}^g \| \boldsymbol{X}^p)$
12: $\mathfrak{D}^r \leftarrow$ sparcification $(\mathfrak{D}^r, \eta, N_{\max}, d_k)$
13: Update GP^g with $\mathfrak{D}^r$
14: **End for**

4. On-line monitoring of an artificial pancreas

4.1. Artificial pancreas challenges

To guarantee an optimally-controlled glucose dynamics, safe implementation of an AP requires proper functioning and communication among all their components at all times. More specifically, it has been concluded that the existing technology still continues to face challenges in terms of sensitivity, stability, calibration, and physiological time-lags (Klonoff, 2007). The reliability of wireless communication between the components also needs to be addressed. Performance degradation in any part of the AP may easily lead to terminating its closed-loop operation.

Though available technology put forward a clinically applicable AP, many issues still prevent safe and optimal operation of closed-loop therapies. For example, continuous subcutaneous insulin infusion (CSII) devices are prone to mechanical failures, being the obstruction of the infusion catheter -because fibrin and cells occlude the inner lumen- the most common event (Bousquet-Rouaud et al., 1993). Blockage of a CSII can lead to ketoacidosis and episodes of hyperglycemia due to a lack of insulin or a life-threatening hypoglycemia event when abruptly releasing an excess of insulin into the body. In turn, implantable CGMS are limited by local factors and substances that can generate unstable output signals or interferences (Jaremko & Rorstad, 1998). As CGMS are placed in the subcutaneous tissue and thus they determine interstitial fluid glucose (IG) rather than blood glucose (BG) concentration, the existence of a BG-to-IG kinetics leads to a time-lag between both concentrations. As this lag grows, due to sensor deterioration, it can give rise to highly correlated errors in the series of glucose readings. Since insulin infusions are based on real-time sensor readings, calibration problems in the CGMS devices can have critical impact in the patient's daily life (Sparacino, Facchinetti, & Cobelli, 2010). Furthermore, as the control algorithm may be run in a mobile-phone device, with a wireless connection to the glucose sensor and the insulin pump, it can be affected by delays or interference in data transmission. Even the control algorithm itself is prone to poor performance whereas errors in calculating the insulin bolus required can lead to severe hypoglycemia events or poor glucose control. As was previously mentioned, excessive glycemic variability contributes to oxidative stress, which has been linked to the development of long-term diabetic complications.

4.2. Overview of faulty scenarios

Proper handling of faulty conditions in an AP is a crucial element in diabetes care. Reliable information about glycemic variations allows physicians and patients to evaluate the success of insulin infusion to maintain BG within the normoglycemic range. Checking if a safety–critical system such as the AP respects its specification is of paramount importance before and after its implementation. Hence, it has to be proven the correctness of the system in any possible scenario. A sort of formalism, to validate an AP operation, is essential in safety–critical systems analysis, since errors can have costly or disastrous consequences. To this aim, Chassin, Wilinska, and Hovorka (2004) suggested a systematic methodology aimed to classify acceptable and unacceptable functioning of a glycemic regulator in a simulation environment. The performance is evaluated in a range of treatment scenarios to guarantee safety and to demonstrate acceptable treatment efficacy. Safety and performance criteria are progressively evaluated. Fig. 8 depicts the validation environment that represents the interaction between a patient with Type 1 diabetes, a measurement device, a glucose controller and an insulin pump with their respective regulatory and technical characteristics. Specification checking takes into account different dynamical settings in order assess the behavior of the AP under faulty operating conditions. The value of the approach lies in the creation of a suitable simulation environment representing truthful conditions. A complete methodology requires specific tools to support clinical monitoring of glucose regulation. Bayesian surprise is a suitable metric to check correctness of AP operation according to the specified system behavior.

4.3. Simulation results

Algorithm 3 is applied to characterize the controlled dynamics using a suboptimal policy, by assuming certain degree of variability σ and a multiple meal regime as in Table 1. The specified dynamic -when the AP is controlled by the optimal policy- is performed using the following setting: variability $\sigma = 0.10$, insulin sensitivity $Sp = 0.5$, calibration error $\xi = 2\%$ and time-lag $\tau = 5$ min. From a total of $l = 200$ random initial states, $N_{\max} = 30$ are selected through sparsification to form the training set $\mathfrak{D}^0$ using $\eta = 0.5$.

Suboptimal regulation of the system is accomplished through a PID algorithm as well as an integrated Fuzzy-PID scheme. PID controllers are reactive to IG measurements and have yet not proven

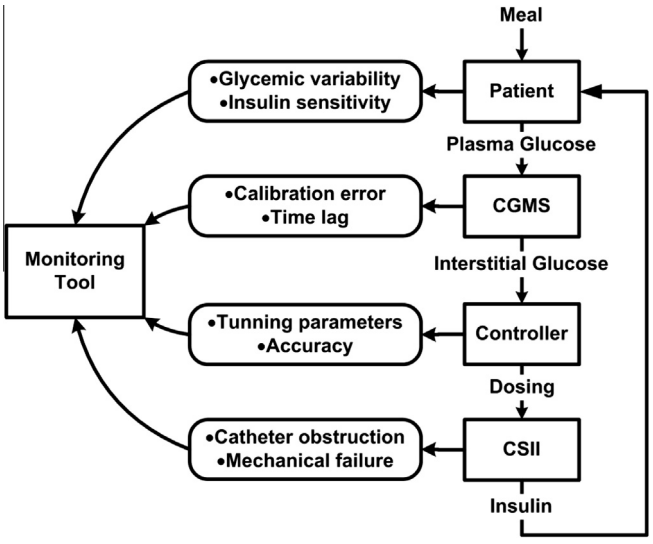

Fig. 8. Overview of the validation methodology for blood glucose regulation in virtual environments.

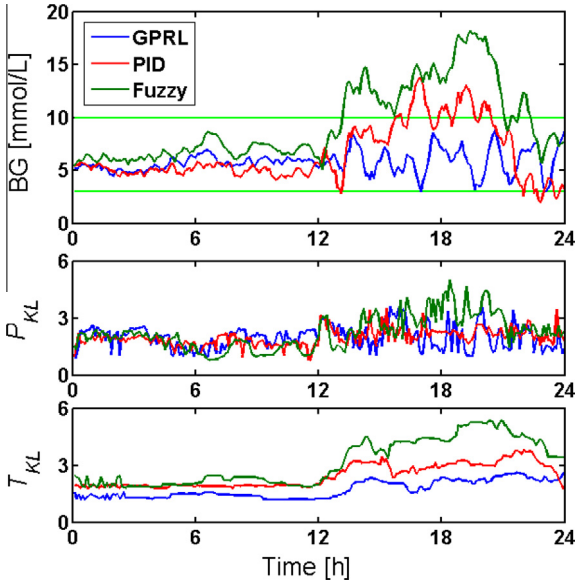

Fig. 9. Assessment of the effect of high glycemic variability using parameter $\sigma = 0.50$.

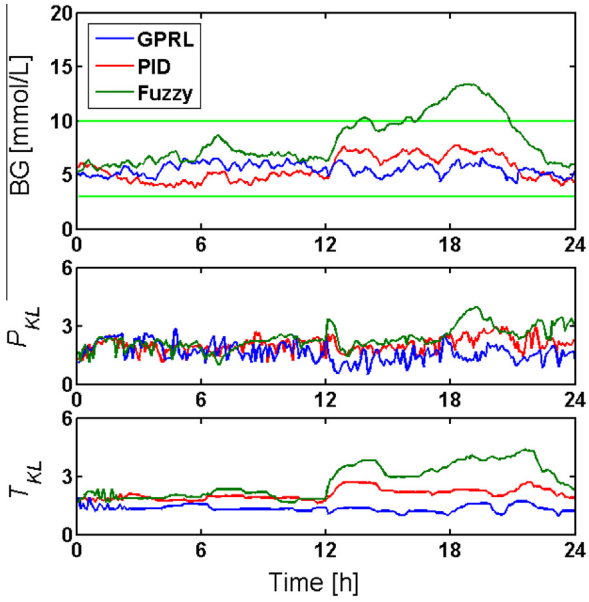

Fig. 10. Assessment of effect of low insulin sensitivity $Sp = 0.1$.

sufficiently efficacy in maintaining normoglycemia after meals. In an automatic insulin delivery in vivo experiment, mean values of glucose readings during closed-loop control employing a PID algorithm was similar to that achieved under standard insulin subcutaneous therapy (Steil, Rebrin, Darwin, Hariri, & Saad, 2006). Hence, PID and Fuzzy-PID algorithms are used as representative examples of performance degradation in glycemic control. Lower performance obtained through suboptimal control, is later assessed to manage IG under faulty conditions against the reference behavior obtained using the GPRL algorithm. This allowed us to evaluate the clinical impact of ordinary glycemic control that is affected by sensor errors as well as time-lags compounded with the effect of patient variability in diabetes management. To this aim, at 12th hour, sensor and insulin sensitivity parameters are significantly changed to simulate performance degradation in the AP

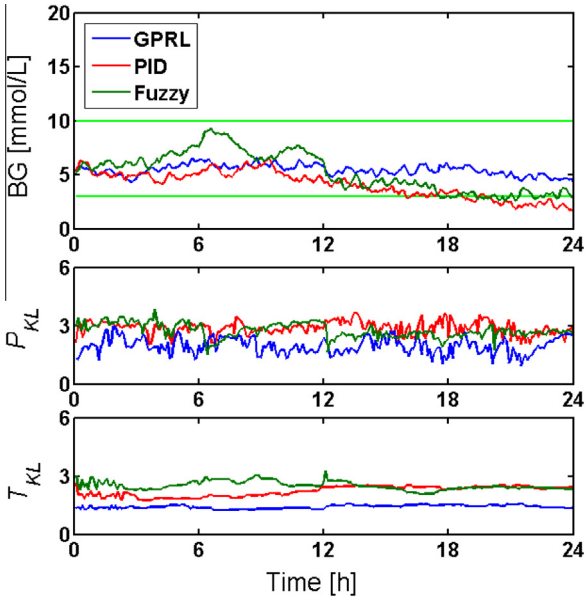

Fig. 11. Assessment of the effect o high insulin sensitivity $Sp = 0.9$.

components, whereas σ is increased to augment glycemic variability.

4.3.1. Glycemic variability and insulin sensitivity

In Fig. 9, the ability of each controller (GPRL, PID and PID-Fuzzy) to control glycemic variability is evaluated. The value of the Ito's parameter of the glucose–insulin model is increased from $\sigma = 0.10$ to 0.50 at 12 h, representing a larger glycemic variability. It is noteworthy how the performance of the closed-loop fast degrades. But even when certain degradation exists in the dynamic of the loop controlled with the GPRL controller, its robustness ensures glycemic levels falling inside the normoglycemic range despite a high level of intrinsic variability. While dissimilarity between the controlled dynamics using each control schemes is exposed by the P_{KL} surprise metric, the effect is more difficult to discern than that for the robust surprise. Initially, the T_{KL} values are no informative until the initial dictionary set $\mathfrak{D}^r$ captures enough points to describe the intrinsic loop variability. This occurs approximately at the 2.5 h for a given dictionary of size $N_{\max} = 30$. Even when the distance between distributions measures by T_{KL} represents a progressive change in the functioning of the loop, the mean values over a determined horizon of the values for P_{KL} should be closed to those obtained using T_{KL}.

In the second scenario, the issue of changes in the patient's insulin sensitive is addressed. As sensitivity varies from person to person, proper glycemic control should demonstrate adaptability to different patients. In Fig. 10, the insulin sensitivity parameter is set to a minimum value of $Sp = 0.1$. It is worth noting that an insulin bolus is less effective to decrease glycemic levels than PID control and even worse compared to the Fuzzy-PID controller. An opposite effect can be observed in Fig. 11 by setting $Sp = 0.9$. As the insulin infusion has a stronger effect over IG levels, suboptimal control actions lead to an apparent hypoglycemic event. As the GPRL controller is based upon a policy then loop behavior is not affected.

4.3.2. Continuous glucose monitor system

Arguably, the CGMS is the most sensitive and challenging component in any promising automatic glucoregulatory system. Implantable sensors help minimize glycemic risks, but such devices usually require regular plasma glucose readings to maintain proper calibration. As CGMS are placed in the subcutaneous tissue, they determine interstitial fluid rather than blood or plasma concentration. Hence, under dynamic conditions IG and BG values can be markedly different because of a time-lag between both concentrations. The magnitude of the lag may be no more than 5 min in optimal conditions, but after prolonged implantation the sensor surface become increasingly fouled with fibrotic substance and the time-lag may progressively increase. To account for the time lag, the approach proposed by Facchinetti, Sparacino, and Cobelli (2010) is applied, where a simulation model is used to assess the effect of sensor calibration errors and time-lags between plasmatic and interstitial glucose resulting from standard glycemic monitoring. In Fig. 12, plasma glucose levels and CGMS responses are depicted for different sensor errors and time-lag values.

Even though BG is usually referred to instead of IG, it is worth recalling that glucose readings accounts for levels obtained by means of a CGMS. Plasma glucose readings affected by sensor miscalibration of $\xi = 10\%$ and time-lag equal to $\tau = 20$ min are depicted in Figs. 13 and 14, respectively. Remind that as the control algorithm responds to the outcome of the CGMS, it actually acts over inaccurate readings rather than true BG levels. As the control loop deviates from the optimal behavior an increase in the Bayesian surprise is observed.

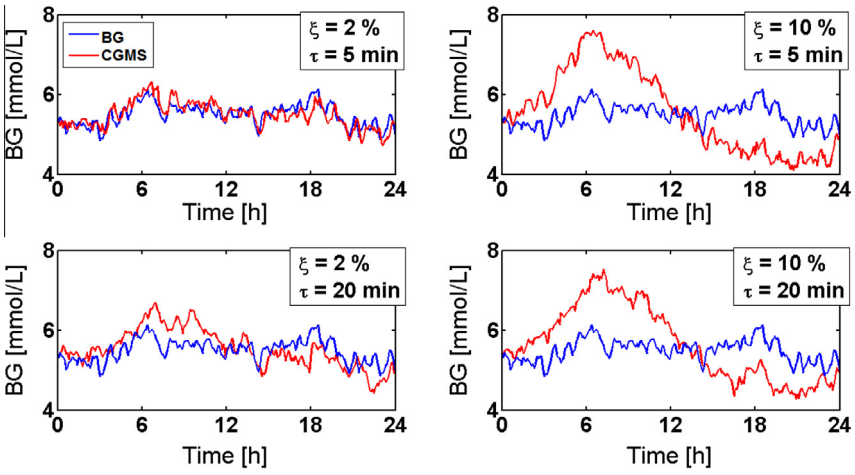

Fig. 12. CGMS reading errors (calibration and time lags) distorting plasma glucose levels.

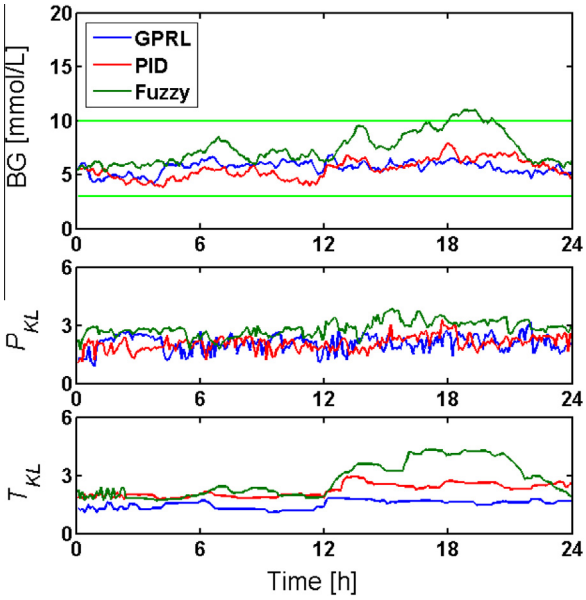

Fig. 13. Performance loss due to glucose sensor miscalibration using $\xi = 10\%$.

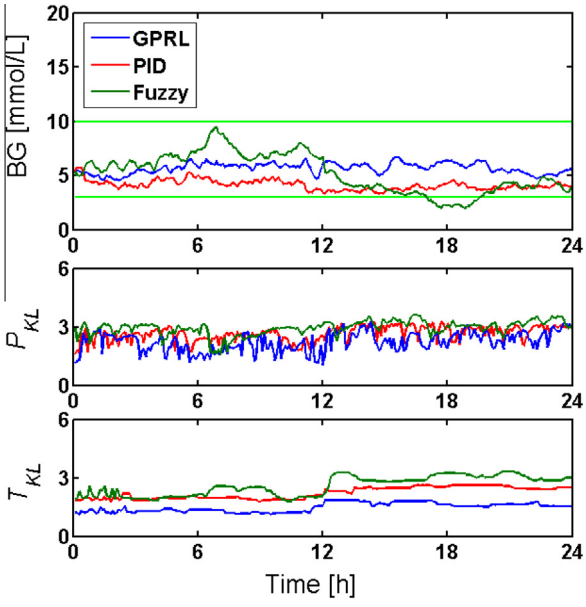

Fig. 14. Performance degradation due a large time-lag using $\tau = 20$ min.

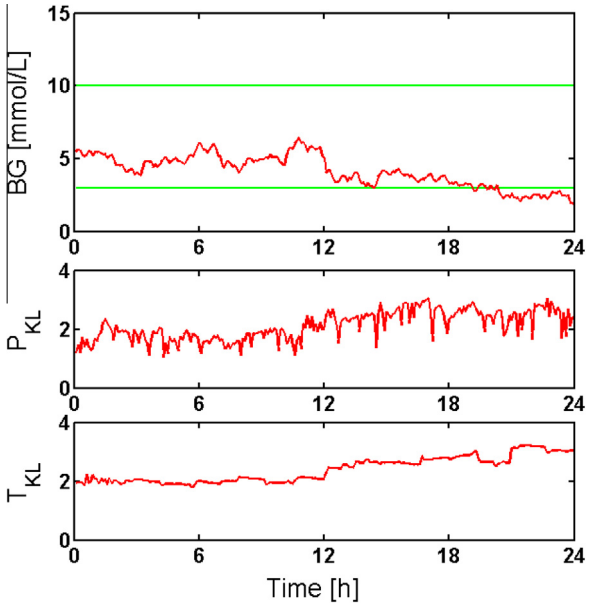

Fig. 15. Performance degradation due to an excessive proportional action in the PID controller.

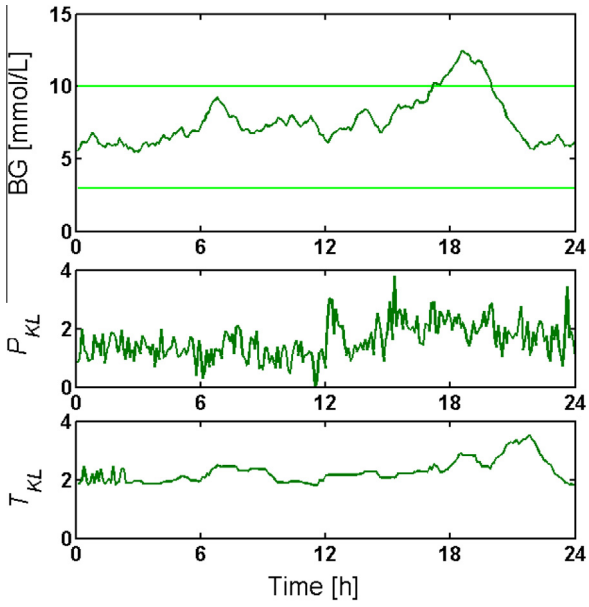

Fig. 16. Fuzzy-PID ill-tuning as a result of restricting the maximum insulin bolus allowed.

4.3.3. Closed-loop controller

As a particular case, let's consider an ill-tuning situation in the gain parameter k_C of a PID controller. Excessive proportional gain results in a large change in the output action for a given change in the error signal $(G(t) - G_x)$ in Eq. (23). If the proportional gain is too high, the system can become unstable. In Fig. 15 the value of the proportional parameter is increased from $k_C = 12$ to $k_C = 30$. Instability caused by the excessive gain gives rise to a hypoglycemic event. In Fig. 16, the sliding scale for the proportional control of Fuzzy-PID scheme, was restricted to prescribe up to a maximum bolus of 4 U/h, reducing at a half the effect of the proportional action. Performance degradation in both system dynamics are clearly revealed by Bayesian surprise, either for a hypoglycemic event due to a excessive control action in the PID loop, or an hyperglycemic event caused by a saturation of the proportional action in the fuzzy controller.

4.3.4. Continuous subcutaneous insulin infusion

Current commercially available insulin pumps deliver insulin continuously and subcutaneously. Intensive insulin therapy with tight glucose control significantly reduces diabetic complications but increases the risk of hypoglycemia. However, there exists local and systemic complications from controlled insulin delivery that include improper dosing due to electronic failures, catheter obstructions, battery depletion and infections (Bousquet-Rouaud et al., 1993). Incidents were diabetic patients, using a CSII device, face diverse complications due to insulin pump malfunctioning were studied in an *in silico* work by Kovatchev, Breton, Dalla Man, and Cobelli (2009). The simulation presented in Fig. 17 tests a faulty scenario in which a catheter blockage occurs during the use of a CSII device. As a result, insulin dose administrated via the pump is reduced significantly (say, 20%) with respect to the

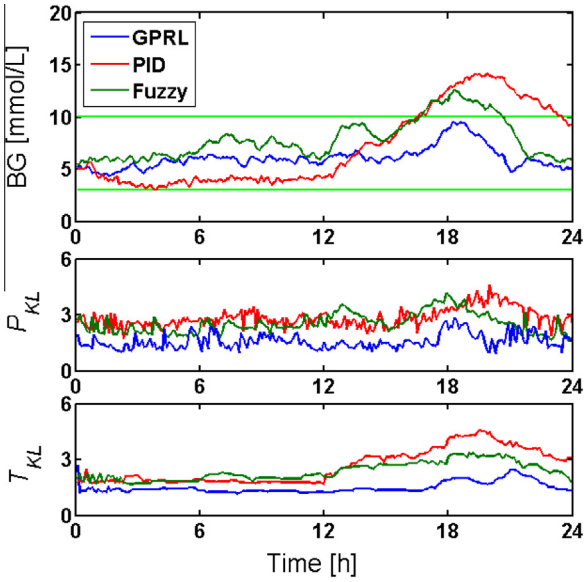

Fig. 17. Performance degradation due to a catheter blockage in an implantable insulin pump.

level prescribed by the control algorithm. This reduces the effect of the bolus and leads glucose levels to unacceptable values. Bayesian surprise quickly pinpoints that a dangerous performance degradation in the AP has occurred.

5. Concluding remarks

Checking if an autonomous agent behavior respects its specification is a key issue to guarantee safety and performance of an increasing number of applications such as driverless cars, drones and artificial pancreas. The main problem for behavior monitoring is defining the specification under the uncertainty the agent should face in its environment. In this work, it is argued that for a class of problems it is possible to approximate an optimal control policy for the agent using reinforcement learning and interactions between the agent and a simulated environment. To obtain the specification different simulated scenarios must be considered where the stochastic variability of the effect of control actions is accounted for. The agent behavior resulting from optimal action selection is then used as reference or expected behavior (benchmark) to assess trace equivalence. The probabilistic nature of the specification calls for a probabilistic characterization of the magnitude of any sub-optimal behavior. The Bayesian framework was then utilized to compute the prior and posterior PDFs, and thus derive surprise.

To assess if an agent implementation respects its specification, its desired behavior is modeled as a prior distribution for state transitions. Bayesian surprise is then used to quantify how observing each successive state in an unfolding sequence affects the prior assumption about optimal behavior. Therefore, we formally measure surprise by quantifying the distance (or divergence) between an agent behavior and the desired optimal policy (specification). A distinctive advantage of computing Bayesian surprise using GPs is that the divergence from the specified behavior can be estimated not only using the expected value of state transitions but also accounting for the corresponding prediction uncertainty for optimal action selection. The surprise metric is based on subjective measures of information and thus the surprise value of a sample depends on the present state of the system being monitored. Since the goal of this paper was to create an evolving system that will use online learning to compare observed state transitions against a benchmark, it is important to detect outliers and prevent overfitting from redundant data.

Two surprise metrics for on-line loop monitoring have been proposed. The stepwise surprise has the ability to rapidly detect unexpected course of actions in the agent implementation, which is crucial in implementing safety–critical systems such as the AP. However, the stepwise surprise is computed step-by-step and thus its value may result too noisy for on-line behavior monitoring. To overcome this drawback, a robust Surprise metric based on a vis-à-vis comparison of the agent implementation and the specification using a sequence of state transitions is proposed. A surprise index based on twin Gaussian processes is introduced in order to detect on-line meaningful indications of an anomalous or deviant behavior. In order to properly characterize the implemented agent behavior, it is necessary to address the relevance of an arriving data stream and using only those state transitions that contribute the most to the accuracy of the GP model that describe the observed agent control policy. Hence, we considered that sparse approximations are useful to compactly represent the support data set for on-line monitoring.

The on-line behavior monitoring approach has been tested on a stochastic model of the artificial pancreas. Different abnormal conditions are simulated to assess the usefulness of Bayesian surprise to pinpoint performance degradation when there are faults in sensor, controller and pump devices as well as abnormal glycemic variability. For this particular application, the concept of a reference behavior is directly linked to optimal action selection in a healthy person. Even though clinical assays are still pending, incorporating the proposed approach in the artificial pancreas will be an outstanding breakthrough for its acceptance and improvement. Moreover, the concept of Bayesian surprise can also be based on a reference behavior where optimal action selection is derived from healthy persons that are somewhat similar to a given patient. This is also an important advantage of the proposed approach since available expert or domain knowledge about the agent specification can be readily integrated in on-line monitoring. Furthermore, a practical advantage of the proposed approach is that the agent implementation is easily testable (decidable from interaction) against optimal action selection using sampled realizations of both (the specification and the implementation) stochastic processes.

The proposed approach has also some weaknesses or limitations that should be addressed. The most important is that it has been assumed that agent's environment is "passive", does not change its response to agent course of actions over time. This prevents applying the proposed approach to multi-agent systems where the specification should account for a Nash or Stackelberg equilibrium. Current research work attempts to extend Bayesian surprise to multi-agent systems by characterizing the desired agent behavior using a game-theoretic perspective. Another, weakness is that on-line monitoring is assumed decoupled from the agent policy for optimal acting. That is, the monitor detects a deviant behavior but no course of action for correcting it is taken. To this aim, we are currently working on tightly integrating the monitor within the agent design in the framework of autonomic systems so that the agent can self-monitor and self-optimize its decision-making policy.

Appendix A. Stochastic model of Type I diabetic patient

A.1. Deterministic model of the glucose–insulin dynamics

The Lehmann and Deutsch (1992) model parameterized as described in Acikgoz and Diwekar (2010) is used as the basis to describe the deterministic glucose–insulin dynamics. The model attempts to account for the underlying physiology of insulin action

and carbohydrate absorption in quantitative terms such as insulin sensitivity, volume of glucose and insulin distribution and maximal rate of gastric emptying.

The model assumes a patient that completely lacks endogenous insulin secretion. It contains a single glucose pool representing extracellular glucose (including blood glucose) into which glucose enters via both intestinal absorption and hepatic glucose production. Glucose is removed from this pool by insulin independent glucose utilization in red blood cells (RBCs) and the central nervous system (CNS) as well as by insulin-dependent glucose utilization in the liver and periphery, the latter taking place mostly in muscle and adipose tissue. In the model, hepatic and peripheral glucose usage are dealt with separately. Glucose excretion takes place above the renal threshold for glucose concentration as a function of the creatinine clearance rate.

Four differential equations along with twelve auxiliary relations and the experimental data from Guyton et al. (1978) constitute the model which is solved by numerical integration. The change in the plasma insulin concentration I is given by the following equation

$$\frac{dI}{dt} = \frac{I_{abs}}{V_I} - k_e I \tag{A.1}$$

where k_e is the first-order rate constant of insulin elimination, I_{abs} is the rate of insulin absorption and V_I is the volume of insulin distribution. The build-up and the deactivation of the active insulin pool I_a is assumed to obey a first-order kinetics

$$\frac{dI_a}{dt} = k_1 I - k_2 I_a \tag{A.2}$$

where k_1 and k_2 are first-order rate constants which serve to describe the delay in insulin action. The rate of insulin absorption is modeled according to Berger and Rodbard (1989) as

$$I_{abs}(t) = \frac{st^s T_{50}^s D}{t[t^s + T_{50}^s]^2} \tag{A.3}$$

where t is the time elapsed from the injection, T_{50} is the time at which 50% of the dose D has been absorbed and s is a preparation-specific parameter defining the insulin absorption pattern of the different types of insulin catered for in the model (regular, intermediate, lente and ultralente). The linear dependency of T_{50} on the insulin dose is defined as

$$T_{50}^s = aD + b \tag{A.4}$$

where a and b are preparation-specific parameters. These values are given in Berger and Rodbard (1989) along with values for s. The steady-state insulin profile, I_{ss}, corresponding to a given regime, is computed by using the superposition principle assuming that three days are enough to reach steady-state conditions, that is, the steady-state response results from the composite effect of injections given for three subsequent days.

$$\begin{aligned} I_{ss}(t) &= I(t) + I(t+24) + I(t+48) \\ I_{a,ss}(t) &= I_a(t) + I_a(t+24) + I_a(t+48) \end{aligned} \tag{A.5}$$

Since the experimental data provided by Guyton et al. (1978) refer to equilibrium conditions, the insulin level equilibrated with the steady-state active insulin is considered when computing the net hepatic glucose balance and peripheral glucose uptake. In other words, at any time during the simulation, we assume steady-state values for I_{ss} and $I_{a,ss}$ and use

$$I_{eq}^*(t) = k_2 I_{a,ss}(t)/k_1 \tag{A.6}$$

as the insulin level responsible for the hepatic and peripheral control action, where I_{eq}^* is the insulin level in equilibrium with $I_{a,ss}$.

Assuming a single compartment for extracellular glucose, the change in glucose concentration with time is given by the differential equation

$$\frac{dG}{dt} = \frac{G_{in}(t) + NHGB(t) - G_{out}(t) - G_{ren}(t)}{V_G} \tag{A.7}$$

where G represents the plasma glucose concentration, G_{in} is the systemic appearance of glucose via glucose absorption from the gut, $NHGB$ is net hepatic glucose balance, G_{out} is the overall rate of peripheral and insulin dependent glucose utilization, G_{ren} is the renal excretion of glucose and V_G is the volume of distribution of glucose.

As the liver produces and also utilizes glucose depending on the blood glucose and insulin levels, hepatic glucose handling is modeled in terms of $NHGB$ which is computed as the sum of gluconeogenesis, glycogen breakdown and glycogen synthesis data derived for different blood glucose and insulin levels from nomograms given in Guyton et al. (1978). Table A1 shows how the $NHGB$ varies as a function of glucose and normalized insulin levels. The hepatic insulin sensitivity parameter S_h (normalized between 0 and 1) is used in the computation of the effective insulin level which controls hepatic glucose usage. The net hepatic glucose balance for any arterial blood glucose level between 1.1 [mmol l^{-1}] and 4.4 [mmol l^{-1}] is computed by interpolation between the values shown on the curves in Table A1. Data shown are based on the steady state plasma insulin level which is normalized with respect to the basal level I_{basal} to obtain an effective plasma insulin level I_{eff}.

Assuming a classical Michaelis–Menten relationship between glucose utilization and the plasma glucose concentration, the insulin concentration is thus reflected in different values of the maximal rate of the transport process as follows

$$G_{out}(G, I_{eq}^*) = \frac{G(cS_p I_{eq}^* G_I)(K_m + G_X)}{G_X(K_m + G)} \tag{A.8}$$

where K_m Michaelis–Menten constant, c is the slope of the peripheral glucose utilization versus insulin level relationship, G_I is the insulin independent glucose utilization and G_X is a reference glucose level. The $NHBG$ value at any combination of G and I_{eq}^* has been derived from the data summarized in Table A1 using $S_h I_{eq}^*$ as the effective insulin level.

The amount of glucose in the gut, G_{gut}, following the ingestion of a meal containing Ch millimoles of glucose equivalent carbohydrate is defined as

$$\frac{dG_{gut}}{dt} = G_{empt} - k_{gabs} G_{gut} \tag{A.9}$$

where k_{gabs} is the rate constant of glucose absorption from the gut into the systemic circulation and G_{empt} is the rate of gastric emptying which is a function of time. The duration of the period T_{max} for

Table A1
NHGB as a function of the arterial blood glucose level AG and the effective plasma insulin level I_{eff}.

I_{eff}	$AG \leqslant 1.1$ mmol l^{-1}	$AG = 3.3$ mmol l^{-1}	$AG \geqslant 4.4$ mmol l^{-1}
0	291.6	160.0	78.3
1	194.6	114.6	53.3
2	129.3	66.0	−1.7
3	95.7	46.3	−54.3
4	85.0	22.6	−76.0
5	76.3	4.3	−85.0
6	69.0	−10.0	−92.0
7	62.0	−25.3	−97.3
8	52.0	−43.3	−101.0
9	48.0	−47.3	−104.0
10	41.7	−49.3	−106.7

Table A2
Patient-independent model parameter values.

$k_e = 5.4$ [h^{-1}]	Insulin elimination rate constant
$k_1 = 0.025$ [h^{-1}]	Parameter for insulin pharmacodynamics
$k_2 = 1.25$ [h^{-1}]	Parameter for insulin pharmacodynamics
$I_{basal} = 10$ [mU l^{-1}]	Reference basal level of insulin
$K_m = 10$ [mmol l^{-1}]	Michaelis constant for enzyme-mediated glucose uptake
$G_I = 0.54$ [mmol h^{-1} kg^{-1}]	Glucose utilization per kg body weight
$G_X = 5.3$ [mmol l^{-1}]	Reference value for glucose utilization
$c = 0.015$ [mmol h^{-1} kg^{-1} mU^{-1} l]	Peripheral glucose utilization versus insulin line
$k_{gabs} = 1$ [h^{-1}]	Rate constant for glucose absorption from the gut
$V_{max} = 120$ [mmol h^{-1}]	Maximal rate of gastric emptying
$V_I = 0.142$ [kg^{-1}]	Volume of distribution for insulin per kg body weight
$V_G = 0.221$ [kg^{-1}]	Volume of distribution for glucose per kg body weight

which gastric empty is constant and maximal is a function of the carbohydrate content of the meal ingested

$$T_{max} = [Ch - 1/2V_{max}2(T_{asc} + T_{des})]/V_{max} \tag{A.10}$$

where V_{max} is the maximal rate of gastric emptying and T_{asc} and T_{des}, are the respective lengths of the ascending and descending branches of the gastric emptying curve which have default values in the model of 30 min.

However, for small quantities of carbohydrate (below approximately 10 g) such values cannot be used because there will never be enough time for the gastric emptying curve to plateau out. In such cases, T_{asc} and T_{des} are defined as

$$T_{asc} = T_{des} = 2Ch/V_{max} \tag{A.11}$$

Based on a triangular function. Eq. (A.11) is only used when the quantity of carbohydrate ingested falls below a critical level Ch_{crit} which is defined as

$$Ch_{crit} = [(T_{asc} + T_{des})V_{max}]/2 \tag{A.12}$$

Using linear interpolation the rate of gastric empty for meals containing Ch millimoles of carbohydrate greater than Ch_{crit} can therefore be defined, according to the time elapsed from the start of the meal, t, as follows

$$\begin{aligned} &G_{empt} = (V_{max}/T_{asc})t \quad \text{if} \quad t < T_{asc} \\ &G_{empt} = V_{max} \quad \text{if} \quad T_{asc} < t \leqslant T_{asc} + T_{max} \\ &G_{empt} = V_{max} - (V_{max}/T_{des})(t - T_{asc} - T_{max}) \\ &\qquad \text{if } T_{asc} + T_{max} \leqslant t < T_{max} + T_{asc} + T_{des} \\ &G_{empt} = 0 \quad \text{elsewhere} \end{aligned} \tag{A.13}$$

Glucose input via the gut wall G_{in} can be modeled by

$$G_{in} = k_{gabs}G_{gut} \tag{A.14}$$

Values for all model parameters, which have been derived from Berger and Rodbard (1989) and Guyton et al. (1978) are given in Table A2. All parameters, except S_p and S_h, are assumed to be patient independent.

The rate of renal glucose excretion, G_{ren} in the model is defined as

$$\begin{aligned} &G_{ren} = GFR(G - RTG) \quad \text{if} \quad G > RTG \\ &G_{ren} = 0 \qquad\qquad\quad \text{elsewhere} \end{aligned} \tag{A.15}$$

for blood glucose values (G) above the renal threshold of glucose (RTG), where GFR is the glomerular filtration (creatinine clearance) rate. Default values have been set to $RTG = 9.0$ [mmol l^{-1}] and $GFR = 100$ [ml min^{-1}]. These default values are used for all patient cases except when renal dysfunction is suspected and the clinical parameters are actually measured. The renal excretion of glucose (G_{ren}) is zero for blood glucose values below its renal threshold.

A.2. Modeling uncertainty in blood glucose dynamics

A deterministic dynamic model of blood glucose is provided by Eqs. (A.1), (A.2), and (A.7). However, there exists uncertainty in the estimation of model parameters that prevents describing intra- and inter-variability among daily values of glucose in patients. Glycemic variability can be represented by introducing a stochastic process in the presented deterministic model.

Considering the simplest generalization of an Ito stochastic process (1951) given by

$$dx = \lambda dt + \sigma dz \tag{A.16}$$

where the quantity dz corresponds to the increment of a Wiener process and is given by

$$dz = \varepsilon_t\sqrt{dt} \rightarrow dx = \lambda dt + \sigma\varepsilon_t\sqrt{dt} \tag{A.17}$$

where λ is called the drift parameter, σ is the variance parameter and ε_t is a random number generated by a normal distribution $\varepsilon_t \sim \mathcal{N}(0,1)$.

By combining (A.7) and (A.17), the degree of uncertainty on the stochastic glucose dynamics is therefore modeled as a result of the dependence of the glycemic model on Ito's process parameters

$$\frac{dG}{dt} = \frac{G_{in}(t) + NHGB(t) - G_{out}(t) - G_{ren}(t)}{V_G} + \frac{\sigma\varepsilon}{\sqrt{dt}} \tag{A.18}$$

Because of this uncertainty, even the same meal and the same amount of physical exercise may result in different blood glucose responses in successive days. This is an important issue because blood glucose dynamics representation using Ito's process is a successful tool to approximate the natural variability of the regulatory system behavior. The proper addition of an Ito's stochastic process using the variance parameter σ provides synthetic patients with a wide range of physiological characteristics.

References

Acikgoz, U., & Diwekar, U. M. (2010). Blood glucose regulation with stochastic optimal control for insulin-dependent diabetic patients. *Chemical Engineering Science, 65*(3), 1227–1236.

Baldi, P. (2002). A computational theory of surprise. In *Information, Coding and Mathematics* (pp. 1–25). Springer.

Baldi, P., & Itti, L. (2010). Of bits and wows: A Bayesian theory of surprise with applications to attention. *Neural networks: The Official Journal of the International Neural Network Society, 23*(5), 649–666. http://dx.doi.org/10.1016/j.neunet.2009.12.007.

Bayarri, M. J., & Berger, J. O. (1997). Measures of surprise in Bayesian analysis. In *Duke university institute of statistics and decision sciences working paper* (pp. 97–46).

Berger, M., & Rodbard, D. (1989). Computer simulation of plasma insulin and glucose dynamics after subcutaneous insulin injection. *Diabetes Care, 12*(10), 725–736.

Bo, L., & Sminchisescu, C. (2010). Twin Gaussian processes for structured prediction. *International Journal of Computer Vision, 87*(1–2), 28–52.

Bousquet-Rouaud, R., Castex, F., Costalat, G., Bastide, M., Hedon, B., Bouanani, M., et al. (1993). Factors involved in catheter obstruction during long-term peritoneal insulin infusion. *Diabetes Care, 16*(5), 801–805.

Bremer, T., & Gough, D. A. (1999). Is blood glucose predictable from previous values? A solicitation for data. *Diabetes, 48*(3), 445–451.

Broggi, A., Medici, P., Zani, P., Coati, A., & Panciroli, M. (2012). Autonomous vehicles control in the VisLab intercontinental autonomous challenge. *Annual Reviews in Control, 36*(1), 161–171.

Chassin, L. J., Wilinska, M. E., & Hovorka, R. (2004). Evaluation of glucose controllers in virtual environment: Methodology and sample application. *Artificial Intelligence in Medicine, 32*(3), 171–181.

Cover, T. M., & Thomas, J. A. (2012). *Elements of information theory*. John Wiley & Sons.

De Paula, M., & Martinez, E. (2012). Probabilistic optimal control of blood glucose under uncertainty. *22nd European symposium on computer aided process engineering* (Vol. 30, pp. 1357–1361). Elsevier. Downloaded from <http://www.sciencedirect.com/science/article/pii/B9780444595201501305>.

Deisenroth, M. P., Rasmussen, C. E., & Peters, J. (2009). Gaussian process dynamic programming. *Neurocomputing, 72*(7–9), 1508–1524. http://dx.doi.org/10.1016/j.neucom.2008.12.019.

Dvijotham, K., & Todorov, E. (2012). Linearly solvable optimal control. In *Reinforcement learning and approximate dynamic programming for feedback control* (pp. 119–141).

Engel, Y., Mannor, S., & Meir, R. (2004). The kernel recursive least-squares algorithm. *IEEE Transactions on Signal Processing, 52*(8), 2275–2285. http://dx.doi.org/10.1109/TSP.2004.830985.

Facchinetti, A., Sparacino, G., & Cobelli, C. (2010). Modeling the error of continuous glucose monitoring sensor data: Critical aspects discussed through simulation studies. *Journal of Diabetes Science and Technology, 4*(1), 4–14.

Farmer, T. G., Edgar, T. F., & Peppas, N. A. (2009). Effectiveness of intravenous infusion algorithms for glucose control in diabetic patients using different simulation models. *Industrial and Engineering Chemistry Research, 48*(9), 4402–4414. http://dx.doi.org/10.1021/ie800871t.

Fernández-Caballero, A., Castillo, J. C., & Rodríguez-Sánchez, J. M. (2012). Human activity monitoring by local and global finite state machines. *Expert Systems with Applications, 39*(8), 6982–6993.

Franco, O. H., Steyerberg, E. W., Hu, F. B., Mackenbach, J., & Nusselder, W. (2007). Associations of diabetes mellitus with total life expectancy and life expectancy with and without cardiovascular disease. *Archives of Internal Medicine, 167*(11), 1145.

Gerich, J. E. (2005). The importance of tight glycemic control. *The American Journal of Medicine, 118*(9), 7–11.

Ghraizi, R. A., Martinez, E., & De Prada, C. (2009). Control loop performance monitoring using the permutation entropy of error residuals. In *International symposium on advanced control of chemical processes*.

Good, I. J. (1956). The surprise index for the multivariate normal distribution. *The Annals of Mathematical Statistics*, 1130–1135.

Good, I. J. (1988). Surprise index. *Encyclopedia of Statistical Sciences*.

Grosman, B., Dassau, E., Zisser, H. C., Jovanoviĉ, L., & Doyle, F. J. (2010). Zone model predictive control: A strategy to minimize hyper-and hypoglycemic events. *Journal of Diabetes Science and Technology, 4*(4), 961.

Günel, S. (2010). *An information-theoretic approach to nonlinear systems – In the view of the Fokker–Planck–Kolmogorov formalism*. Germany: LAP Lambert Academic Publishing.

Guyton, J. R., Foster, R. O., Soeldner, J. S., Tan, M. H., Kahn, C. B., Koncz, L., et al. (1978). A model of glucose–insulin homeostasis in man that incorporates the heterogeneous fast pool theory of pancreatic insulin release. *Diabetes, 27*(10), 1027–1042.

Hasanbelliu, E., Kampa, K., Principe, J. C., & Cobb, J. T. (2012). Online learning using a Bayesian surprise metric. In *The 2012 international joint conference on neural networks (IJCNN)* (pp. 1–8). IEEE.

Ito, K. (1951). *On Stochastic Differential Equations*. Memoirs of the American Mathematical Society. Downloaded from <http://www.archive.org/details/onstochasticdiff029540mbp>.

Itti, L., & Baldi, P. F. (2005). Bayesian surprise attracts human attention. In *Advances in neural information processing systems* (pp. 547–554).

Itti, L., & Baldi, P. (2005a). A principled approach to detecting surprising events in video. In *CVPR 2005: IEEE computer society conference on computer vision and pattern recognition, 2005* (Vol. 1, pp. 631–637). IEEE.

Jaremko, J., & Rorstad, O. (1998). Advances toward the implantable artificial pancreas for treatment of diabetes. *Diabetes Care, 21*(3), 444–450.

Jaynes, E. T. (1988). *How does the brain do plausible reasoning?* Springer.

Klonoff, D. C. (2007). The artificial pancreas: how sweet engineering will solve bitter problems. *Journal of Diabetes Science and Technology (Online), 1*(1), 72.

Kovatchev, B. P., Breton, M., Dalla Man, C., & Cobelli, C. (2009). Biosimulation modeling for diabetes: in silico preclinical trials: a proof of concept in closed-loop control of type 1 diabetes. *Journal of Diabetes Science and Technology (Online), 3*(1), 44.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics, 22*(1), 79–86.

Lehmann, E. D., & Deutsch, T. (1992). A physiological model of glucose–insulin interaction in type 1 diabetes mellitus. *Journal of Biomedical Engineering, 14*(3), 235–242. doi:16/0141-5425(92)90058-S.

Majda, A., Kleeman, R., & Cai, D. (2002). A mathematical framework for quantifying predictability through relative entropy. *Methods and Applications of Analysis, 9*(3), 425–444.

Marling, C. R., Shubrook, J. H., Vernier, S. J., Wiley, M. T., & Schwartz, F. L. (2011). Characterizing blood glucose variability using new metrics with continuous glucose monitoring data. *Journal of Diabetes Science and Technology, 5*(4), 871–878.

Nguyen-Tuong, D., & Peters, J. (2011). Incremental online sparsification for model learning in real-time robot control. *Neurocomputing, 74*(11), 1859–1867. http://dx.doi.org/10.1016/j.neucom.2010.06.033.

Plastino, A. R., Miller, H. G., & Plastino, A. (1997). Minimum Kullback entropy approach to the Fokker–Planck equation. *Physical Review E, 56*(4), 3927.

Ranganathan, A., & Dellaert, F. (2009). Bayesian surprise and landmark detection. In *ICRA'09. IEEE International Conference on* (pp. 2017–2023). IEEE.

Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian processes for machine learning*. MIT Press.

Salge, M., & Milling, P. M. (2006). Who is to blame, the operator or the designer? Two stages of human failure in the Chernobyl accident. *System Dynamics Review, 22*(2), 89–112.

Sanger, T. D. (2011). Distributed control of uncertain systems using superpositions of linear operators. *Neural Computation, 23*(8), 1911–1934.

Siegelaar, S. E., Holleman, F., Hoekstra, J. B. L., & DeVries, J. H. (2010). Glucose variability; does it matter? *Endocrine Reviews, 31*(2), 171–182. http://dx.doi.org/10.1210/er.2009-0021.

Sparacino, G., Facchinetti, A., & Cobelli, C. (2010). «Smart» continuous glucose monitoring sensors: On-line signal processing issues. *Sensors, 10*(7), 6751–6772.

Steil, G. M., Rebrin, K., Darwin, C., Hariri, F., & Saad, M. F. (2006). Feasibility of automating insulin delivery for the treatment of type 1 diabetes. *Diabetes, 55*(12), 3344–3350.

Steil, G. M., Rebrin, K., Hariri, F., Jinagonda, S., Tadros, S., Darwin, C., et al. (2005). Interstitial fluid glucose dynamics during insulin-induced hypoglycaemia. *Diabetologia, 48*(9), 1833–1840.

Susanto-Lee, R., Fernando, T., & Sreeram, V. (2008). Simulation of fuzzy-modified expert PID algorithms for blood glucose control. In *ICARCV 2008: 10th international conference on control, automation, robotics and vision, 2008* (pp. 1583–1589). IEEE.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. MIT Press.

Thimbleby, H. (2009). Contributing to safety and due diligence in safety–critical interactive systems development by generating and analyzing finite state models. In *Proceedings of the 1st ACM SIGCHI symposium on engineering interactive computing systems* (pp. 221–230). ACM.

Todorov, E. (2009). Efficient computation of optimal actions. *Proceedings of the National Academy of Sciences of the United States of America, 106*(28), 11478–11483.

Weaver, W. (1966). Probability, rarity, interest, and surprise. *Pediatrics, 38*(4), 667–670.

West, G. M., McArthur, S. D. J., & Towle, D. (2012). Industrial implementation of intelligent system techniques for nuclear power plant condition monitoring. *Expert Systems with Applications, 39*(8), 7432–7440.