Towards the Generation of Visual Qualia in Artificial Cognitive Architectures

Raúl Arrabales, Agapito Ledezma and Araceli Sanchis

Abstract The nature and the generation of qualia in machines is a highly controversial issue. Even the existence of such a concept in the realm of artificial systems is often neglected or denied. In this work, we adopt a pragmatic approach to this problem using the Synthetic Phenomenology perspective. Specifically, we explore the generation of visual qualia in an artificial cognitive architecture inspired on the Global Workspace Theory (GWT). We argue that preliminary results obtained as part of this research line will help to characterize and identify artificial qualia as the direct products of conscious perception in machines. Additionally, we provide a computational model for integrated covert and overt perception in the framework of the GWT. A simple form of the apparent motion effect is used as a preliminary experimental context and a practical case study for the generation of synthetic visual experience. Thanks to an internal inspection subsystem, we are able to analyze both covert and overt percepts generated by our system when confronted with visual stimuli. The inspection of the internal states generated within the cognitive architecture enable us to discuss possible analogies with human cognition processes.

1 Introduction

Understanding what qualia are and whether they have a functional role is one of the key issues in consciousness studies. Although much research work has been done in the domain of human consciousness, especially from the point of view of philosophy, little empirical work has been carried out in the field of Machine Consciousness (MC). This is not surprising given the elusive character of the phenomenal aspects of consciousness and the relative immaturity of the MC research field. Synthetic Phenomenology (SP), understood as the simulation or modeling of phenomenal states in artifacts [1], provides a pragmatic framework for the study of qualia in machines. In other words, the specification of the contents of subjective experience in machines, whether or not they can be claimed to be real pheno-
menal states, is a valuable method for making progress in the understanding of qualia. At least, as illustrated in this paper, the functional dimension of qualia can be analyzed.

Although empirical results obtained using machines are not directly applicable to humans, building models or simulations of experiential states can potentially shed some light on the problem of human conscious perception. In this particular case, as we have used an architecture inspired on the GWT, a plausible functional explanation of how conscious experience could be generated out of a global workspace is provided. Hence, in addition to the feedback that could be provided in terms of human cognition modeling, we argue that this SP approach can also provide a useful contribution for the design of functionally conscious machines.

Since the approach adopted in this work has a strong focus on functionality, no claims are made in terms of the presence of phenomenal states in the machine. We take no position as to what the required substrate for phenomenal states might be. In fact, our proposal is fully based on a computational model, which could be implemented as a virtual machine [2], in any standard Von Neumann computer (see Section 3 for a description of the current implementation). Consequently, the main working hypothesis proposed in this work is that qualia are the only contents of subjective experience and they have a clear functional role. Whether or not the proposed process for the generation of these explicit contents conveys associated phenomenal states remains to be proved.

Other authors have also approached the problem of qualia somewhat similarly, but focusing more on neuroscience or mathematical aspects of the so-called qualia space. For instance, Lehky and Sejnowski designed and artificial neural network able to map a variable wavelength input into a color space in such a way that the appearance of the white color quale (which is known not to be directly related to a particular wavelength) could be predicted [3]. Qualia have been also characterized from the point of view of the information integration theory. Balduzzi and Tononi have recently proposed a mathematical representation for characterizing the informational relationships of the qualia space [4].

In the remainder of this paper we describe and analyze the implementation of a minimal architecture for the generation of artificial visual qualia inspired on the GWT. In section 2, we briefly introduce the GWT and the associated minimal computational model that we have used. Section 3 covers the corresponding implementation as a cognitive architecture for robotics. The experimental setting is defined in section 4, and the corresponding preliminary results in section 5. Finally, we draw some conclusions in light of the obtained preliminary results.
2 A Computational Model Inspired on the Global Workspace Theory

Baars originally proposed the GWT as a metaphor for the access dimension of consciousness [5]. GWT explains access consciousness using the intuitive idea of a “theater” (see [6] for a definition of access consciousness in contrast with the phenomenal consciousness concept). Taking this theory as inspiration, we have built a computational model for the specification of the contents of subjective experience. In the following, we briefly introduce the GWT and then describe the derived computational model that we have used as framework for this research work.

2.1 Global Workspace Theory

Baars’ theater is a functional explanation of consciousness which can be considered as a contrary vision of the dualism advocated by the “Cartesian theater” idea. In Baars’ metaphorical theater, the scene corresponds to the working memory and the spotlight on the scene represents the focus of attention. The selection of the position of the spotlight is primarily done behind the scenes by the play director – executive guidance processes. The (conscious) action taking place under the bright spot on stage is formed thanks to a large set of specialized (unconscious) processors – the metaphorical audience in the dark – that can form coalitions and contribute their output to the workspace. There is a permanent competition between individual processors or coalitions of processors to contribute to the workspace. Contextual systems behind the scenes shape the content under the spotlight which will be globally available; i.e., once the content is shaped under the spotlight it is broadcasted to the audience (see Fig. 1).

In short, GWT is based on the idea that functional (or access) consciousness is produced in the brain thanks to the operation of a blackboard architecture [7]. There seems to be evidence that such a global access mechanism actually takes place in the brain. Some known neural mechanisms have been found to correlate with the functional roles described in the GWT [8, 9].
Fig. 1 Global Workspace Theory. The conscious contents of the mind are formed under the bright spot and then broadcasted to both the audience (specialized processors) and the management team behind the scenes (context formation and executive guidance processes). Thanks to the broadcast mechanism, the specialized processors “see” the action taking place under the focal point on stage. Depending on the information received by the processors and their potential contribution, these processors may form interim coalitions to build a new possible elaborated contribution to the next steps of the performance. Individual processors can also provide their processed outputs. All contributions from the audience compete for the appearance in the brightly lit area of the scene. The winning content that will show up as the main action of the play is finally shaped under the influence of active contexts and guidance from the director, script writer, scene designer, etc.

2.2 Proposed Computational Model

The GWT has inspired a number of MC models and implementations [10-12], including the one we have used for this research work: CERA-CRANIUM [13].

While GWT provides just a metaphorical description of how a blackboard system could operate within a mind, a computational model inspired on this theory has to go beyond the metaphor and describe the same processes but in terms of engineering design. This is to say, the computational model must provide a fine level of description which serves as a design guideline for a real implementation. Furthermore, in the particular case of this research work, which is focused in the generation of qualia, the computational design has to provide an actual mechanism for the identification and specification of the explicit contents (overt perception) of the artificial mind.

As pointed out in the introductory section, at this stage of the research we are not aiming at describing a mechanism for the production of phenomenal states, but a mechanism for the generation and specification of the contents of overt perception (see [14] for a thoughtful discussion about this characterization of artificial
Towards the Generation of Visual Qualia in Artificial Cognitive Architectures

qualia). In short, we characterize artificial qualia as the contents that appear under the spotlight in a GWT-inspired implementation. We also argue that the explicit content specification produced by our system can be compared to the experiential content reported by humans when confronted with the same stimuli. We think this exercise can be useful in two different but related ways: on one hand, the computational model can be improved in order to better match human qualia production; on the other hand, a significant success in the former task might help understanding how conscious contents are generated in humans.

2.2.1 Workspace Dynamics

In computational terms, we interpret GWT as follows: there is a single serial thread that corresponds to conscious perception and gives access to a very limited set of highly elaborated representations. These selected integrated contents, which here we characterize as artificial qualia [14], are formed thanks to the collaboration and competition of a vast number of processors that run concurrently. Working memory – or short term memory – is modeled as a specific type of blackboard system whose operation is modulated by control signals sent from a set of coordination processes (see Fig. 2).

GWT broadcast mechanism is modeled as messages being sent from the single serial thread to the rest of processes running in the system (specialized processors and coordination processes). Analogously, possible contributions from processors are defined as messages sent from the processors to the workspace (these messages are called submissions – see Fig. 2). Note that the broadcast message from the spotlight to the specialized processors is indeed a submission message to the workspace.

Each piece of content sent to the workspace using a submission message is temporarily stored and potentially accessible to any specialized processor. These pieces of content are generically called percepts, as they represent the interim and partial elaborations that can potentially be integrated in a more complex and integrated percept. Whenever a new percept enters the workspace, all relevant processors can be notified. Notification messages provide this functionality: they are used to send newly created percepts to a variable list of specialized processors. Selection of processors eligible to receive a notification message is based on current active context and the contents of the corresponding percept. Integrated percepts coming from a serial thread submission are sent to all available processors, i.e., broadcasted.

Active contexts are generated dynamically as a function of current explicit perception (the sequence of artificial qualia percepts) and active system’s goals. Contexts are defined as a set of criteria. For instance, a possible context based on the relative location criterion could be “right side”, meaning that percepts coming from the right field of view are more likely to be active, and therefore notified to specialized processors.
Fig. 2 CRANIUM Computational Model. Global workspace dynamics are modeled as an information processing system that takes a vast amount of raw sensory data as input and generates a serial and much lower bandwidth summary of integrated information (artificial qualia). Information filtering and integration is achieved by the distributed processing carried out by the specialized processors. The specific way these processors have access to the information is regulated by the application of contexts and other commands sent to the workspace from the context formation and coordination processes.

The generation and application of multimodal contexts is explained in detail elsewhere [15]. In short, active contexts help solving the problem of percept selection, i.e. directs the attention of the explicit or “conscious” perception. This means that in the competition for access to the serial thread, those percepts which better fit the active context are more likely to be selected.

Other mechanisms exist for modulating the workspace dynamics and allowing lower level feedback loops, like the modeling of unconscious reflexes (see [13]).

2.2.1 Software Architecture Design

Our computational model is based on the principles of Service-Oriented Architectures – SOA (see [16]). The SOA-based model permits us to define a mesh of services that can be executed independently and distributed across a number of computers. Each service is defined in terms of its associated functionality (service contract), and the whole model can be seen as a collection of services and relations between them. These relations include sending messages like the submission and notification messages introduced above. The definition of these messages is based on the Decentralized Software Services Protocol – DSSP [17], and the asynchronous input/output is managed using the Concurrency and Coordination Runtime – CCR [18]. Independent services have been defined for specialized pro-
cessors, the workspace, the spotlight serial thread, and coordination processes (see Fig. 3).

![Diagram of GWT model](image)

**Fig. 3** The GWT model is based on a service-oriented architecture that follows a REST (REpresentation State Transfer) style. Behavior of each service is defined as a service contract, which basically is a set of operations. The messages sent between services invoke these operations and carry partial or full service state (state representations are denoted using parentheses in the figure) using the DSSP protocol. Workspace service state corresponds to the working memory.

### 2.2.2 Unconscious competition and coalition

Competition and coalition mechanisms developed by specialized processors in the GTW are not directly represented in the proposed computational model, i.e. there are no direct communication between processors in relation with these mechanisms (indeed, there is no direct communication at all between specialized processors).

The synthesis of more complex and elaborated information out of incoming sensory data is based on the concept of percept. The main role of the system described in Fig. 2 is to dynamically generate integrated percepts by iteratively processing and combining more simple percepts. The content available in working memory, which is essentially represented in the form of single percepts and complex percepts, is both the input and the output of this iterative integration process.

In sum, the functionality associated with processors coalition is the collaboration of two or more specialized processor to generate a new more integrated percept out of a number of other existing percepts. This functionality is implemented in the model by means of percept reiterative generation using the workspace shared memory. Instead of sending partial elaborations directly from one processor to another, they are temporarily stored in the workspace. From there, they are also distributed to other processors, which thus participate in an ad hoc coalition as illustrated in Fig. 2. See [15, 19] for specific examples of multimodal percept generation in the domain of autonomous mobile robotics.

The competition between processors is also modeled by means of the dynamically generated percepts that are stored temporarily in the workspace. The aim of this competition process is to select the contents that will be “illuminated” by the spotlight; therefore this task can be carried out by the application of contexts. At any given time a specific context is active, inducing a bias in the workspace. The
application of a context in the workspace implies that only those percepts that match the criteria of the context are likely to be sent to the processors. It could be said that the competition takes place between the percepts, and they compete for being further processed and hopefully become a part of the finally selected complex percept (which will be sent to the spotlight service and also used as input for the definition of the next active context).

3 Minimal Implementation of CERA-CRANIUM Architecture

The CERA-CRANIUM cognitive architecture has been designed to serve as a MC research test bed where two main components can be distinguished: CERA is a cognitive architecture structured in layers, and CRANIUM is basically an implementation of the functional consciousness model described above. CERA uses the services provided by CRANIUM in order to control an autonomous robot. Although the present implementation covers more aspects, like component reusability across robotic platforms, here we will primarily focus on CRANIUM, describing just a minimal part of CERA. Analogously, although other specialized processors have been implemented for other modalities, at this time we will focus exclusively on visual processing. See [13] for a comprehensive description of the CERA-CRANIUM.

Current definition of CERA comprises four layers (see Fig. 4):
- Sensory-motor services layer encloses sensor and actuator drivers.
- Physical layer hosts a CRANIUM workspace and manages representations directly related with existing physical body of the robot.
- Mission-specific layer hosts a CRANIUM workspace and manages problem domain-dependent representations.
- Core layer regulates the operation of lower level workspaces and manages domain-independent representations.

CERA layers are also defined as services, managing the access to CRANIUM services (partly described in Fig. 3) and establishing a hierarchy that enables the use of several workspaces. Just one workspace located in the CERA physical layer has been used in this work for the sake of simplicity.

For preliminary experiments on the generation of artificial qualia a minimal configuration of the cognitive architecture has been used. Basically, no actuators are used, mission-specific layer is not used, and the core layer consists of a minimal implementation for context definition. The perceptual information flow is limited to visual modality.

Image bitmaps from the digital camera are acquired periodically thanks to a camera driver located in the CERA sensory-motor services layer (sensor services in Fig. 4). Proprioceptive sensing data is also acquired thanks to specific services located in the sensory-motor services layer. In the case of vision, relative location and current orientation of the camera is provided. All sensing data is sent to the
Towards the Generation of Visual Qualia in Artificial Cognitive Architectures

physical layer – note that CERA higher layers do not have direct access to raw sensory data. As soon as sensor data are received in the physical layer, first single percepts are created and submitted to the workspace service. Basically, these initial single percepts are data packages combining exteroceptive sensing data with the associated proprioceptive data.

For instance, one single percept might contain an image bitmap plus the camera orientation that was logged when that bitmap was acquired by the CCD sensor. Data structures enclosed in single percepts are used to implement a depictive representation of percepts in the sense described by Aleksander and Dunmall [20].

4 Experimental Settings and Methodology

A minimal architecture along the lines described above has been implemented for the generation of visual qualia. Although the system is designed to work both with real and simulated robots, a simulated environment has been set up initially. The simulated robot, designed using Robotics Developer Studio (RDS) [21], is a modified version of a MobileRobots Pioneer 3 DX8 robot in which actuators are disabled, and the only sensor available is a single simulated camera oriented to look over the robot’s front side (providing 320x200 pixels bitmaps). A specific sensor service for the camera has been written, so a predefined image sequence can be injected at will (instead of ingesting the synthetic visual feed from the simulator or the feed from a real camera connected to the system).

Only one meta-goal has been defined in the CERA core layer, which is based on detecting saliencies. Therefore, the operation of the workspace will be modulated by relative location contexts pointing to novelties. The kind of novelties being detected depends on the perception process; that is to say, the particular processors implemented in the system will determine the sort of novelties that can be potentially detected. Two types of processors have been used for the experiments described in this work: motion detectors and region of interest (ROI) detectors.
The following types of basic visual stimuli have been used for preliminary testing:
- \textit{S1}. Static white object in a dark background.
- \textit{S2}. White object moving along a rectilinear trajectory.
- \textit{S3}. Two stationary white blinking rounded spots.

\textit{S1} and \textit{S2} are generated using the RDS simulator, while \textit{S3} is generated using the image sequence injection mechanism. It consists of a sequence designed to induce apparent motion effect in humans [22]. Humans can consciously perceive motion not only from real moving objects, but also from series of images containing spatially segregated visual stimuli like \textit{S3} (see Fig. 6).

The ability to report qualia with grounded meaning is one of the key features of conscious machines [23]. In order to study artificial qualia, we have to analyze what kind of integrated content a system is able to report at any given level [14]. The system we have implemented for this work is not yet endowed with an accurate reporting system. However, we have devised an inner state inspection mechanism: the CERA viewer (see Fig. 5).

![Fig. 5](image)

**Fig. 5** Artificial qualia specification produced by CERA-CRANIUM can be compared with human conscious content report using the CERA viewer. A human observer can judge the similarity between his/her own experiential content and the representation displayed on screen.

Although the output of a viewer cannot be directly compared with human accurate verbal report, alternative strategies can be adopted in order to compare the contents of conscious perception in humans with the content specification of explicit perception in our proposed system. For instance, the same human observer can confirm whether or not the content of his/her visual experience matches the integrated percepts represented using CERA viewer.

As pointed out by the authors elsewhere [14], the proposed comparison scheme is just an initial step towards a more complete synthetic phenomenology approach. Three stages have been defined as partial but complementary definitions of artificial qualia. Current implementation addresses the so-called Stage 1 (perceptual content representation) and to some extent Stage 3 (self-modulation and reportability). It is our aim to work towards a full implementation that addresses the three defined steps, including Stage 2 (introspective perceptual meta-representation).

Preliminary experiments were conducted as follows: both a human subject without previous knowledge of the domain (\textit{H}) and the minimal CERA-CRANIUM implementation (\textit{MCC}) were exposed to \textit{S1}, \textit{S2}, and \textit{S3} visual stimuli. \textit{H} was asked to pay attention to white objects and verbally report the action per-
ceived when looking at the computer screen. *MCC* was given just one meta-goal: focus on saliencies; that is to say, core layer was programmed to generate spatial contexts pointing to either ROI or detected movement. CERA viewer was programmed to generate a camera field of view screen in which the complex percepts coming out the physical layer are represented. Only two specialized processors were activated: a specific ROI detector for white objects and a motion detector based on pixel changes. Consequently, current version of CERA viewer is only able to pinpoint the location of complex percepts corresponding to integrated ROIs (using red color pixels), and the direction of movement (using a black color mark indicating the direction of movement).

Working memory span (maximum age of percepts temporarily stored in the workspace and hence available to processors) was configured to 500 msec. *S3* white ball stimulus duration was 100 msec. and inter-stimulus interval (ISI) was 50 msec.

### 5 Preliminary Results

When exposed to *S1* stimulus, *H* reported a static white object resting on the ground, located near to the center of the screen. The output of CERA viewer when *MCC* was exposed to the same visual stimulus matched part of *H*’s report (Fig. 6a). Given the current implementation, all meaning that *MCC* can represent is exclusively about white objects and movement. Therefore no representation for such a concept as ground can appear in the viewer.

When exposed to *S2*, *H* reported a round object moving uniformly from the right to the left. *MCC* viewer representations again matched *H*’s report in part (Fig. 6b). Given that motion detector processor does not provide any measure related with speed, the uniform speed was not perceived by *MCC*.

As expected, when *H* was exposed to *S3* she reported a ball continuously moving back and forth from the left to the right and vice versa. However, *MCC* did not produce a matching motion representation (Fig. 6c). *MCC* viewer showed motion marks to the left and to the right as expected (this behavior ceases if working memory span is shorter than ISI), but no continuity existed during black ISI after every dot stimulus. In short, the generated representation did not correspond to the smooth visual experience reported by *H*. 
Fig. 6 S1 (a), S2 (b), and S3 (c) visual stimuli. From top to bottom: simulated scenario, robot camera image, and CERA viewer output. Corresponding H’s report is illustrated using speech bubbles. The white object in scenarios (a) and (b) is a simulated iRobot Create robot.

6 Conclusions and Future Work

The research line introduced in this paper is well underway. Preliminary results indicate that proposed GWT-based implementation needs to be enhanced in order to be able to specify accurately human-like visual experience. Future work also includes improving the CERA viewer interface so overt perception content specification can be interpreted easily by humans.

It is well known that human perception is dramatically affected by expectations. Therefore, next steps planned for the enhancement of the architecture, which we think will lead to better results, include the generation of expectation-based percepts. Our hypothesis is that the use of expectations will contribute to a more robust system when we progress on to testing with noisy real world images. Additionally, better results might be obtained in terms of reproducing some human optical illusions. If that is the case, it will help to demonstrate that the presence of perceptual illusions correlates with better perception accuracy in noisy environments, and therefore illusions could be considered a by-product of an outstanding perception systems selected by evolution. Whether or not robots with human-level visual recognition skills will inevitably experience similar optical illusions re-
Towards the Generation of Visual Qualia in Artificial Cognitive Architectures

mains to be seen. Once we have the expectation generation mechanism in place, we aim to test other typical visual phenomena like color-phi effect, attentional blink, flash-lag effect, etc. Additionally, we are working on improving the proposed reporting mechanism in order to effectively represent counterfactual stimuli providing an appropriate representation of the possible qualia space.

Acknowledgements

We wish to thank the anonymous reviewers for their useful and constructive criticism. This work was supported in part by the Spanish Ministry of Education under CICYT grant TRA2007-67374-C02-02.

References


