MODELS OF CONSCIOUSNESS 2

A CONFERENCE SERIES ON FORMAL APPROACHES TO THE MIND-MATTER RELATION

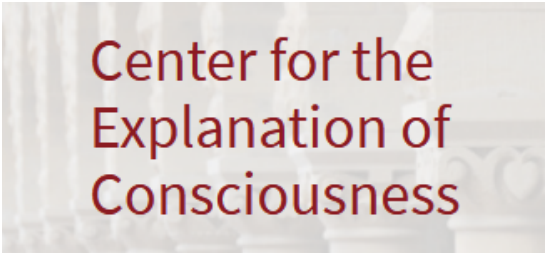
ONLINE
SEPTEMBER 27-30, 2021

Association for Mathematical Consciousness Science

ORGANIZING AND SUPPORTING INSTITUTIONS

OMCAN
Oxford Mathematics of Consciousness and Applications Network

mc/Mp
Munich Center for Mathematical Philosophy



CENTER FOR THE **FUTURE MIND**

CONTENTS

WELCOME

We are delighted to bring you Models of Consciousness 2 (MoC2) which follows the highly successful MoC1 conference held at the University of Oxford in September 2019. When planning began for the second Models of Consciousness conference (MoC2), Covid-19 was unknown. Its rapid spread in the spring of 2020 quickly made any hope of an in-person conference that year impossible. Once vaccines became widely available and travel restrictions began to ease this past spring, our hope was to be able to hold MoC2 in-person this year. But venue difficulties and the spread of the Delta variant forced us to put our next in-person meeting off until 2022 (see details on MoC3 on page 11). It was our wish, however, to not leave three years between meetings, particularly given the rapid expansion of our field. We are therefore excited to host this online version of MoC2. We are also very excited that MoC2 represents the first conference of the newly formed Association for Mathematical Consciousness Science (AMCS).

The AMCS is an international association of scientists and philosophers devoted to exploring the application of rigorous mathematical methods to the scientific study of consciousness. Its stated aim, which is the further development of such methods, recognizes the fact that there is a certain point at which a scientific field of study reaches the stage at which rigorous, formal approaches become necessary. With the number of conferences, workshops, online seminar series, and journal special issues devoted to these methods growing, there has also become a need to bring these efforts together to create a community. To that end, AMCS exists to provide opportunities for researchers working in this field to connect with one another.

From all of the conference organisers and advisors, welcome to MoC2!

CONFERENCE TEAM

ORGANISERS

Johannes Kleiner
LMU Munich

Jonathan Mason
University of Oxford

Wanja Wiese
Ruhr University Bochum

Robert Prentner
LMU Munich & FAU

Robin Lorenz
Cambridge Quantum Computing

ADVISORY BOARD

Paul Skokowski
Stanford University & University of Oxford

Ian Durham
Saint Anselm College

Kobi Kremnizer
University of Oxford

EVENT INFORMATION

ACCESSING THE CONFERENCE

To make sure you have everything you need whenever you need it, we have collected all access codes and links to the conference at the following webpage:

https://amcs-community.org/event-information/
Password: amcs-moc2

In case there are any important technical announcements, e.g. if links have to be changed, you will also find them at that page. So if anything does not work on your end, we recommend checking there.

GATHER.TOWN

We care deeply about making this conference as personal as possible. In order to give you a chance to meet others, and in order to facilitate lively discussions, we are making use of a new platform called Gather.Town.

You can find the link to our conference space above. If you have never used Gather.Town before, please watch this short 5-minute introductory video that explains all you need to know to have an enjoyable experience at MoC2:

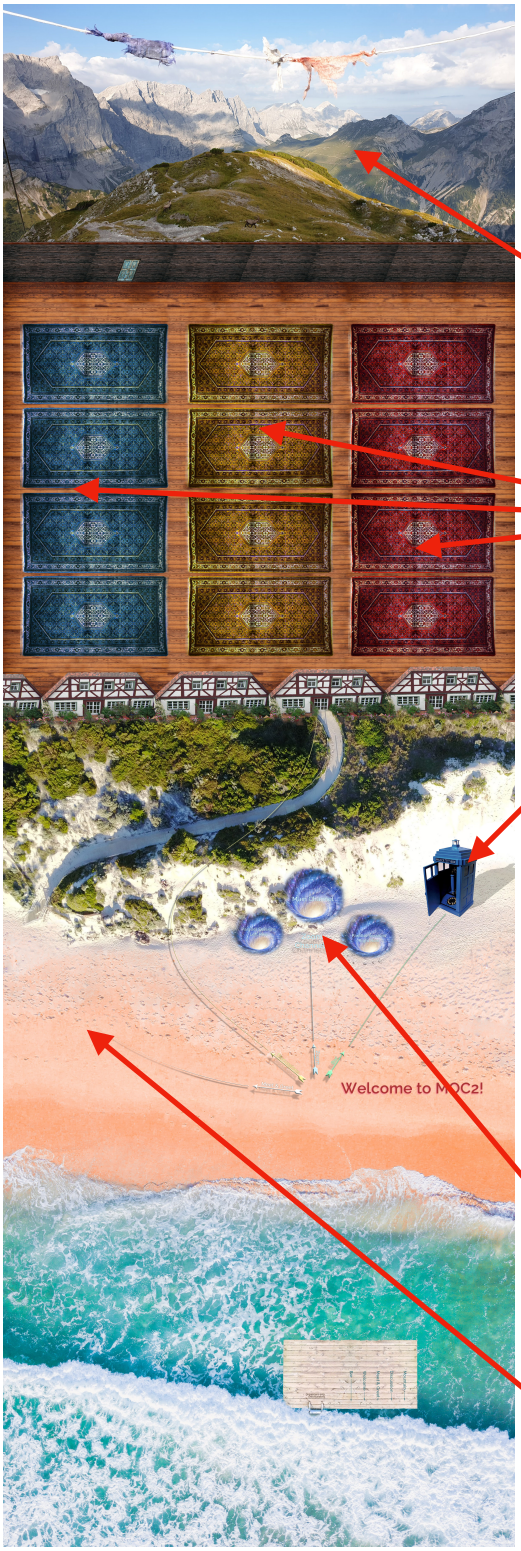
https://www.youtube.com/watch?v=89at5EvCEvk

When joining for the first time, you will first see a short tutorial before being connected to the conference. Please use your full name to sign in to the conference.

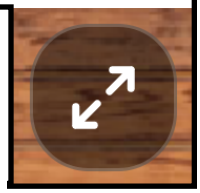
Currently, Gather.Town is optimised for the Chrome and Firefox browsers. Other browsers may not be supported in certain spaces.

A link and map of the MoC2 Gather.Town space appears on the next page.

GATHER.TOWN MAP, LINK, AND HINTS



This little button is very helpful. It's just to the right of the video streams at the top of your window, and easily missed. Click it to enlarge the video feeds, which greatly helps in discussions.



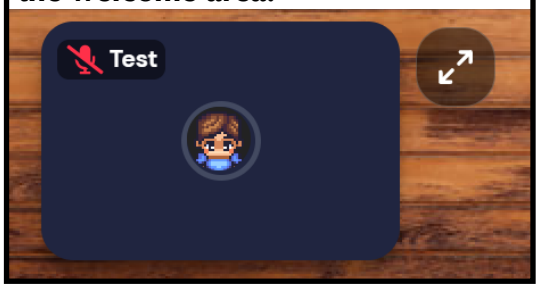
Virtual mountain view

Discussion session rooms

Portal to virtual bar

An easy way to navigate the map is to simply double-click on the place you wish to go. Your avatar will then move itself to that spot.

If you get lost, click on your name at the bottom of the screen and choose "Respawn". You'll land in the welcome area.



Portal to Zoom link

Meet & Greet Area
(includes virtual whiteboards)

https://gather.town/app/fvZo7jDZRTVKhgbJ/moc2amcs

DISCUSSION SESSIONS

We strongly believe that conferences and workshops which bring together many brilliant minds should not function like a diode when it comes to the flow of information. This is why we have experimented with alternative formats for many years.

One of the most successful formats we have tried so far are discussion sessions in small, parallel groups. In this conference, we will try this for the first time in a large online format.

To this end, we will make use of the Gather Town platform (cf. above), as well as a tool that allows you to propose questions that could be discussed, and vote on which questions you think are more important. You can find the links to both of these tools on our event-information website linked to above.

We will provide you with everything else that you need to know for participating at the discussions during the conference.

In advance, we would only like to share the following statements with you, which motivate us in each and every discussion we organize:

First and foremost, there must be ease, relaxation, and a general sense of permissiveness. The world in general disapproves of creativity, and to be creative in public is particularly bad. Even to speculate in public is rather worrisome. The individuals must, therefore, have the feeling that the others won't object. (...) It seems necessary to me, then, that all people at a session be willing to sound foolish and listen to others sound foolish.

Isaac Asimov, "How do people get new ideas", 1959

At the center of a discussion with Werner Heisenberg was "the shared problem and the desire to grasp and clarify it. One carefully approached it, passed it to the other, like in a friendly table tennis game, where both made sure that the ball remained in play. All the attention was focused on truly understanding the other and to avoid letting him stumble sophistically over his poor and inadequate expression. One could stutter, one could speak vaguely, even incomprehensibly, and he would guess what one actually wanted to say, would repeat it in his own different words, so that one could often exclaim with pleasure: 'Yes, exactly that...!'. During such an (...) intense exchange of thoughts, the ideas and concepts sharpened, so that their contours became recognizable more clearly."

A former collaborator of W. Heisenberg

PROPOSING AND VOTING FOR DISCUSSION TOPICS

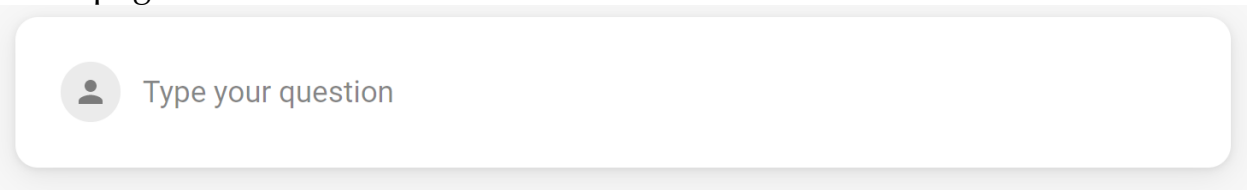
Discussion sessions are about the questions and topics you personally find important. This is why it is important to us that you can vote and propose topics for discussion.

In order to make this possible, we're working with a platform called sli.do. To find our event, simply go to https://www.sli.do/ and enter the code **#moc2amcs**. Alternatively, you can click on the link provided on our Event Information page.

Once you opened our event on sli.do, you can vote on questions by pressing this button at the top right of each question:

The number shows how many people have voted on that question already. If there are many votes, we know that many people think this question is good and we'll try to schedule it for a session.

To propose a topic for discussion, simply type it into this field on the very same page:



It's good to write concise proposals so that people can easily understand it.

We look forward to your questions!

MODERATING DISCUSSION SESSIONS

Having an inspiring discussion is a hard job. In our experience, it can be made a lot easier by having someone moderate the discussion session, and we would cordially invite you to moderate sessions as well. If you do, here is some general advice that we hope could prove useful.

We are completely aware that too much "structuring" can also negatively influence a creative process, so please don't feel "parented" by the below suggestions but rather see them as just that - advice.

We found that it can be very helpful to **clarify at the start** the directions and (if possible) the goal of the discussion. Topic descriptions are short and people might have different ideas in mind about what it is that is being discussed. So:

- We recommend taking 10 minutes to just get everyone's input on what the focus of the discussion should be, or what the most important aspect of the question is. That means 1 minute per person, so as a moderator, please emphasize that everyone should be brief.
- Nobody should hold back his/her ideas and intentions at this point, and everybody should be listened to. Try not to "criticize" at this stage.
- Try to make sure people don't go into details yet, but first gather all ideas. Refer to us organizers if you need someone to blame for cutting people short ;-)

If different directions emerge, don't be afraid to split your group (people can easily move to other tables in Gather.Town.) It would of course be nice if you joined together again at the end to gather your individual outcomes.

After the initial phase it's up to you and the group to see where the interest flows. As a moderator, we recommend you try to make sure that those who are less extrovert have a chance to express their thoughts. But you will see that everything happens quite naturally. During the main phase of the discussion we found these things to be helpful:

- Try to focus on basic ideas and principles as well as fundamental questions.
- Going into technical details usually takes a lot of time which you usually will not have at this point. The "experts" among you can always get together at a later point during the workshop.
- The discussion sessions are not there for you to learn a topic, but rather to discuss deep issues and to develop new ideas.

Finally, before the end of the discussion, we recommend you to call for a closing round in which every participant can summarize what was essential for them. This is to the benefit of all, it is sometimes very surprising how different the perspectives are that emerge from one and the same conversation. Also, this helps the minute-taker (cf. below) to do his/her job easily.

Please do encourage people to hang out and continue the discussion if they like, even after the official end of the slot. In most cases, we've managed to schedule a break after the discussion slot.

MINUTE-TAKING FOR DISCUSSION SESSIONS

There will be many amazing discussions during that conference, and chances are high that you will be part of quite a number of them.

To preserve some of what has been said, we would like to ask every group to determine one participant who briefly writes up the main points that came up in the discussion and pastes them into our shared folder. Details of how to access this folder is given in the online information page for the conference as mentioned earlier in this booklet. This allows everyone to see what else has been discussed for a given topic, and crucially is some reference for you to come back to in the future in case you're thinking about this discussion.

Crucially, the write-up **does not need to be long** and **does not need to be polished**. **Anything goes really**. A few sentences are enough, and in whatever language or form seems appropriate. No names are required (unless you want to).

AMCS WIKI & MoC3

AMCS WIKI PROJECT

The AMCS has a wiki project and the wiki will include an account of the conference. Selected details taken during the discussion sessions will also be included. Therefore, where discussions produce materials that will clearly be of interest to a wide audience, these ideas will be recorded on the wiki. If you would like to help work on the wiki after the conference then please add a note with your details to one of the discussion documents in the shared folder for the conference discussions as mentioned above.

MODELS OF CONSCIOUSNESS 3

Having been invited by the Center for the Explanation of Consciousness at Stanford University, MoC3 will take place in-person at Stanford's beautiful Alumni Center from 5-9 September, 2022. MoC3 will consist of a balanced mix of invited talks, contributed talks, poster sessions, and discussion sessions. Details will be posted on the AMCS website in the coming months and will also be circulated via e-mail.

STANFORD UNIVERSITY ALUMNI CENTER

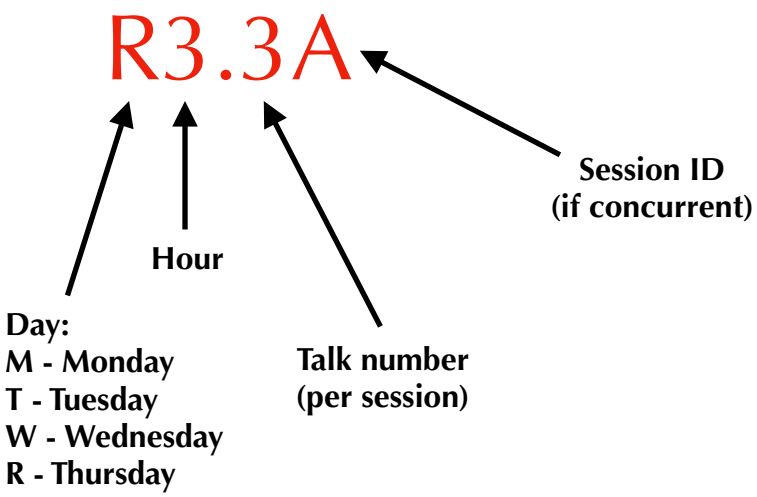
Time (UTC)	Monday, September 27		Tuesday, September 28	
1:45 PM - 2 PM	Opening address (President Lenore Blum + Organizers)			
Hour 1 2 PM - 3 PM	**Karl Friston** *Deep inference* (M1.1)		**Lenore Blum** *A Theoretical Computer Science Perspective on Consciousness* (T1.1A)	**Vincent Wang-Mascianica** *Talking space: inference from spatial linguistic meanings* (T1.1B)
			Manuel Blum *Insights from the Conscious Turing machine* (T1.2A)	**Camilo Miguel Signorelli** *Reasoning about conscious experience with axiomatic and graphical mathematics* (T1.2B)
			Theodor Nenu *The strange loops inside our brains* (T1.3A)	**Quanlong Wang** *Mathematical framework for consciousness-only—modeling playa consciousness in process theory* (T1.3B)
Hour 2 3 PM - 4 PM	**Majid D. Beni** *Neither a qualia realist nor an illusionist be* (M2.1A)	**Carlotta Langer** *How morphological computation shapes integrated information in embodied agents* (M2.1B)	**Discussion sessions**	
	Ariel Zeleznikov-Johnston *Theoretical and empirical developments towards a category-theoretic framework for phenomenal holism* (M2.2A)	**Pedro Mediano** *What it is like to be a bit: An Integrated Information Decomposition account of emergent mental phenomena* (M2.2B)		
	Natesh Ganesh *No substitute for functionalism* (M2.3A)	**Ignacio Cea** *A vector model of how consciousness as integrated information makes a causal difference* (M2.3B)		
Break				
Hour 3 5 PM - 6 PM	**Discussion sessions**		**Adrian Kent** *Beyond IIT: can we model the evolution of consciousness?* (T3.1)	
			Johannes Kleiner *A no-go theorem for the closure of the physical* (T3.2)	
			Larissa Albantakis *Why a little bit of causal structure is necessary… even for functionalists* (T3.3)	
Hour 4 6 PM - 7 PM	**Kobi Kremnitzer** *Collapse and the closure of the physical* (M4.1A)	**William Marshall** *A Measure for Intrinsic Information* (M4.1B)	**Discussion sessions**	
	Tim Palmer *Free will and consciousness - a new perspective based on invariant set theory* (M4.2A)	**Adam Safron** *Integrated World Modeling Theory (IWMT) and the physical and computational substrates of consciousness* (M4.2B)		
	Pedro Resende *Revisiting the measurement problem and qualia* (M4.3A)	**Hans Georg Zimmerman** *Extending mathematical models of artificial intelligence to consciousness* (M4.3B)		

Time (UTC)	Wednesday, September 29	Thursday, September 30	
Hour 1 **2 PM - 3 PM**	**Eva Jablonka** *Consciousness as we know it: a view from biology* (W1.1)	**Discussion sessions**	
Hour 2 **3 PM - 4 PM**	**Discussion sessions**	**Discussion sessions**	
Break			
Hour 3 **5 PM - 6 PM**	**Shimon Edelman** *Autodiagnosis and the Dynamical Emergence Theory of Basic Consciousness* (W3.1)	**Andrea Luppi** *A synergistic workspace for human consciousness and cognition revealed by Integrated Information decomposition* (R3.1A)	**Sophie Taylor** *Consciousness from pro arrows: a double-categorical framework for constructive cognitive architectures* (R3.1B)
		Marco Fabus *Hysteron modeling of anaesthetic slow-wave power* (R3.2A)	**Alex Maier** *Cause-effect structure of cortical columnar responses* (R3.2B)
		Matteo Grasso *Of maps and grids* (R3.3A)	**Shanna Dobson** *Making Up Our Minds* (R3.3B)
Hour 4 **6 PM - 7 PM**	**David Chalmers** *Consciousness and the collapse of the wave function* (W4.1)	**Gualtiero Piccinini** *Qualitativism: Consciousness Consist of Physical Qualities* (R4.1)	
7 PM - 7:10 PM	**Susanne Still** *Thermodynamics of real world observers* (W5.1)	**Closing remarks**	
7:10 PM - 8 PM			

Colour key:

Invited
Contributed
Discussion

Abstract locator number:



TITLES AND ABSTRACTS

MONDAY, SEPTEMBER 27

M1.1 **Karl Friston** (University College London)

Deep inference

In the cognitive neurosciences and machine learning, we have formal ways of understanding and characterising perception and decision-making; however, the approaches appear very different: current formulations of perceptual synthesis call on theories like predictive coding and Bayesian brain hypothesis. Conversely, formulations of decision-making and choice behaviour often appeal to reinforcement learning and the Bellman optimality principle. On the one hand, the brain seems to be in the game of optimising beliefs about how its sensations are caused; while, on the other hand, our choices and decisions appear to be governed by value functions and reward. Are these formulations irreconcilable, or is there some underlying information theoretic imperative that renders perceptual inference and decision-making two sides of the same coin? And does a model of consciousness entail a model of how we make choices?

M2.1A **Majid D. Beni** (Middle East Technical University)

Neither a qualia realist nor an illusionist be

The talk explores two extant philosophical interpretations of a theory of consciousness under the free energy principle—these are qualia realist illusionist (or rather semi-illusionist) interpretations—and argues that none of them is completely convincing. Then the paper develops and defends a new philosophical interpretation of the free energy account of consciousness along the lines of strong

intentionalism (as developed in works of Tim Crane), which is in line with structural realism.

M2.2A **Ariel Zeleznikow-Johnston** (Monash University)

Theoretical and empirical developments towards a category-theoretic framework for phenomenal holism

Phenomenal holism is the proposition that an experience is defined not by its intrinsic properties but by its relationship to other experiences a subject could possess. The proposition bears a striking resemblance to the Yoneda lemma in category theory, a statement showing that any object in a category is defined by its relationship to all the other objects in that category. Inspired by these ideas, we have obtained preliminary empirical support for using this framework to characterise phenomenology. Specifically, we have shown that colour experiences across the visual field are equivalent through capturing the subjective similarity relationships between these experiences at different locations. Nonetheless, theoretical concerns remain around how this approach can respond to the 'inverted qualia' thought experiments. A specific worry is that if experiences are not defined intrinsically, it seems that specific experiences can become unbound from the substrate upon which they supervene. We will explore potential solutions to this problem that do not entail eliminativism. Ultimately, development of this framework will help in verification of the neurophenomenal structuralism hypothesis, which holds that the relationships between neural correlates of consciousness should structurally resemble their phenomenological counterparts.

M2.3A **Natesh Ganesh** (NIST/CU Boulder)

No Substitute for Functionalism

In this talk, the author will expand on the model of falsification proposed in 'Falsification & Consciousness' by Kleiner & Hoel, allowing us to identify two different types of 'variation' — Type-1 and 2. We will show that only variations of Type-2 lead to the falsification results in the original paper. Motivated by examples

from neural networks, finite state automata and Turing machines, we will prove that substitutions based on Type-2 variations do not exist for a very broad class of functionalist theories, rendering them immune to the substitution argument. We will briefly discuss implications for both the substitution argument and functionalist theories of consciousness.

M2.1B **Carlotta Langer** (Hamburg University of Technology)

How morphological computation shapes integrated information in embodied agents

The Integrated Information Theory provides a quantitative approach to consciousness and can be applied to neural networks. An embodied agent controlled by such a network influences and is being influenced by its environment. We present a technique combining different methods in order to examine the information flows among and within the body, the brain and the environment of an agent. This allows us to relate various information flows to each other. We demonstrate the implications of this framework within a simple experimental setup. There, the optimal policy for goal-directed behavior is determined based on the "planning as inference" method, in which the information-geometric em-algorithm is used to optimize the likelihood of the goal. Morphological computation and integrated information are then calculated with respect to the optimal policies. Comparing the dynamics of these measures under changing morphological circumstances highlights the antagonistic relationship between these two concepts. The more morphological computation is involved, the less information integration within the brain is required. Furthermore, we argue that it is necessary to additionally measure the information flow to and from the brain in order to determine the influence of the brain on the behavior of the agent.

M2.2B **Pedro Mediano** (University of Cambridge)

What it is like to be a bit: An Integrated Information Decomposition account of emergent mental phenomena

In this talk we outline an approach to consciousness science that combines insights from Integrated Information Theory (IIT) and the mathematical formalism of Partial Information Decomposition (PID), which we term Integrated Information Decomposition (Φ_{ID}). Through this approach, we lay out a formal argument relating consciousness and causal emergence in any given system based on their respective information-theoretic compositions — providing a principled answer to the long-standing dispute on the relationship between consciousness and emergence. Furthermore, leveraging the empirical tractability afforded by Φ_{ID}, we build a revised measure of integrated information, Φ_R, and argue that a multi-dimensional description of information dynamics based on Φ_R and related quantities may provide a principled way to compare the neural basis of different aspects of consciousness. The 'modes of consciousness' outlined by Φ_{ID} establish a common space for mapping the phenomenology of different conscious states, which as an example we explore through the concept of selfhood. Overall, Integrated Information Decomposition yields rich new ways to explore the relationship between information, consciousness, and emergence in a way that is mathematically rigorous, empirically driven, and ontologically innocent.

M2.3B **Ignacio Cea**

A vector model of how consciousness as integrated information makes a causal difference

In this talk, we present a vector model of how consciousness, as depicted by Integrated Information Theory (IIT), can exert genuine causal powers on its physical substrate. IIT identifies consciousness with an irreducible structure of causal powers that a system of mechanisms in its current state specifies. So, IIT asserts that consciousness is "supremely causal" (Tononi 2012, p. 309). Moreover, IIT's first axiom is that consciousness exists, and one of its

key theoretical assumptions is that to exist is to have causal power. Therefore, consciousness must have causal power, on pain of internal incoherence. However, there is a tension between consciousness being causally powerful and the fact that, in IIT's framework, the behavior of any conscious system seems fully determined by its non-phenomenal properties.To resolve this, we argue that IIT's interventionist framework should be complemented with a "causal powers" metaphysical view of causation, and propose a mathematically formulated method to translate IIT's irreducible causal structure (=consciousness) to a vector model where phenomenal powers interact and determine what state of activation the system is more disposed to be in the immediate future. In this way, we could advance our understanding of how consciousness as integrated information makes a causal difference.

M4.1A **Kobi Kremnizer** (University of Oxford)

Collapse and the closure of the physical

In this talk I will describe recent joint work with Johannes Kleiner. I will give a precise definition of the closure of the physical and then look at the implication to physics if the (current) physical is not closed: quantum collapse theories. I will discuss the implications of this to scientific theories of consciousness.

M4.2A **Tim Palmer** (University of Oxford)

Free will and consciousness — a new perspective based on Invariant Set Theory

Motivated by chaos theory, Invariant Set Theory posits that the universe is a deterministic system which evolves precisely on a high-dimensional fractal invariant set in cosmological state space. As such, it is proposed that the laws of physics at their most primitive describe the geometry of this fractal invariant set. This postulate violates the so-called Statistical Independence assumption in Bell's Theorem and as such can account for quantum entanglement without the need for indeterminism or violation of local causality. We propose that as a consequence of the brain's extraordinary

energy efficiency, it makes use of quantum processes which would otherwise be too energetically expensive (e.g. to ensure a rapid enough flow of ions across axon membranes). We interpret the brain as a hybrid quantum/classical system from the perspective of Invariant Set Theory. The key idea discussed in this presentation is based around the notion that we have a weak sense of neighbouring state-space trajectories on the invariant set. This provides novel explanations of why, despite determinism, we feel so deeply that we have free will, and indeed why perceive ourselves as conscious beings.

M4.3A **Pedro Resende** (Instituto Superior Técnico)

Revisiting the measurement problem and qualia

The measurement problem in quantum mechanics hinges on a description of quantum systems in terms of their states (wave functions), with measurements being operations that correlate with unavoidable state changes. In this talk I present a geometric approach to measurements which takes them to be fundamental processes that occur interdependently with the origination of classical information. This approach can be regarded as a mathematical formalization of Wheeler's "it from bit," and it is based on a topological notion of space of measurements whose algebraic structure caters for a primitive notion of time and causality. Such a model emphasizes the mathematical structure of measurements, whereas states and observers are derived entities. I will focus on the definition and basic properties of measurement spaces and revisit the idea, following a previous talk at Models of Consciousness 1 (2019), that measurements should be identified with qualia.

M4.1B **William Marshall** (Brock University)

A Measure for Intrinsic Information

We introduce an information measure that reflects the intrinsic perspective of a receiver or sender of a symbol, who has no access to the communication channel and its source or target. The measure satisfies three desired properties — causality, specificity, intrinsicality

— and is shown to be unique. Causality means that symbols must be transmitted with probability greater than chance. Specificity means that information must be transmitted by an individual symbol. Intrinsicality means that a symbol must be taken as such and cannot be decomposed into signal and noise. It follows that the intrinsic information carried by a symbol increases if the repertoire of symbols increases without noise (expansion) and decreases if it does so without signal (dilution). An optimal balance between expansion and dilution is relevant for systems whose elements must assess their inputs and outputs from the intrinsic perspective, such as neurons in a network. The measure and its implications are discussed in the context of the integrated information theory of consciousness.

M4.2B **Adam Safron** (Johns Hopkins University)

Integrated World Modeling Theory (IWMT) and the physical and computational substrates of consciousness

This presentation will discuss the physical and computational substrates of consciousness as suggested by Integrated World Modeling Theory (IWMT). I will first review how neural synchrony may entail marginalization over Bayesian networks, implementing the implicit calculations of joint beliefs and establishment of marginal message-passing regimes within a predictive processing context. I will describe how association cortices may correspond to shared latent (work)spaces within an autoencoding framework, potentially structured according to principles of geometric deep learning. Along these lines, I will further describe how Scott Aaronson's "expander graph" critique of Integrated Information Theory may actually provide surprisingly strong support for the use of Phi in identifying necessary (but not sufficient) conditions for realizing potentially conscious systems. I will then discuss how IWMT's original proposal of the realization of consciousness via alpha-synchronized subnetworks may have contained an error, with subjective experience potentially entailed by beta-complexes generating inferences at faster speeds with more restricted scopes. Finally, time permitting, I will present a model of how pleasure and pain may correspond to conflicting predictions within these

subnetworks, potentially suggesting mechanisms by which such processes could be modified with relatively low-tech interventions.

M4.3B **Hans Georg Zimmerman** (Fraunhofer IIS)

Extending Mathematical Models of Artificial Intelligence to Consciousness

Artificial Intelligence (AI) and Human Intelligence (HI) try to solve similar problems (perception, understanding, action), even if they use different solution approaches. Obviously, the AI side is accessible to mathematical modeling. Under different circumstances both sides have different advantages or disadvantages.Is it possible to extend the above AI mathematics to consciousness? The author starts with a conception of the entities: world, mind, self and an analysis of their interconnection. To model the mind as an interface between world and self we can use an extended version of the above AI model. This mathematical concept allows a consistent description of perception, understanding, action even if one favors a dualistic picture of the world. One contribution of the talk is, that we do not need a hardware bridge between self, mind, world. Their alignment can be explained as a result of a learning process. This talk is on the interface between Computer Science and Philosophy, but still has consequences for Neurophysiology.

TUESDAY, SEPTEMBER 28

T1.1A **Lenore Blum** (Carnegie Mellon University)

A Theoretical Computer Science Perspective on Consciousness

The Conscious Turing Machine (CTM) is a machine model of consciousness inspired by Alan Turing's simple yet powerful model of a computer and Bernhard Baars' Global Workspace model of consciousness. In this brief presentation, we explore how the CTM might experience phenomena generally associated with consciousness and, while we do not claim this is how humans experience these phenomena, we suggest it provides some high-level understanding of how such experiences might be generated.

We start with three examples related to vision: blindsight, inattentional blindness, and change blindness. Then we consider illusions, dreams and free will. This is joint work of Lenore, Manuel and Avrim Blum.

T1.2A **Manuel Blum** (UC Berkeley)

Insights from the Conscious Turing Machine

The quest to understand consciousness, once the purview of philosophers and theologians, is now actively pursued by scientists of many stripes. In this talk, we discuss consciousness from the perspective of theoretical computer science (TCS), a branch of mathematics concerned with understanding the underlying principles of computation and complexity, especially the implications of resource limitations. In the manner of TCS, we formalize the Global Workspace Theory (GWT) originated by cognitive neuroscientist Bernard Baars and further developed by him, Stanislas Dehaene, and others. Our principal contribution lies in the precise formal definition of a Conscious Turing Machine (CTM). We define the CTM in the spirit of Alan Turing's simple yet powerful definition of a computer, the Turing Machine (TM). We are not looking for a complex model of the brain nor of cognition but for a simple model of (the admittedly complex concept of) consciousness. After defining CTM, we give a formal definition of consciousness in CTM. We then suggest why the CTM has the feeling of consciousness. The perspective given here provides a simple formal framework to employ tools from computational complexity theory and machine learning to further the understanding of consciousness.

T1.3A **Theodor Nenu** (University of Bristol)

The Strange Loops Inside Our Brains

In this talk, we critically investigate Douglas Hofstadter's (1979, 2007) analogical appeals to Kurt Gödel's (1931) First Incompleteness Theorem, whose 'diagonal' proof supposedly contains the key ideas required for understanding both consciousness and mental

causation. We conclude that there are simply too many weighty details left unfilled in Hofstadter's proposal which really need to be fleshed out before we can even hope to say that our understanding of classical mind-body problems has been advanced through metamathematical parallels. We maintain that bringing Mathematical Logic into play cannot furnish mentalistic insights which would otherwise be unavailable—on the contrary, the Gödelian discourse may be seen as a distraction from the various insightful points that Hofstadter makes.

T1.1B **Vincent Wang-Mascianica** (University of Oxford)

Talking Space: inference from spatial linguistic meanings

The physical space around us is a fundamental cognitive model for conceptual and metaphorical meaning. We propose a formal mechanism for how space and linguistic structure interact in a matching compositional fashion. Once the model is in place, we show how inferences drawing from the structure of physical space can be made, resulting in a rich compositional model of meaning close to human embodied experience in the world.

T1.2B **Camilo Miguel Signorelli** (University of Oxford)

Reasoning about conscious experience with axiomatic and graphical mathematics

We cast aspects of consciousness in axiomatic mathematical terms, using the graphical calculus of symmetric monoidal categories and Frobenius algebras. This calculus exploits the ontological neutrality of category theory. A toy example using the axiomatic calculus is given to show the power of this approach, recovering other aspects of conscious experience, such as external and internal subjective distinction, privacy or unreadability of personal subjective experience, and phenomenal unity, one of the main issues for scientific studies of consciousness. In fact, these features naturally arise from the compositional nature of axiomatic calculus.

T1.3B **Quanlong Wang** (Cambridge Quantum Computing)

Mathematical framework for Consciousness-only—Modeling Alaya consciousness in process theory

Yogacara (also known as Consciousness Only) is one of the two main branches of Mahayana (Great Vehicle) Buddhism (the other one is Madhyamaka (Middle Way)). The key feature of the Yogacara philosophy is consciousness-only which means there is nothing existing completely independent of all sentient beings' consciousnesses. Via the view of consciousness-only, the so-called hard problem of consciousness can be avoided. However, another difficult problem is unavoidable: how could the existence of the physical world and bodies be explained by consciousness? This problem has been systematically investigated in the Yogacara literatures in some way philosophically, especially under the concept of Alaya consciousness(also called storehouse consciousness or seed consciousnes). In this talk, we propose a mathematical framework based on process theory (also known as symmetric monoidal category theory) for modelling the Alaya consciousness. In particular, we demonstrate that the proposed model satisfies the six characteristics of seeds of Alaya consciousness which can be seen as a standard axiomatic characterisation of those seeds in Yogacara philosophy. This model paves a way for accounting for the phenomenon of physical laws from the aspect of consciousness.

T3.1 **Adrian Kent** (University of Cambridge)

Beyond IIT: can we model the evolution of consciousness?

Tononi et al.'s ``integrated information theory'' (IIT) postulates rules for assigning measures Φ and qualia types Q of consciousness to classical information networks. We consider whether IIT is compatible with Darwinian evolution. We argue that an IIT-like theory that assigns consciousness to physical systems by relatively simple mathematical rules poses extraordinary fine-tuning problems. We introduce IIT+, a class of extensions of IIT, and argue that IIT+-like theories offer at least partial explanations of how some key features of consciousness evolved, which IIT-like theories cannot

explain. We conclude that if one takes seriously Darwinian evolution and the case for an IIT-like theory, one has to take seriously the case for an IIT+-like theory.

T3.2 **Johannes Kleiner** (LMU Munich/MCMP)

A no-go theorem for the closure of the physical

We analyze the implications of the closure of the physical for experiments in the scientific study of consciousness when all the details are considered, especially how measurement results relate to physical events. It turns out that the closure of the physical implies that no experiment can distinguish between two theories of consciousness that obey this assumption. Therefore, the closure of the physical is incompatible with scientific practice. This conclusion points to a fundamental flaw in the paradigm underlying most of the experiments conducted to date.

T3.3 **Larissa Albantakis** (University of Wisconsin-Madison)

Why a little bit of causal structure is necessary... even for functionalists

One of the easiest ways of testing a theory of consciousness is to apply it to a room full of people — 3 awake adults, for example. If the theory cannot identify (at least) three consciousnesses there is a problem. Yet, almost no proposal to date has the theoretical tools to approach this problem without presupposing brains as the seats of experience (IIT being a notable exception, see also Fekete et al. (2016)). Crucially, causal analysis is the only way to address this individuation problem as I will argue by means of a thought experiment inspired by Searle's Chinese Room Argument, which I titled "The Greek Cave". The reason is that identifying conscious individuals in general requires intervening upon the system under study. The implication is that no theory of consciousness can do entirely without assessing causal structure. Any functionalist account

that completely lacks reference to causal structure is inadequate, because the same (complex) input-output function can always be distributed across different numbers of consciousnesses.

WEDNESDAY, SEPTEMBER 29

W1.1 **Eva Jablonka**

Consciousness as we know it: a view from biology

Although there are many debates about the distribution of consciousness (or subjective experiencing), there is a general agreement among biologists that the only consciousness that we are currently aware of is the consciousness of living beings. I present a view of biological consciousness developed by Simona Ginsburg and myself, which conceptualizes consciousness as a biological mode of being and investigates it using an evolutionary approach. We adopt a methodology suggested by the Hungarian system-chemist Tibor Gánti for the study of minimal life to the study of minimal consciousness. We start by listing the set of capacities that are deemed necessary for minimal consciousness by biology-oriented scholars, and follow it by identifying a single capacity, an evolutionary transition marker, which, when present, entails the complete set of the consciousness-characterizing capacities and points to the completion of the evolutionary transition to a new sustainable mode of being (minimal consciousness in our case). The evolutionary transition marker that we suggest is a form of domain-general, open-ended associative learning, which we call unlimited associative learning (UAL). We argue that the organizational dynamics of UAL constitute the dynamics of minimal consciousness and put forward a toy model of UAL. We suggest that the neural implementation of UAL evolved over time and that a comparative study of UAL in different animal groups can resolve some of the current debates about the necessary and sufficient conditions for consciousness and inform us about the different forms it takes. I end with some theoretical generalizations and predictions stemming from the UAL model.

W3.1 **Shimon Edelman**

Autodiagnosis and the Dynamical Emergence Theory of Basic Consciousness

Phenomenal awareness — the basic, selfless kind of consciousness — must involve discernment among states of affairs: "this, not that" rather than simply "this" (which would be meaningless on its own). Furthermore, such discernment must be intrinsic to the system: the distinctions among states must arise from its own dynamics, rather than through outside interpretation. These two considerations suggest that basic consciousness, with or without a self-model, may amount to dynamic autodiagnosis: a process whereby the system tells apart its own states, such that distinct ones result in qualitatively different state-space trajectories, as dictated by the system's dynamics. Accordingly, our Dynamical Emergence Theory of consciousness (DET) defines the amount and structure of phenomenal experience in terms of the intrinsic topology and geometry of a physical system's collective dynamics. In particular, we posit that distinct perceptual states correspond to autodiagnosed coarse-grained macrostates reflecting a self-consistent partitioning of the system's state space — a notion that aligns with several ideas and results from computational neuroscience and cognitive psychology.

W4.1 **David Chalmers** (New York University)

Consciousness and the collapse of the wave function

Does consciousness collapse the quantum wave function? This idea was taken seriously by John von Neumann and Eugene Wigner but is now widely dismissed. We develop the idea by combining a mathematical theory of consciousness (integrated information theory) with an account of quantum collapse dynamics (continuous spontaneous localization). Simple versions of the theory are falsified by the quantum Zeno effect, but more complex versions remain compatible with empirical evidence. In principle, versions of the theory can be tested by experiments with quantum computers. The upshot is not that consciousness-collapse interpretations are clearly correct, but that there is a research program here worth exploring.

W5.1 **Susanne Still** (University of Hawaii)

Thermodynamics of real world observers

Observers are real world entities and as such subjected to the laws of physics. They can acquire, process, and use information. There are fundamental bounds governing how efficient, how fast, how reliable observers can function. Knowing these physical limits puts us in a situation where we can ask: what are the design principles ensuring that observers can potentially reach the fundamental limits? This talk explores one such limit: energetic efficiency. We find that minimally dissipative observers need to perform predictive inference. I will introduce generalized, partially observable information engines as the theoretical construct for studying intelligent information processing by real world observers, and discuss a canonical example that shows that optimal observer memories would be hard to guess without use of this theoretical framework.

THURSDAY, SEPTEMBER 30

R3.1A **Andrea Luppi** (University of Cambridge)

A synergistic workspace for human consciousness and cognition revealed by Integrated Information Decomposition

A central goal of neuroscience is to understand how the brain synthesises information into a unified conscious experience. Here, we address two fundamental questions: how is the human information-processing architecture organised, and how does it support consciousness? Developing an information-resolved approach to functional connectivity in the human brain, we reveal that sensorimotor processing is supported by redundant interactions, whereas integrative processes rely on synergistic information, which is more prevalent in humans than non-human primates, with high-synergy regions exhibiting the highest degree of evolutionary cortical expansion and synaptic density. We delineate a "synergistic global workspace" architecture comprising gateway regions that

gather information from specialised modules, which is then integrated within the workspace and widely distributed via broadcaster regions. Through functional MRI analysis, we show that gateway regions of the workspace correspond to the brain's default network, whereas broadcasters coincide with the executive control network.Remarkably, loss of consciousness due to anaesthesia or brain injury corresponds to reduced integrated information between gateway regions of the synergistic workspace, which is restored upon recovery. Thus, loss of consciousness may coincide with a breakdown of information integration accessing the synergistic workspace of the human brain, providing an avenue to reconcile prominent theoretical accounts of consciousness.

R3.2A **Marco Fabus** (University of Oxford)

Hysteron Modeling of Anaesthetic Slow-wave Power

During anaesthetic loss of consciousness, the brain's electrical activity changes dramatically. Cortical neurons oscillate at ~1Hz, producing a spatiotemporally-varying electric potential at the scalp, measurable on an electroencephalogram (EEG). This slow-wave activity (SWA) shows saturation (SWAS) as anaesthetic concentration increases. At even higher, but still clinically relevant anaesthetic concentrations, SWA is seen to decrease and shows hysteresis. This is one example of 'neural inertia', i.e. brain state asymmetry between losing and regaining consciousness. This talk will present a novel, physiologically-motivated model of anaesthetic slow-wave power that can account for these phenomena: the hysteron model. The model fits available clinical data significantly better than commonly used sigmoid models. I will outline the model's link to Preisach hysteresis, underlying neuron behaviour, and complexity-based theories of consciousness.

R3.3A **Matteo Grasso** (University of Wisconsin-Madison)

Of maps and grids

In neuroscience, consciousness is usually approached in functional terms: the goal is to understand how the brain represents

information, accesses that information, and acts on it. But this functional, information-processing approach leaves out the subjective structure of experience: how experience feels. Here we consider a simple model of how a "grid-like" network meant to resemble posterior cortical areas can represent spatial information and act on it to perform a simple "fixation" function. We show how the model represents topographically the retinal position of a stimulus and triggers eye muscles to fixate or follow it. Encoding, decoding, and tuning functions of model units fully explain what the model does. However, these functional properties leave out the fact that a human fixating a stimulus would also experience it at a location in space. Using Integrated Information Theory (IIT), we show how the subjective properties of experienced space—its extendedness—can be accounted for in objective, neuroscientific terms by the cause-effect structure specified by the grid-like cortical area. By contrast, a "map-like" network without lateral connections, meant to resemble a pretectal circuit is functionally equivalent to the grid-like system with respect to representation, action, and fixation, but cannot account for the phenomenal properties of space.

R3.1B **Sophie Taylor** (Queensland University of Technology)

Consciousness from proarrows: A double-categorical framework for constructive cognitive architectures

Hitherto, the most comprehensive formalisations of cognition have been in the form of computational cognitive architectures; large computer programs comprising of a number of disparate cognitive mechanisms working in tandem to produce extremely emergent behaviour. However, the study of the mathematical foundations of thesesystems has been relatively neglected.We present an approach to the categorification of cognitive architectures based on virtual equipments, double-pushout rewriting, and profunctor optics. Further, we examine a number of highly suggestive questions that arise naturally in the process, such as a possible explanation of the utility of quantum probability theory in cognitive science.

R3.2B **Alex Maier** (Vanderbilt University)

Cause-Effect Structure of Cortical Columnar Responses

Formalized phenomenology is a revolutionary approach to the scientific study of consciousness. The goal of this research program is to derive mathematically formalized structures of conscious perceptual (phenomenal) experience. The great promise of this approach is that phenomenological structures can be linked to brain activity in ways that go beyond simple correlation. Once both phenomenal and brain activity structures are formalized, well-developed mathematical tools such as category theory can be used to study the exact relationship (i.e., functors) between them. In other words, formalized phenomenology offers the unique potential to uncover mathematically formalized laws that explain phenomenal experience as brain activity. Recent years have seen great progress in formalizing phenomenological structure. In comparison, mathematically formalized structures of brain activity are less well developed. Here we demonstrate that Integrated Information Theory (IIT) can be used to derive formal cause-effect structures of columnar neuronal responses in visual cortex. Specifically, we characterize the internal causal structure of neuronal up and down states measured in different layers of primary visual cortex using linear multielectrode arrays (LMAs), the basic technique behind Neuropixels and Elon Musk's Neuralink. Deriving neural cause-effect structures in this way marks an important step towards linking formalized phenomenology and associated brain responses.

R3.3B **Shanna Dobson** (CSULA/UC Riverside)

Making Up Our Minds

Disciplines including computer and cognitive science, developmental and evolutionary biology, and fundamental physics have advanced empirically-supported formal models in which perceived objects are “made up” in the sense of being observer-relative inferential constructs. In these models, the cognitive ontology an agent employs to organize its behavior is conditionally independent of the “real” ontology, if any, of its environment. Here we claim that these same models apply, without modification, to any agent’s metacognition about itself. “Minds” and in particular

thoughts, beliefs, desires, and memories are just as "made up" as the external world and its objects, and are made up using the same mechanisms and for the same reasons. We describe this process in the physical language of holographic encoding on a boundary, and then focus on the case of time perception, episodic memory, and planning using the mathematical language of sheaf theory and condensed objects. We summarize a formal framework in which both retrospective and prospective memory are pure constructs from the present, reconstructed on demand moment to moment. Our approach makes a specific empirical prediction: that the neural mechanisms that generate retrospective and prospective memories implement a particular functor acting on sheaves of categorizations of observed events.

R4.1 **Gualtiero Piccinini**

Qualitativism: Consciousness Consists of Physical Qualities

I argue that consciousness consists at least in part of physical qualities. I begin by introducing two types of natural property: physical qualities and causal powers. I introduces levels of composition and realization. I introduce mechanisms and the notions of multiple realizability and medium independence. I argue that physical computation is a medium-independent notion. I introduce living organisms and the teleological functions that organismic traits and artifacts have. Finally, I argue that cognition is largely medium-independent and hence a matter of computation but phenomenal consciousness most likely involves physical qualities, which are aspects of physical reality that outstrip its causal powers.