International Journal of Machine Consciousness
Vol. 1, No. 1 (2009) 53–70
© World Scientific Publishing Company

World Scientific
www.worldscientific.com

SYNTHETIC PHENOMENOLOGY

RON CHRISLEY

*Centre for Research in Cognitive Science and Department of Informatics,
University of Sussex, Falmer, BN1 9QH, United Kingdom
ronc@sussex.ac.uk*

The term "synthetic phenomenology" refers to: 1) any attempt to characterize the phenomenal states possessed, or modeled by, an artefact (such as a robot); or 2) any attempt to use an artefact to help specify phenomenal states (independently of whether such states are possessed by a naturally conscious being or an artefact). The notion of synthetic phenomenology is clarified, and distinguished from some related notions. It is argued that much work in machine consciousness would benefit from being more cognizant of the need for synthetic phenomenology of the first type, and of the possible forms it may take. It is then argued that synthetic phenomenology of the second type looks set to resolve some problems confronted by standard, non-synthetic attempts at characterizing phenomenal states. An example of the second form of synthetic phenomenology is given.

Keywords: Synthetic phenomenology; machine consciousness; content specification.

1. Machines and the Scientific Explanation of Consciousness

A machine need not be conscious for it to be a contribution to the field of machine consciousness. The field of machine consciousness, as a whole, aims not only to create artefacts that themselves have phenomenal states, but also to create artefacts that can help us understand or explain natural phenomenal states in humans and other animals, even if such artefacts are not themselves conscious.[13,30] An important goal of the field of machine consciousness, then, is to make substantial contributions to the science of consciousness.

Any science requires an ability to specify its *explananda* (facts, events, etc., to be explained) and its *explanantia* (states, facts, events, properties, laws, etc., that do the explaining). In a science of consciousness, particular conscious states (experiences) may be expected to play both of those roles. A science of consciousness, then, has a double need for a way to specify experiences. There is reason to believe that work in machine consciousness can provide and benefit from such experiential specification. In order to see how, some philosophical background must first be provided.

1.1. *Specifying the contents of consciousness*

What one takes to be involved in a precise specification of experiences will depend, in general, on one's theory of consciousness. In particular, theories vary concerning the extent to which consciousness is representational. Some theories take phenomenal states to be exhausted by their representational character,[32] some take them to have both representational and non-representational aspects,[26] while others (e.g., some sense-data theories) take phenomenal states to be wholly and essentially non-representational. Since the latter view is problematic and largely discredited, the approach of this article will be to assume that experiences are at least partly representational, and that one can characterize the representational aspects of an experience more or less independently of the non-representational aspects, if any.

It follows that at least part of what is essential to most, if not all, experiences is their representational *content*. Recent work in the general theory of representation takes foundational work in linguistic meaning as its starting point. We know from Frege that linguistic significance includes not just *reference* (what the representation is *of*), but also *sense* (the *way* that referent is represented). In general theories of representation, which aspire to give an account not just of linguistic meaning, but non-linguistic, mental representation, the notion of sense is generalized to representational content. Thus, the content of an experience is the way the experience presents the world as being. The question then becomes: "How can we specify the content of particular experiences?"

The standard way of specifying the content of mental states is by use of "that" clauses. For example, "Bob believes that snow is white" ascribes to Bob a belief, the content of which is the same as, and is therefore specified by, the phrase following the word "that": i.e., "snow is white". This means of specifying content is called linguistic expression because the content is specified not by finding a piece of language that refers to the content in question (as does the specification "the content of the experience that Bob had 2.5 minutes ago"), but rather by finding words that have or *express* that very content. Although linguistic expression works for specifying the content of linguistically and conceptually structured mental states (such as those involved in explicit reasoning, logical thought, etc.), there is reason to believe that some aspects of mentality (e.g., some aspects of perceptual experience) have content that is not conceptually structured.[10,16,17,27] Insofar as language carries only conceptual content, linguistic expression will not be able to specify the non-conceptual content of experience.[a] As has been pointed out before,[10] an alternative means of specification is needed.

[a] Actually, the situation may be worse than that. There is reason to believe that linguistic expression cannot even allow us to specify all the conceptual, linguistically-structured contents we might have need of. If externalists like Putnam[29] and Burge[5] are right in claiming that the content of linguistic expressions depends on the environment of a subject, then one will not, in general, be able to specify the content of the conceptual, linguistically-structured mental states of a subject whose environment is sufficiently different from one.

1.2. *Constraints on an alternative means of specification*

A successful alternative means of specifying the content of experience must meet several constraints:

- If it is to be of use in intersubjective science, then, like linguistic expression, it must be communicable (not private).
- If it is to permit the application of law-like generalizations, then unlike "the content of the experience that Bob had 2.5 minutes ago", it must specify the content of experience *canonically*, by virtue of the experience's essential, rather than accidental, properties[16] (or by reference to the essential properties of a non-phenomenally characterized system to which the experience is either identical or lawfully related; see Sec. 1.2.2).
- If it is to respect the Fregean insight sketched above, it cannot, in general, consist merely in a specification of what the experience is about; it must also specify the way that referent is being represented.
- If it is to be a genuine alternative to linguistic expression, then it must be capable, if not of specifying a superset of the contents specifiable by linguistic expression, then at least of specifying contents that are not specifiable by linguistic expression (e.g., non-conceptual experiential contents).[b]

Given the pervasiveness and easy, facile nature of linguistic expression, it can be difficult to imagine what an alternative means of content specification could be. But the need for an alternative to linguistic expression specifications has been acknowledged before. Peacocke offers scenarios, ways of filling out the space around a subject, as a means of specifying a kind of non-conceptual content: scenario content.[27] Bermudez carefully considers the problem of specifying what he calls "non-linguistic" content, and offers some strategies for doing so.[4] Other work, although perhaps not conceived of by its authors as potential solutions to this problem, can nevertheless be considered as such. In particular the work of Lehar,[22] which offers cartoons and sketches as a way of highlighting striking structural aspects of human experience, might be thought of as providing non-symbolic evocative specifications of experiential content (see Sec. 1.2.1). But there has been little systematic investigation into the possible forms that content specification might take. To be clear on this, it may be helpful to step back and consider the question: what is it to specify content, anyway? There seems to be three main ways how content specification is achieved: evocative, referential, and mixed[14,15] (but an earlier treatment[10] provides a different way of classifying the possibilities).

1.2.1. *Evocative specification*

Evocative modes of content specification aim to create in the recipient of the specification a mental state with the same content as the content to be specified. This can

[b] On this point, aficionados of non-conceptual content are referred to App. A.

be done symbolically, non-symbolically, or imperatively:

- **Symbolic:** Linguistic expression is a kind of symbolic evocative specification: by virtue of understanding the language used in the specification, the recipient enters a mental state with the very same content as the one to be specified. As argued above, this means of specification is too limited for the purposes of a science of consciousness.
- **Non-symbolic:** Specifications that cause the recipient to go into a mental state with the content to be specified by means other than by virtue of the recipient interpreting the specification symbolically/linguistically will be evocative, but non-symbolic specifications. For example, images can be used as a form of non-symbolic evocative specification of the content of visual experience (in which case they are referred to as *depictions*[14,15]): by viewing the depiction, the recipient has a visual experience the content of which is the same as that to be specified.
- **Imperative:** Another possibility is imperative evocative specification, in which one provides the recipient with a set of instructions that, if followed, would result in the recipient entering into a mental state with the content to be specified. Following the instructions themselves will typically involve making use of symbolic or non-symbolic evocative specifications. An example is "roll up a sheet of paper into a tube, hold it against your left hand with the palm facing you, and look through the tube with your right eye, keeping your left eye open".

  Following this "experience recipe" will produce in normal subjects a visual experience we might try to specify symbolically as "seeing a hole in one's hand". It can therefore count as a specification of such an experience (although it is unlikely to meet the "canonicalness" condition necessary for scientific enquiry, mentioned above). Given the imaginative power of the typical specification recipient, some imperative evocative specifications may not need to be actually carried out; rather, it may be sufficient for the recipient merely to imagine carrying out the actions in the specification for them to enter into a mental state with the desired content.

1.2.2. *Referential specification*

Evocative depictions can succeed independently of one's theory of the relation between non-experiential and experiential states, assuming one has such a theory at all. Referential specifications, on the other hand, instead specify the content of experience in a way that is not independent of one's understanding of the relation between the experiential and the non-experiential. Referential specifications do not specify the content of an experience by virtue of causing the recipient to have an experience with that content (although they may, at times, have such an effect). Rather, they aim to give the recipient discriminating knowledge of the content in question (Evans' "knowledge which", distinct from "knowledge that" and "knowledge how"[17]), knowledge that gives the recipient the ability to distinguish the specified experiential content from all other experiential contents. An example would be "the content of the experience realized by neural state N", where N is a canonical specification of a neural state. Clearly this only specifies a content relative to some theory relating neural and experiential states (since it assumes, *inter alia*, that neural states are sufficient for experiential contents). Because of this theory-mediation, referential specifications can leave the recipient with a feeling of epistemic dissatisfaction: "I know that the experience being specified is the experience realized by neural state $N214$, but which experience is that?"

1.2.3. *Mixed evocative/referential specification*

Mixed specifications aim to retain the best features of evocative and referential specifications while avoiding their respective limitations. For example, in the case of specifying the content of visual experience, depictions can assist the recipient in visualizing the structures employed in a referential specification, structures that according to the associated theory determine the experiential content of interest. Depictively presenting these structures *in the right way* may be crucial to the success of the specification. For example, if one assumes an expectation-based theory of perceptual experience[14,15] (in which the non-conceptual content of a visual experience consists in the set of sub-personal expectations that a subject's visual system has relative to a set of relevant possible actions, such as eye movements), then one could attempt to specify a particular visual experiential content by compiling a list of these sub-personal expectations: a list of various actions a subject having that experience might perform, and the expected visual stimulation (input) that would result. In some sense, reference to the correct content would have been secured, but not in a way that is of use to the recipient (the "epistemic dissatisfaction" mentioned at the end of the previous paragraph). If instead one arrayed designators of expected inputs spatially, where the location of a designator depended on the spatial properties of the potential action that generates the expectation (as done in Sec. 2.2.4), then the recipient may be much more likely to know which set of expectations, and thus which content, is being specified. (Compare giving someone a list of 1s and 0s corresponding to the binary contents of a jpeg file, as opposed to giving them the picture that file encodes.) If the theory being used is a correct account of human visual experience, the recipient will be employing the very same abilities and structures in perceiving the depictive component of the specification as are employed by the specification to (referentially) indicate the content of interest. Such mixed specifications that exploit a fit between the sensory-motor contingencies[25] that determine the content to be specified and the sensory-motor abilities employed by the recipient in perceiving the specification itself are called *enactive depictions.*[14,15]

Like referential specifications, enactive depictions are theory-mediated, and can thus be seen as a special case of that class. However, because they are associated with a particular kind of theory of experience — theories that take action-indexed expectations of sensory input to determine the content of experience — they have available to them a particular mode of conveying the abilities that determine a content, and therefore the content itself. Specifically, enactive depictions present these expectations in a spatially-indexed way, isomorphic to the spatial relations of their associated actions. Thus, any recipient of such a specification will themselves come to have a set of expectations that are isomorphic to the expectation set of a subject with the experiential content being specified. It follows that, according to the expectation-based theory of experience being assumed, the recipient of such a specification will have an experience with a content that is structurally isomorphic to the content being specified. Knowledge of this fact, and acquaintance with their own experiences, allows the recipients to "enact" the relevant content, and therefore to know which experience is being specified. In Sec. 2.2 it will be suggested that enactive depictions can be more interactive, generated on-the-fly in response to a recipient's probing by making use of an embodied, robotic system. This potentially permits specifications of content of substantially greater temporal and conceptual sophistication.

2. Synthetic Phenomenology

With the above background in place, we can at last begin a discussion of synthetic phenomenology itself. There are two types:

- **Type I synthetic phenomenology:** Any attempt to characterize the phenomenal states possessed, or modeled by, an artefact (such as a robot);
- **Type II synthetic phenomenology:** Any attempt to use an artefact to help specify phenomenal states (independently of whether such states are possessed by a naturally conscious being or an artefact).

Note that the Types are not exclusive; a Type II means of specification could be used to specify the experiences of an artificial agent, and thus be an instance of Type I as well. Type I synthetic phenomenology is by far the more common of the two Types.

There are some who use the term "synthetic phenomenology" in a way different to the above. See App. B for details.

2.1. *Type I synthetic phenomenology*

Involving as it does the notion of machine consciousness, Type I synthetic phenomenology invites the consideration of a range of thorny philosophical issues: What criteria must an artificial system meet in order for it to be correctly attributed conscious experiences at all? What criteria must an artificial system meet in order merely to model one particular conscious experience, as opposed to another? In what ways might it be easier or more difficult to ascertain which experiential state an artificial system is in, compared to doing so for a human? Along these lines, Gamez says: "It is impossible to describe the phenomenology of a system that is not *capable* of consciousness, and so the first challenge faced by synthetic phenomenology is to identify the systems that are capable of phenomenal states."[18] But it is not clear that this is so. For example, the original phenomenologists (e.g., Husserl) certainly thought it possible to distinguish the contents of consciousness in a precise manner without knowing the criteria that one's brain or body had to meet in order for one to be in a mental state with that content. True, Type I synthetic phenomenology, unlike traditional phenomenology, is not essentially a first-person enterprise, but the possibility of traditional phenomenology establishes that in principle, content specification can proceed in advance of an elaborated psychophysical theory. In fact, the points made at the beginning of Sec. 1 suggest that there is unlikely to be much progress in the development of a psychophysical theory, even for robots, without having some means of specifying the experiences in question. More likely, then, theory and means of specification will develop in parallel, each constraining the other; waiting for a completed psychophysical theory before embarking on Type I synthetic phenomenology seems counter-productive. Of course, such interleaving requires some idea, however inchoate, of which phenomenal states a system would have, were it to have any phenomenal states at all. Specifically, at least one form of synthetic phenomenology (cf. Sec. 1.2.3), involving as it does referential specifications of content, relies on a basic psychophysical theory; but it is a discriminative theory, not the constitutive one Gamez asks for (for more on the discriminative/constitutive distinction,[14,15] see Sec. 2.2.1).

For the case of the conceptual content of conscious experiences possessed or modeled by robots or other artificial systems, Type I synthetic phenomenology can proceed with methods already developed for conscious states in humans, such as linguistic expression. On the other hand, even if all complex conceptual contents are linguaform or propositional (a dubious claim), it remains possible that at least some of the conceptual primitives out of which such structures are made have a graded, metric, multi-dimensional character; hence the notion of *conceptual spaces*.[19] Thus, the geometric notational systems that have been developed in the theory of conceptual spaces may be extended quite straightforwardly, from the case of humans to that of robots, to specify experiences with this kind of conceptual content.[7,8]

In Sec. 1.1, the case for the existence of experiences that cannot be adequately captured by conceptual means (e.g., linguistic expression and conceptual space diagrams) relied, more or less, on reflection on our own, human case. But there is no reason to believe that such considerations do not also apply to animals, artificial consciousnesses (should they ever appear), or artefacts that are intended merely to model consciousness (which are here already). Therefore doing Type I synthetic phenomenology will, in general, require an alternative to linguistic expression, for the reasons given before.

Although the project of machine consciousness is relatively new, and use of the term "synthetic phenomenology" even more so, there have already been several instances of roboticists devising customized means for specifying the content of experience-like states modeled by their robots. A full treatment of Type I synthetic phenomenology would include a detailed analysis of these examples in the light of the preceding discussion, but there is no space for that in the present article. All that can be done here is to mention a selection of these cases with a brief comment, with the hope that the interested reader will consult the cited work for more detail.

- Mel's MURPHY[24] controlled a robot arm and a camera that looked down on it; a crucial part of MURPHY learning to use the arm involved off-line considerations of various joint angles and the expected camera input that would result from them. Mel used a depiction of these expectations as a way of specifying the content of these imaginative states.
- Aleksander and Dunmall's Neural Representational Modeller[2] displays the activities of simulated cells taken to be constitutive of visual awareness in a way that allows one to see the contents of modeled visual and imaginative experience changing over time.
- Stening, Jakobsson and Ziemke[31] go beyond straightforward depictions of imaginative states. Using a "conceptual inversion" method,[23] they create diagrams of experiential sequences of perceptually grounded concepts to better compare and contrast the way the world is experienced by the robot during imagination and real-time sensory-motor interaction. They use the term "synthetic phenomenology" to describe this work. Similar work was earlier reported by Holland and Goodman.[20]
- Holland's SIMNOS is a virtual reality system employed by the CRONOS robot to imagine, plan, etc.[21] Thus, standard virtual reality renderings of the state of SIMNOS serve as vivid and detailed depictions of the phenomenal states the CRONOS-plus-SIMNOS system models.

2.2. *Type II synthetic phenomenology*

The general upshot of the consideration, in Sec. 1.2, of alternatives to standard content specification (for humans or robots) was that a mixed evocative/referential specification seemed most promising. For the case of visual experiences, a kind of mixed specification, enactive depicition, was proffered, and it was suggested that the use of robots might itself facilitate such specifications; this proposal amounts to an instance of Type II synthetic phenomenology.

SEER-3 is a robotic system that has been designed to specify particular experiences via enactive depiction. This section explains how this kind of Type II synthetic phenomenology is achieved, in the following way. First, a simple discriminative theory of experience is assumed, without argument, for the purposes of illustration only. Then the SEER-3 robot itself is described, showing how, in the light of the assumed discriminative theory of experience, it can function as a model of experience. With these pieces in place, it is shown how SEER-3 can be used to dynamically generate enactive depictions to specify to a theorist the experiential states being modeled.

2.2.1. *A discriminative theory of experience*

The approach to specifying experience employed here relies on the use of a robotic system that models the experience of some hypothetical subject. It therefore requires a theory of consciousness that relates the states of the model to experiential states. This theory need not be a constitutive theory: it need not give necessary and sufficient conditions for a system to be an experiencer in general (that is, it need not solve the Hard Problem[6]). For present purposes, it can be assumed that the hypothetical subject is an experiencer, and that the robotic system models the aspects of that subject that are relevant to it being an experiencer. What is required is a discriminative theory of experience: a theory that determines, given the modeled facts on an occasion concerning a subject that is experiencing, exactly which experience the subject has on that occasion. These points can be made clear by considering the particular theory to be used in conjunction with SEER-3: the sensory-motor expectation theory of experience mentioned in Sec. 1.2.3. According to the simplest version of this theory (which is loosely inspired by, but not meant to be faithful to, the work of O'Regan and Noë[25]), (part of) the (non-conceptual) content of the visual experience of a subject at a time t is a spatially distributed conjunction of:

1. A subset of the sensory information being received at t ("foveal" input);
2. The foveal inputs the agent would (sub-personally) expect to have were it to perform at t, an action a, drawn from a privileged set of actions, spatially displaced from the location of the sensations in (1) in a way isomorphic to the spatial properties of the action a.

The theory implies that the content of the visual experience of a humanoid subject would be expected inputs in a two-dimensional spatial array, with the current foveal inputs at the center; the visual input that the subject would (sub-personally) expect to have, were it to look up and to the left, would be located up and to the left in the array; the visual input that the subject would (sub-personally) expect to have, were it to look to the right, would be located to the right in the array; and so on.[c]

The theory assumed here is simple and flawed. However, this does not count against it for present purposes; recall that this particular theory is put forward only because some theory or other is required in order to provide an example of enactive depiction specifications of experience.

2.2.2. *The SEER-3 robotic model*

That the purpose here is demonstration of a technique, rather than arguing for a particular theory or model, allows the normal modeling relation to be inverted. Rather than choosing a subject of experience and attempting to construct a robotic system that models it, we can start with a convenient robotic system, and assume

[c] Talk of an input s being spatially located at l is metaphorical shorthand; what is thereby referred to is a content c (an abstract object with no location) that presents the world as being s-like at l.

that there is some experiencing subject for which the robotic system is an adequate model. In the case of SEER-3, the robotic platform used is an off-the-shelf commercial zoomorphic robot (the Sony AIBO ERS-7), suitably programmed to implement a basic model of visual experience. The robot has a single, fixed-position video camera mounted at the tip of its nose, and in these demonstrations, only the head of the robot moves. Thus the head plays the role of a large, saccading eye in an otherwise stationary subject with monocular vision.

The hardware and software together comprise a model with the following components relevant to the present discussion:

- A visual processing component, that transforms the raw camera signal by introducing a blind spot, reducing acuity outside the foveal area, etc. In the current implementation, the non-foveal/foveal distinction consists in monochrome versus color input, and reduced versus full intensity.
- An expectation-maintenance mechanism, which uses currently received foveal inputs only (but see Sec. 2.2.3) to modify the robot's expectation of what sensations it would receive were it to make this or that head movement.

The expectation maintenance mechanism could be implemented in a simple recurrent neural network; work by the author on the CNM system[9] is an example of such an implementation that learns, and employs in planning, sensory expectations. A possible advantage of such an approach is that the automatic generalization features of neural networks would result in the extrapolation and interpolation of expectations to actions never before performed in the current context, resulting in an experienced visual field that spans the entire action space from the outset. The SEER-3 demonstrations reported here do not employ such an implementation for the expectation maintenance mechanism; rather, a kind of look-up array is used. The array, corresponding in extent to possible eye coordinates, is initially undefined, signifying the absence of any expectations. After performing a given action, such as the fixation of gaze x degrees to the right and y degrees up, the robot will modify its expectations for any action that would result in the robot receiving sensations from any point within the foveal radius r of (x, y). The expectations for any location within that circle will be changed to be whatever input it is now receiving at that location. This will, in general, alter at least some of the expectations for all changes of gaze fixation to any point less than $2r$ from the current point of fixation.

This approach, in conjunction with the expectation-based discriminative theory assumed in Sec. 2.2.1, results in the field of visual experience starting from an initial, foveated region, and expanding as more of the visual environment is explored. Although many would think this to be an unlikely feature of the visual experience of any actual organism, it is, by definition, a feature of the experience of the hypothetical subject the SEER-3 system models. Despite the developmental differences between the neural network and look-up table implementations, the steady-state extent of the modeled experiential visual field will be the same. Specifically, the modeled visual field will be a superset of the field of current visual sensation, delimited by where the eye can saccade to.

2.2.3. *Recent extensions to SEER-3: Time, affect and foveation*

Recent work has included modeling and depicting experiential contents with a endogenous dynamics. Specifically, the concept of the *intensity* of an expectation, which may be identified with subjective probability of outcome, confidence, and/or salience, depending on one's theory, has been introduced. It is then possible for the intensity of any given expectation to change over time; in particular, the intensity of expectations in SEER-3 are now initialized at some maximum at the moment of expectation creation (perception), monotonically decreasing over time (fading).

A simple way of modeling affect in SEER-3 is to think of the environment as supplying a negative reinforcement input of varying intensity. This input is distinct from non-affective inputs in that it is inherently undesirable: any action which the system expects to result in a negative reinforcement input of a certain intensity is *ipso facto*, and to that extent, less likely to be selected than actions without such an expectation (though this aversion may, in some systems, be overridden). SEER-3 can acquire such expectations by interacting with the world in the usual way. These expectations could fade in intensity in the manner described above. A similar extension could be made to include inherently positively reinforcing inputs. (The current SEER-3 configuration does not fully implement these affective features; the affective aspects of the depiction in Fig. 1 are mock-ups for illustrative purposes only.)

The original design of the expectation management system was as stated in Sec. 2.2.2: only current received foveal inputs could affect the expectational state of the system. SEER-3 has been extended so that (monochrome, reduced intensity) non-foveal inputs affect the expectational state in the same way foveal inputs do.

2.2.4. *Using SEER-3 to specify experiences*

According to the discriminative theory presented in Sec. 2.2.1, the set of visual expectations an agent has plus its current visual input determines the content of its visual experience. It follows that a robot can model the having of a visual experience by having analogous expectations and input. Independently of whether the theory or model is correct, we can conclude, as per the discussion in Secs. 1.2.2 and 1.2.3, that conveying the robot's expectations and its current visual input in the right way (although there may be more than one right way, even for a single recipient) can serve as a specification of the content of the modeled experience. The present hypothesis is that one right way of doing this is via enactive depictions (cf. Sec. 1.2.3). In the case of SEER-3, such depictions are constructed as follows. At each time step, for each expectation that the robot has at that time, depict the expectation as follows:

- **Visual experience:** The expectation to be depicted is that the robot would receive, at the point of fixation, a sensation of hue h were it to fix its gaze y degrees up from and x degrees to the right of the origin of the axes of head movement.

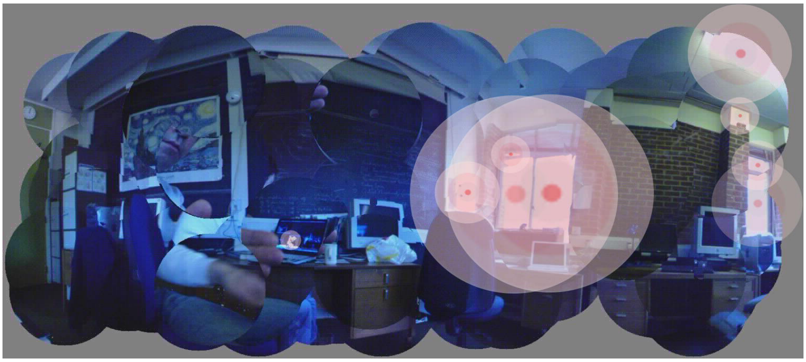

Fig. 1. An enactive depiction of the content of a visual experience modeled by the SEER-3 robot.

Suppose also that this expectation has intensity i. A depiction of this expectation would consist in placing a mark of hue h with intensity i at the location y degrees up from and x degrees to the right of the center of the depictive frame.

- **Affective experience:** The expectation to be depicted is that the robot would receive a negative reinforcement input of intensity i were it to fix its gaze y degrees up from and x degrees to the right of the origin of the axes of head movement. (To make sense of this, imagine that the robot is modeling an organism that finds fixating on excessively bright light sources painful.) A depiction of this expectation would consist in placing a mark of some distinctive hue (in the case of Fig. 1, pink) with intensity i at the location y degrees up from and x degrees to the right of the center of the depictive frame.

An example of a SEER-3 depiction produced by this process appears in Fig. 1.[d] As per the model described in Sec. 2.2.2, the pixel color value at any point indicates a set of expectations based on recent previous experiences. For example, a blue pixel value at a location (e.g., the back of the chair) indicates not only an expectation to receive a blue input at the point of fixation if the robot were to fixate on the location corresponding to the back of the chair, but also an expectation to receive a blue input down and to the left of the point of fixation if the robot were move its gaze to the location corresponding to the window. Similarly, a pink pixel value at a location (e.g., the window) indicates an expectation to receive a negatively reinforcing input ("pain") if the robot were to look in the direction of the window. (This creates a depictive ambiguity for pink, but this is merely a consequence of displaying all kinds of content in one depictive frame; instead, one could have distinct layers of depiction for each kind of content.)

[d] A full-color version can be found at http://www.cogs.susx.ac.uk/users/ronc/depiction.jpg; a movie that depicts the dynamic, fading aspect of the visual experiential content modeled by SEER-3 can be found at http://www.cogs.susx.ac.uk/users/ronc/fading.mov.

In the depiction in Fig. 1, there is an exception to the interpretation of color just mentioned: the absence of any expectations for a location is indicated in grey. A grey pixel is therefore ambiguous: it could mean either an expectation to receive grey input, or the lack of any expectation at all for that location. Context should enable the theorist to resolve any ambiguity.

2.2.5. *SEER-3: Discussion*

The image in Fig. 1, then, is an enactive depiction, in the sense of Sec. 1.2.3. It should be clear that it is not a simple, non-symbolic evocative specification (cf. Sec. 1.2.1), in that it is also referential (cf. Sec. 1.2.2), and thus relies on the recipient knowing a theory, in this case, the discriminative, expectation-based theory being employed in SEER-3 (cf. Sec. 2.2.1). Because the depictions in Fig. 1 present these expectations in a spatially-indexed way, isomorphic to the spatial relations of their associated actions, viewers of Fig. 1 will, according to the theory being assumed, themselves come to have a set of expectations that are isomorphic to the expectation set of the subject with the experiential content being specified. It follows that viewers will have an experience with a content that is structurally isomorphic to the content being specified. Knowledge of this fact, and acquaintance with their own experiences, allows viewers to enact the relevant content, and therefore to know which experience is being specified (cf. Sec. 1.2.2).

Of course, the image in Fig. 1 is only a snapshot of a SEER-3 enactive depiction, which is temporally extended, dynamically generated, constantly updated. This in itself makes the robotic aspect of specification system indispensable. Proposed extensions to SEER-3 allowing interactive probing by the recipient (cf. Sec. 2.2.6) will serve to further exploit the robotic implementation, thus making SEER-3 an even clearer example of Type II synthetic phenomenology.

SEER-3 specifications of experiential content meet the desiderata outlined in Sec. 1.2:

- They are communicable.
- They specify the content of experience *canonically*, by reference to the essential properties of a non-phenomenally characterized system (expectational state) to which the experience is lawfully related, according to the theory assumed in Sec. 2.2.1.
- They specify the content of experiences, not (just) what the experiences are about.
- They are capable of specifying contents not specifiable by conceptual means, such as linguistic expression. A striking feature of SEER-3 depictions, like that in Fig. 1, is their fragmented nature: contours and object boundaries are not respected. This reveals the modeled experience to be non-conceptual, in that the relevant expectations are created and maintained in a way that is not guided by the concepts, e.g., **straight line**, **chair**, **hand**, etc.

2.2.6. *SEER-3: Future work*

One can distinguish the temporal aspects of the depiction from the temporal aspects of the content. On an expectation-based view of content, there are two temporal dimensions to content, corresponding to the two temporal dimensions of expectations: the time of the having of the expectation, and the time that the expectation is about.[12] Both of these dimensions might be accommodated by using movies, rather than static depictions such as the one in Fig. 1, thus aligning the temporality of the depiction and the content depicted. But the temporality of the depiction can be used to capture non-temporal dimensions of content as well. In a way, this is already the case with the depiction in Fig. 1; in order to know which content is being specified, the recipient must spend time scanning the image, saccading here and there over it. In general, the use of immersive virtual reality techniques might be an effective way to exploit the temporality of the recipient's experience to handle the multiple dimensions of experiential content.[10] It should be made clear that such a virtual system would not replace the robotic system, but rather provide the creators and recipients of specifications the ability to selectively highlight, probe and explore the expectational state.

This would not only enable the specification of experiences of substantially greater temporal and conceptual sophistication, but also pre-objective experiences which have until now eluded precise specification. For example, consider the visual experience of a young infant looking at a mug. Surely their experience shares many commonalities with the one an adult would have in that situation: the basic color and perspectival shape of the mug, etc. But the adult experiences the mug as an object, as something with an unseen back that is probably of the same color as the visible front, etc. Such differences of experience cannot easily be represented in a depiction, like that in Fig. 1, if at all. But in the case of an interactive virtual reality system, the crucial differences between the infant's and adult's expectations could be made manifest. In the case of an extended SEER-3 specification of the adult's experience, if the specification recipient were to move the point of view to a location behind the mug, and turn around to face the other way (resulting in a view depicting the adult's expectations of what would be seen if a similar move were actually made in the real world), the back of the mug, with appropriate color, would be displayed. In the case of a specification of the infant's experience, no such mug back would be seen, thus revealing differences in the two experiences that could not be depicted with a single snapshot, but could only be enacted by allowing the recipient to actively explore the expectational space of the subject.

The power of SEER-3 specifications would be greatly increased if its ability to specify non-conceptual experiences could be augmented with an ability to specify more conceptualized contents. The possibility of such unification will depend crucially on one's theory of concepts. If, to have an experience involving application of the concept **chair** is just to have a certain set of low-level expectations (e.g., ones that, unlike those depicted in Fig. 1, respect the boundaries of chairs), then specifications of conceptual content might be continuous with the specifications provided here. But if conceptual content involves a different kind of expectation altogether (e.g., the expectation that one will see a chair, rather than the expectation that one will receive this or that low-level sensory input), then a distinct approach might be needed. In such a case, the use of distinct depictive layers, touched on briefly in Sec. 2.2.4, might be employed.

Acknowledgments

Thanks to Mike Beaton, Rob Clowes, Joel Parthemore, Anil Seth and the members of the PAICS Research Group in the Department of Informatics at the Univeristy of Sussex for valuable comments and suggestions. Thanks also to Joel Parthemore for the implementing the robot controller and depiction generation software for SEER-3.

Appendix A. Notes on Non-Conceptual Content

The mere possibility of a "superset" means of specification that can specify both conceptual and non-conceptual contents is a good enough reason to reject a common, but confusing and problematic, definition of non-conceptual content; e.g., the first line of the entry "Nonconceptual Mental Content" in the *Stanford Encyclopedia of Philosophy*:

> "*The central idea behind the theory of nonconceptual mental content is that some mental states can represent the world even though the bearer of those mental states need not possess the concepts required to specify their content.*"[3]

If a "superset" means of specification exists, then it is possible that for most, if not all contents, including those that we would intuitively take to be conceptual, there will be a way to specify them such that the subject need not possess the concepts used in that specification. This would presumably render them non-conceptual according to the standard definition. At the very least it would render a content's conceptual status as relative to a means of specification. One could try to patch up this problem by defining non-conceptual content as "any content c such that there is no way to specify c in terms of concepts that the subject must possess in order to have mental states with content c". But other problems with this specification-relative way of characterizing non-conceptual content persist, even on the patched version. For example, by defining non-conceptual content in terms of the concepts required to specify or possess a content, the standard construal of non-conceptual content presupposes, rather than explains, what the conceptual/non-conceptual distinction is. Once this dependency on *a prior* notion of concept is acknowledged, the problems of specification-relativity can be removed by defining non-conceptual content to be, simply, "content that is not entirely composed of concepts", and providing a positive account of what it is for a content constituent to be a concept.[11] Peacocke has drawn similar conclusions.[28]

Appendix B. Terminological Notes

The first known use of the term "synthetic phenomenology" in a sense anywhere near the one employed here was by Scott Jordan in October 1998, in a talk given at the Max Planck Institute for Psychological Research in Munich, entitled "Synthetic phenomenology? Perhaps, but not via information processing".

There is a common, non-technical use of the word "phenomenology" that does not refer to the precise specification of the content of phenomenal experiences, but to the experiences themselves. Correspondingly, there have already been a few uses of the term "synthetic phenomenology" intended to mean neither Type I nor Type II specifications of experiences, but the experiences of artificial agents themselves. I think this confusion should be resisted, and the term "synthetic phenomenology" used as done so in this paper. The experiences of the artificial agents themselves could instead be referred to as "synthetic *phenomenality*",[14,15] or just "artificial consciousness".

Gamez, who has made some notable contributions to understanding synthetic phenomenology, often uses the term "synthetic phenomenology" as it is used in this paper, e.g., when he says: "[T]he synthetic phenomenological project is the description of machine consciousness."[18] But sometimes, Gamez uses "synthetic phenomenology" to mean "synthetic phenomenality"; for example, consider the use of the word "phenomenology" in this passage from the same article:

> "*Within the machine consciousness community*, '*synthetic phenomenology' is now more generally used to refer to the determination whether artificial systems are capable of conscious states and the description of their phenomenology when and if this occurs*, *and it is in this sense that I will be using it here*."[18]

Such usage in the context of a sentence attempting to define the term "synthetic phenomenology" can only lead to confusion. Also, as argued in Sec. 2.1, the former activity, of determining whether artificial systems are capable of conscious states, is best seen as a distinct enterprise not falling under the term "synthetic phenomenology".

At first glance, the following passage from Aleksander and Morton suggests that they, too, are using "phenomenological" to mean "phenomenal"; "To be synthetically phenomenological, a system S must contain machinery that represents what the world and the system S within it seem like, from the point of view of S."[1] But on more careful inspection, they are reserving the term "phenomenological" for systems that are not just capable of being in phenomenal states, but that are also capable of *representing* what it is like to be in such states. A more charitable interpretation, then, would be that they are indeed using "synthetic phenomenology" in the sense advocated here, and not in the sense of "synthetic phenomenality". (Earlier work has clarified the differences between the present use of the term "depiction" and the use of the same term by Aleksander and his colleagues.[14,15])

References

1. I. Aleksander and H. Morton, Why axiomatic models of being conscious? *J. Consci. Stud.* **14** (2007) 15–27.
2. I. Aleksander, H. Morton and B. Dunmall, Seeing is believing: Depictive neuromodeling of visual awareness, *IWANN'01: Proc. 6th Int. Work-Conf. on Artificial and Natural Neural Networks* (Springer-Verlag, London, UK, 2001), pp. 765–771.
3. J. L. Bermúdez, Non-conceptual mental content, `http://plato.stanford.edu/entries/content-nonconceptual/index.html`
4. J. L. Bermúdez, *Thinking Without Words* (Oxford University Press, Oxford, 2003).
5. T. Burge, Individualism and the mental, *Studies in Metaphysics*, Vol. 4, 1979.
6. D. Chalmers, *The Conscious Mind: In Search of a Fundamental Theory* (Oxford University Press, Oxford, 1996).
7. A. Chella, M. Frixione and S. Gaglio, Conceptual spaces for computer vision representations, *Artif. Intell. Rev.* **16** (2001) 137–152.
8. A. Chella, M. Frixione and S. Gaglio, A cognitive architecture for robot self-consciousness, *Artif. Intell. Med.* **44** (2008) 147–154.
9. R. Chrisley, Cognitive map construction and use: A parallel distributed processing approach, in *Connectionist Models: Proc. 1990 Connectionist Models Summer School.* eds. D. Touretzky, J. Elman, T. Sejnowski and G. Hinton (Morgan Kaufmann, San Mateo, 1990).
10. R. Chrisley, Taking embodiment seriously: Non-conceptual content and robotics, in *Android Epistemology*, eds. K. Ford, C. Glymour and P. Hayes (MIT Press, Cambridge, 1995), pp. 141–166.
11. R. Chrisley, *Non-Conceptual Psychological Explanation: Content & Computation*, Phd thesis, University of Oxford, 1996.
12. R. Chrisley, Some foundational issues concerning anticipatory systems, *Int. J. Comput. Anticipatory Syst.* **11** (2002) 3–15.
13. R. Chrisley, Philosophical foundations of artificial consciousness, *Artif. Intell. Med.* **44** (2008) 119–137.
14. R. Chrisley and J. Parthemore, Robotic specification of the non-conceptual content of visual experience, in *Proc. AAAI Fall Symp. on "Consciousness and Artificial Intelligence: Theoretical Foundations and Current Approaches"*, eds. A. Chella and R. Manzotti (AAAI Press, 2007), pp. 36–42.
15. R. Chrisley and J. Parthemore, Synthetic phenomenology: Exploiting embodiment to specify the non-conceptual content of visual experience, *J. Consci. Stud.* **14** (2007) 44–58.
16. A. Cussins, The connectionist construction of concepts, in *The Philosophy of Artificial Intelligence*, ed. M. Boden (Oxford University Press, Oxford, 1990), pp. 368–440.
17. G. Evans, *The Varieties of Reference* (Oxford University Press, Oxford, 1982).
18. D. Gamez, Progress in machine consciousness, *Consciousness and Cognition* **17** (2008) 887–910.
19. P. Gärdenfors, *Conceptual Spaces: The Geometry of Thought* (MIT Press, Cambridge, 2000).
20. O. Holland and R. Goodman, Robots with internal models: A route to machine consciousness? in *Machine Consciousness*, ed. O. Holland (Imprint Academic, Exeter, 2003).
21. O. Holland, R. Knight and R. Newcombe, A robot-based approach to machine consciousness, in *Artificial Consciousness*, eds. A. Chella and R. Manzotti (Imprint Academic, Exeter, 2007).
22. S. Lehar, A cartoon epistemology, `http://cns-alumni.bu.edu/slehar/cartoonepist/ cartoonepist.html`

23. F. Linåker and L. Niklasson, Extraction and inversion of abstract sensory flow representations, in *Proc. 6th Int. Conf. Simulation of Adaptive Behavior: From Animals to Animats 6*, eds. J. Meyer, A. Bethoz, D. Floreano, H. Roitblat and S. W. Wilson (MIT Press, Cambridge, 2000), pp. 199−208.
24. B. W. Mel, *MURPHY: A Neurally-Inspired Connectionist Approach to Learning and Performance in Vision-Based Robot Motion Planning*, PhD thesis, Urbana, IL, USA, 1989.
25. K. O'Regan and A. Noë, A sensorimotor account of vision and visual consciousness, *Behavioral and Brain Sciences* **24** (2002) 939−973.
26. C. Peacocke, *Sense and Content: Experience, Thought and their Relations* (Clarendon Press, Oxford, 1983).
27. C. Peacocke, *A Study of Concepts* (MIT Press, Cambridge, 1992).
28. C. Peacocke, Non-conceptual content: Kinds, rationales, and relations, *Mind and Language* **9** (1994) 419−429.
29. H. Putnam, *Representation and Reality* (MIT Press, Cambridge, 1988).
30. A. Seth, The strength of weak artificial consciousness, *Int. J. Mach. Consci.* **1** (2009) 71−82.
31. J. Stening, H. Jacobsson and T. Ziemke, Imagination abstraction of sensorimotor flow: Towards a robot model, in *Proc. AISB'05 Symp. on Next Generation Approaches to Machine Consciousness, Hatfield, UK*, eds. R. Chrisley, R. Clowes and S. Torrance, 2005.
32. M. Tye, *Ten Problems of Consciousness: A Representational Theory of the Phenomenal Mind* (MIT Press, Cambridge, 1995).