Transcript of:
Active Inference GuestStream #024.1
Stephen Grossberg, "Explainable and Reliable AI"

Stephen Grossberg
"Explainable and Reliable AI"
https://sites.bu.edu/steveg/

Active Inference Institute
https://www.activeinference.org/
https://www.youtube.com/c/ActiveInference

============ Session 1 ==============
Session #024.1, June 28, 2022
"Explainable and Reliable AI"
Livestreamed on Jun 29, 2022
Duration: 1:46:39
https://www.youtube.com/watch?v=tOFA7FODn8w

Speakers in Session
Prof. Stephen Grossberg
Daniel Ari Friedman
Ali Rahmjoo

[sp: DANIEL FRIEDMAN]
00:01 Hello and welcome, everyone. It is June 28th, 2022 and we are here in ActInf Lab GuestStream number 24.1. Today we're here with Professor Stephen Grossberg and the agenda will be as follows. First, Ali will provide a short introduction. We will then play a 45-minute pre-recorded video, followed by a Q&A. So thanks, everyone, for joining, and Professor Grossberg, really appreciate joining, and I'll pass to Ali for the introduction.

[sp: ALI RAHMJOO]
00:42 Hello and welcome. I'm Ali. I'm an independent researcher from Iran. I'm very happy and excited to be here and be able to speak with Professor Grossberg today. So I'd like to thank Professor Grossberg for joining us. Stephen Grossberg is the Wang Professor of Cognitive and Neural Systems and a Professor Emeritus of Mathematics & Statistics, Psychological & Brain Sciences, and Biomedical Engineering at Boston

University.

01:11 For more than 50 years, he has led pioneering research in discovering and developing neural design principles for autonomous adaptive intelligence based on biological and machine learning. His neural network models have been applied to many large-scale problems in engineering and technology, including the design of increasingly autonomous adaptive algorithms and mobile agents. In fact, this is what Karl Friston says about him:

01:38 "Whenever you claim to be ‘the first to do’ this or that in artificial intelligence, it is customary—and correct—to add ‘with the exception of Stephen Grossberg.’ Quite simply, Stephen is a living giant and foundational architect of the field." Professor Grossberg is the recipient of the 2015 Norman Anderson Lifetime Achievement Award of the Society of Experimental Psychologists, the 2017 Frank Rosenblatt Award of the IEEE Computational Intelligence Society, and the 2019 Donald O. Hebb Award of the International Neural Network Society.

02:17 His latest book, *Conscious MIND Resonant BRAIN (Grossberg, 2021)*, as a culmination of his decades long research written in a rather non-technical and conversational style, is published in 2021 by Oxford University Press and is the winner of the Association of American Publishers’ 2022 PROSE Award for the Best Book of the Year in Neuroscience. Now I'll pass it to Professor Grossberg and then we'll continue with the 45-minute pre-recorded lecture.

[sp: Friedman]
02:49 If you'd like to say hi. Otherwise, I'll begin the recorded video.

[sp: STEPHEN GROSSBERG]
02:53 Oh, I just saw my face frozen on the screen! Well, I'm delighted to be here, and I hope you find some points of interest in the lecture. And I'll look forward to the Q&A. Ali has prepared a series of questions that I've thought about and have some prepared, sketched answers. And then after that, if you're still interested, I'm happy to do a live Q&A about anything related to the topics of the day.

[sp: Friedman]
03:34 OK. Onto the main course! I will play the video now and you won't hear anything on the live stream. I’ll crop it and the audio will be coming through fine now.

[sp: Grossberg]
03:53 Hello. I'm delighted to be able to speak to you today about a topic concerning

artificial intelligence, which, as you know, is very much in the news these days. And I'll be contrasting two very different approaches to artificial intelligence. But to do that, I need to pull up my PowerPoint slides and share them with you and let me maximize them and minimize my face.

04:32 So my topic today is “Explainable and Reliable AI: Comparing Deep Learning with Adaptive Resonance.” This lecture is based on the following article from this year (Grossberg, 2020), which is both open access and on my web page. The article summarizes core problems of Deep Learning, such as its untrustworthiness, because it's unexplainable, and its unreliability, because it experiences catastrophic forgetting. The article explains how Adaptive Resonance overcomes these problems and indeed overcomes 17 problems of Deep Learning and outlines a blueprint for achieving autonomous adaptive intelligence.

05:27 The article is part of a *Frontiers in Neurorobotics* special issue about explainable AI (“Explainable Artificial Intelligence and Neuroscience: Cross-Disciplinary Perspectives,” n.d.), whose editors wrote and I quote: “Though Deep Learning is the main pillar of current AI techniques and is ubiquitous in basic science in real-world applications, it is also flagged by AI researchers for its black-box problem: it is easy to fool us and it also cannot explain how it makes a prediction or decision.”

06:01 In other words, Deep Learning is not trustworthy. No life or death decision, such as a medical or financial decision, can confidently be made based upon a Deep Learning prediction. Deep Learning uses the back propagation algorithm for learning how to predict output vectors in response to input vectors. Back propagation was based on perceptron learning principles that Frank Rosenblatt started to introduce in the 1950s (Rosenblatt, 1958).

06:38 It has a complicated history, which Jürgen Schmidhuber beautifully reviewed in an article from this year (Schmidhuber, 2020). Major contributors include Shun-Ichi Amari (Amari, 1972), Paul Werbos (Werbos, 1974), and David Parker (Parker & Sloan School of Management, 1985). Perhaps one would say that it reached its modern form with simulated applications in Paul's 1974 paper (Werbos, 1974) before being popularized 12 years later by Rumelhart, Hinton and Williams (Rumelhart et al., 1986). This schematic of a back propagation circuit [is] printed from a survey article by Gail Carpenter of neural network models (Carpenter, 1989).

07:20 In it, information flows feedforward from an input stage to an output stage. Learning is supervised by an external teacher who on each trial defines a target or desired output. The teaching signal is the error or mismatch between the actual and the target outputs. The teaching signal in level F_3 of adaptive weights in level F_2 have no network pathway whereby to reach from F_3 to F_2 within the algorithm.

08:00 So the algorithm uses an artifice called *weight transport*, which physically lifts the weights from here and moves them there so that they can be used to control learning. Well, this is clearly a non-local operation as well as being clearly non-biological.

08:25 Back propagation learns through slow learning, which means that the adaptive weights change just a little to reduce error on each learning trial. That requires many trials, that is to say, many repetitions of the whole database to learn possibly hundreds or thousands of trials. This is to be contrasted with fast learning where adaptive weights zero error signals on each trial, just as we can learn a face that we see just once and remember it for a long time.

08:59 If backprop [tries] to use fast learning, it would become wildly unstable. Catastrophic forgetting also occurs in backprop, so that during any learning trial, an unpredictable part of its learned memory can unexpectedly collapse. So Deep Learning, which is based on back propagation, is thus neither reliable nor trustworthy. But why is this? One reason is that all inputs are processed by a shared set of learned weights. The algorithm cannot selectively buffer learned weights that are still predictively useful.

09:41 In particular, there's no attention mechanism. This problem occurs in any learning algorithm whose shared weight updates follow the gradient of the error in response to the current batch of data points, while ignoring past batches. There’ve been multiple efforts to fix back propagation. One is to selectively slow learning on the weights important for learning by optimizing parameters using the Bayes’ rule, as Kirkpatrick et al. (Kirkpatrick et al., 2017) suggested a few years ago.

10:16 But that assumes an omniscient observer who can discover and alter the important weights as well as non-local computation such as the Bayesian computation. The same problem occurs with evolutionary algorithms and diffusion-based neuromodulation and other approaches to try to fix backprop. These efforts to overcome catastrophic forgetting created additional conceptual and computational problems. I view them as adding epicycles to ameliorate a fundamental flaw in the model, which to me is reminiscent of adding epicycles to correct problems in the Ptolemaic model of the solar system.

11:06 As we all know, the Copernican model, that we now accept, didn't require epicycles. Perhaps this is why Geoffrey Hinton, who played a key role in developing both backprop and Deep Learning, said in an Axios interview (LeVine, 2017) a few years ago that, quote, he's “deeply suspicious of back propagation… I don't think it's how the brain works… We clearly don't need all the labeled data…

11:35 My view is, throw it all away and start over.” I would claim we don't have to start over, because these problems were solved in the 1970s and 1980s. In particular in the first issue of the journal *Neural Networks* in 1988, I had an article that listed 17 problems of back propagation that are overcome by Adaptive Resonance (Grossberg, 1988). And here they are.

12:07 With regard to not needing all the labeled data, I noted in the third item here that self-organized unsupervised or supervised learning frees us from needing labels all the time. As to slow learning, I noted that in ART you can have fast or slow learning. Indeed, ART can learn to classify an entire database using fast learning on a single learning trial, as Gail Carpenter and I showed in the 1980s (Carpenter & Grossberg, 1987, 1988).

12:42 Moreover, ART overcomes all 17 problems of back propagation without epicycles. Furthermore, all the core ART predictions have been supported by subsequent psychological and neurobiological data. Indeed, ART is a principled biological and technological theory, unlike backprop and Deep Learning, which are just algorithms. ART has explained data from hundreds of experiments, and it's made scores of predictions that have subsequently received experimental support.

13:28 But why has ART been so successful? There are a number of reasons, but one of them is that ART can be derived from a thought experiment about a universal problem in error correction that I published 40 years ago in *Psychological Review (Grossberg, 1980)*. The thought experiment asks the question, “How can a coding error be corrected if no individual cell knows that one has occurred?”

13:56 Let me quote from my paper: “The importance of this issue becomes clear when we realize that erroneous cues can accidentally be incorporated into a code when our interactions with the environment are simple and will only become evident when our environmental expectations become more demanding. [And] even if our code perfectly matched a given environment, we would certainly make errors as the environment itself fluctuates.”

14:28 So I was talking about autonomous local learning in a changing world. A purely logical inquiry into error correction is translated at every step of the thought experiment into processes learning autonomously in real time with only locally computed quantities. Moreover, the thought experiment uses familiar environmental facts about how we learn as its hypotheses and ART circuits naturally emerge, where these facts are familiar, because they’re ubiquitous environmental constraints on the evolution of our brains, and since we're living with them all the time, they become familiar.

15:18 Because of this universality, ART circuits may thus, in some form, be embodied in all future truly autonomous adaptive intelligent devices, whether biological or artificial. ART has, probably for this reason, already been used in many large-scale engineering and technological applications.

15:44 In fact, almost immediately after ART was introduced, it began being used because it succeeded in benchmark studies against machine learning, back propagation, statistical methods and genetic algorithms, either getting much better accuracy or much faster training speed, or both. It's also used in applications where other algorithms totally fail, such as the Boeing Company's part design reuse and inventory compression application.

16:21 That's just one of many large-scale applications in engineering and technology, some of which can be found on our Tech Lab webpage at bu.edu (http://techlab.bu.edu/resources/articles/C5.html). The Boeing parts design retrieval system in particular was used to help design the Boeing 777. And to do that, you needed fast learning and stable memory to learn and search a huge and continually growing non-stationary parts inventory.

16:51 At the time of this application, there were already 16 million 1-million dimensional vectors that were used to describe each of the parts and you have to be able to quickly search the inventory if you wanted to find a part to use in a new plane.

17:10 Especially if your new design might have a part in the inventory that was similar to it, finding it and slightly modifying [the] design could save millions of dollars in fabrication costs. Satellite remote sensing is another large-scale application that ART was used for very soon, and Gail Carpenter and her colleagues took the lead here (Carpenter et al., 1999). For example, using a very small number of pixels of ground truth of 17 vegetation classes, they used ART to automatically complete these maps using remote sensing data.

17:56 ART did it in a day, rapidly and automatically. It gave a confidence map for each pixel and the pixels with 30 meters in scale, which was small enough to see roads. This contrasted with an AI expert system which took a whole year to do it, and it had to derive ad hoc rules from experts. You had to correct upwards of a quarter of a million site labels, and even so, the pixel size was an order of magnitude larger.

18:31 Gail went on with her colleagues to study information fusion in remote sensing. Let's say you have multiple observers. Each of them may be using different labels. The labels may also be incomplete or missing or even incorrect. And the task was to derive consistent knowledge from potentially inconsistent data to automatically learn and stably store one-to-many mappings.

19:01 And along the way, Gail and her colleagues showed how to self-organize a hierarchy of cognitive rules, including confidence measures between these different levels of the hierarchy. There's been continual work on ART. Some more recent work was summarized in a special issue of *Neural Networks* just in December 2019 (*Neural Networks: Special Issue in Honor of the 80th Birthday of Stephen Grossberg*, n.d.) that was edited by Donald Wunsch, who started the special issue with a general overview of neural network models that I and my colleagues developed (Wunsch, 2019), and then went on in a long and detailed article with several collaborators to provide a survey of Adaptive Resonance Theory neural network models for engineering applications to the present time (Brito da Silva et al., 2019).

19:53 So back propagation and Deep Learning are feedforward adaptive filter[s]. But ART is more than that. In fact, ART is an explainable self-organizing production system in a non-stationary world. What do these words mean? ART is self-organizing, because it can autonomously carry out arbitrary combinations of unsupervised or supervised learning trials with the world as its only teacher. It's a production system, because it uses hypothesis testing to discover and learn rules by a top-down matching process that focuses attention on critical feature patterns. These are the patterns that predict behavioral success while suppressing irrelevant features.

20:44 ART is explainable using both its activities, or short-term memory (STM) traces, and its adaptive weights, or long-term memory (LTM) traces: activation dynamics, learning dynamics. Observing the STM traces in a critical feature pattern explain[s] what recognition categories will learn to code, and what features predict goal-oriented actions. In particular, the long-term memory traces in the fuzzy ARTMAP algorithm translate into explicit fuzzy IF-THEN rules that code what combinations of critical features in what numerical ranges effectively control predictions, thereby illustrating one of many examples where neural networks can learn rule-based behaviors.

21:49 ART includes a bottom-up adaptive filter, a feedforward neural network, as I've observed already, but that's supplemented by top-down learned expectations and two types of recurrent inhibitory feedback interactions that help to choose the recognition categories and the critical features.

22:10 Notably, top-down expectations use what Gail Carpenter and I call the ART matching rule to learn how to focus attention on critical features that control predictive success. The ART matching rule is another way of talking computationally about the process of object attention: how we pay attention to salient objects in the world. And we show how it stabilizes learning and thereby avoids catastrophic forgetting.

22:44 Remarkably, and this has been supported by many data, the ART matching rule can be realized by a top-down, modulatory on-center, off-surround network. What does this mean? Well, let's say we have bottom-up inputs from external features to feature selective cells that get stored in short-term memory. Let's say the bottom-up inputs activate a recognition category which has previously been learned and tries to read out its learned excitatory prototype.

23:21 Well, it can't fully do so, because it also reads out an inhibitory off-surround that's broader than the prototype. So this is approximately one excitatory against one inhibitory. It can only give you a modulatory on-center. But if you have both bottom-up inputs and the top-down expectation simultaneously active, then within the bounds of the prototype, if you also have a bottom-up feature input, you have two excitatory against one inhibitory and those features can be selected, gain-amplified, and synchronized, to start focusing attention on this critical feature pattern, while outlier features, the ones that aren't within the prototype, only have one excitation against one inhibition, or suppressed.

24:19 And in 1999 (Grossberg, 1999), I was able to begin to understand how laminar cortical circuits carry out object attention. In particular, layer VI of a higher cortical area can activate layer VI of a lower cortical area, either directly or via layer V, and then it can send a bottom-up input from layer VI to layer IV, what I call 'folded feedback' to modulate an on-center, and to inhibit an off-surround.

24:50 So attention acts via a top-down, modulatory on-center, off-surround network via folded feedback within laminar neocortex. And this is one example of the paradigm of laminar computing that I introduced, which asks “Why are all neocortical circuits organized in layers? And how do laminar circuits give rise to all kinds of biological intelligence?” Adaptive Resonance answers this story because attended feature clusters reactivate their bottom-up pathways, activated categories reactivate their top-down pathways, closing an excitatory feedback loop between features and categories, giving rise to a feature-category resonance that synchronizes, amplifies, and prolongs system response between the attended critical features and the category to which they are bound.

25:59 And it's this resonance that triggers fast learning in the bottom-up and top-down adaptive weights, which is why I have called the theory Adaptive Resonance Theory. Moreover, I've done a lot of work since then, showing that all conscious states are resonant states and these feature-category resonances are one example of that, one that supports conscious recognition of visual objects and scenes.

26:32 There's a lot of data support for ART predictions. It's well-known that attention does have an on-center off-surround circuit behind it, and that attention can facilitate matched bottom-up signals, many other data as well. So now we can say more about why ART is explainable or trustworthy. In short-term memory, it's because the critical feature patterns determine the attentional focus that controls information processing and you can just read off what those features are.

27:09 In long-term memory, again, it's the critical feature patterns that determine the adaptive weights learned by the bottom-up adaptive filter and the top-down learned expectation. So, you know also what these weights are encoding. ART is reliable and avoids catastrophic forgetting because outlier features that are not in the critical feature pattern are suppressed. So that only the predictive features are processed and coded.

27:40 ART is a production system, because it carries out a kind of hypothesis testing, and this is nicely illustrated in the simplest ART model called ART 1 that Gail Carpenter and I published in 1987 (Carpenter & Grossberg, 1987). ART 1 has an attentional system that does all the category learning and the expectation learning and the paying of attention that interacts with an orienting system which is activated when there are big enough matches in the attentional system and thereby drives a reset and search for either a better-matching category or a new category to learn about novel information.

28:19 Here's a schematic of the ART hypothesis testing and learning cycle. So let's say you have a bottom-up feature pattern coming in. There may be many, many active bottom-up features, but I'll draw just one arrow here for simplicity. But that vector of input features can activate a distributed pattern of feature detector cells.

28:45 Some may be very active, some not so active, some not active at all. And as this is happening, each of these active pathways is trying to turn on the orienting system. So there might be quite a few inputs converging here, but as the features are activated, each of them tries to inhibit the orienting system and there are as many features as there are inputs.

29:12 So this excitation and inhibition are balanced, keeping the orienting system quiet as the feature pattern goes through the adaptive filter and chooses a category. That category reads out a learned top-down expectation that obeys the ART matching rule, which can suppress some mismatched features, thereby reducing the amount of inhibition on the orienting system and raising the question when you have too little inhibition and too much excitation, how big a mismatch will activate the orienting system and cause reset?

29:50 And that ratio is determined by what's called vigilance, which I'll say more about soon. But if you don't have enough inhibition, then the orienting system gets activated. It equally activates all the cells in the category layer because it doesn't know which cell may be active or not.

30:16 So it causes a novelty-sensitive, non-specific burst of arousal (novel events are arousing), thereby selectively shutting off the active category, eliminating its top-down expectation, and unmasking the original feature pattern, which can again go through the adaptive filter. However, now this previously disconfirmed category remains off and the category level is renormalized. So it responds to the same input pattern with a new category and you go through this cycle of resonance and reset until you get a good enough match to either learn a new category or select a previously learned category.

31:08 And it's a theorem that as categories are learned through this matching process, search automatically disengages leading to direct access without search to the globally best-matching category, explaining, for example, how we can quickly recognize familiar objects like your mother's face, even if, as we get older, we store enormous numbers of additional memories. So you don't have to search your whole repertoire.

31:41 When you see Mom, you get direct access and quickly say, "Hi, Mom." There's a lot of support for the hypothesis-testing cycle. One source of support is from event-related potentials, also called human scalp potentials, which shows correlated sequences of three different evoked potentials during oddball learning tasks, an experiment that Jean-Paul Banquet and I reported in the 80s (Banquet & Grossberg, 1987), where you'll get a P120 for a mismatch, an N200 for the arousal activated by the orienting system, and a P300 for the short-term memory reset of the category layer, thereby supporting the processing stages of the search cycle.

32:31 There was also physiological data from inferotemporal cortex where categories are learned early on from the lab of Bob Desimone (Miller et al., 1991), who showed an active matching process that's reset between trials during this kind of event. There's also classical data about hippocampal mismatch dynamics. It's known that novelty potentials subside as learning proceeds from numerous experiments. This is as the orienting system is disengaged. And there's more recent data using multiple-electrode studies from the lab of Earl Miller (Brincat & Miller, 2015), from prefrontal cortex and simultaneous recordings in hippocampus.

33:19 And they show this rapid object associative learning may occur in prefrontal cortex, which is a projection of inferotemporal cortex, one of the stages of category learning, while the hippocampus may guide neocortical plasticity by signaling success or failure. Well, this is just what happens when the attentional system interacts with the orienting system.

33:48 There's also complementary computing in ART. In particular, the attentional and orienting system laws are complementary as manifested by the fact that two event-related potentials are complementary: processing negativity and N200. Processing negativity is activated when there's a top-down match in the attentional system. The N200, as I just noted, is activated when there's a mismatch that activates the orienting system. And you can just look across these four rows and see that these two kinds of ERP potentials are manifestly complementary, as illustrative of the complementarity of the attentional and orienting systems.

34:40 So this leads us to discuss another paradigm introduced, which I call complementary computing, that asks “What is the nature of brain specialization?” Complementary computing introduces new principles of uncertainty and complementarity that clarify why there are multiple parallel processing streams with multiple processing stages in our brains. And a beautiful example of that is this famous image of the macro-circuit of the visual system from David Van Essen and his colleagues (Felleman & Van Essen, 1991), where you can see these multiple parallel-processing streams and the multiple stages needed to achieve what I call hierarchical resolution of uncertainty.

35:32 So what are complementary properties? There are analogies: like a key fits into a lock, or puzzle pieces fitting together. In words, computing one set of properties at a processing stage prevents that stage from computing a complementary set of properties. These complementary parallel processing streams are balanced against one another. It's a very Yin-Yang kind of situation, and interactions between the streams overcome their complementary weaknesses.

36:09 In fact, there are many complementary processes that are known in the brain that have been modeled. Here [are] just five of them. There are many more. So this is a basic principle of brain organization. So in summary, so far backpropagation and Deep Learning do not have short-term memory activation patterns, including critical feature patterns, so they can't pay attention.

36:37 Indeed, they don't have any fast information processing, nor do they have long-term memory top-down learned expectations. So they can’t carry out hypothesis testing using [interacting] short-term and long-term memory traces. Indeed, there's no neural architecture. There's just an algorithm, which greatly contrasts with complementary computing, which discusses the global organization of our brains. From the very start, it was shown how easy it is to get catastrophic forgetting.

37:12 And Carpenter and I showed it in ART when we would shut down the ART matching rule (Carpenter & Grossberg, 1987). Then we demonstrated you could get catastrophic forgetting if you had just four input vectors A, B, C, D presented in the order ABCAD, ABCAD, and so on, if they obeyed very simple subset relationships. And here's a computer simulation of that. Here you don't have the ART matching rule. Here is ABCAD, ABCAD. And you see A is coded by category 1 here, by category 2 here, by category 1 here, 2 here, never settles down.

38:01 But as soon as you impose the ART matching rule, learning is complete by the second trial and after that point you get direct access to the globally best-matching category. Well, let's say a little more about vigilance. Vigilance determines what features are learned in the critical feature pattern. It clarifies how our brains learn concrete knowledge for some tasks and abstract knowledge for others.

38:32 So in particular, high vigilance leads to learning of narrow, concrete categories, like a category that fires selectively to a frontal view of your mother's face. Low vigilance leads to learning of broad and abstract categories, like everyone has a face. It should be emphasized that critical feature patterns are explainable at every level of vigilance. It's known from physiological experiments by Desimone again (Spitzer et al., 1988) that there’s vigilance control in the inferotemporal cortex, which they showed by studying easy versus difficult discriminations in monkeys.

39:17 And in the difficult condition which you'd assume would give you high vigilance, as expected, you had enhancement of the responses and sharpened selectivity for the attended stimuli. How is vigilance computed? Well, let's say an input vector instates a vector of activities at feature detectors, at the same time as it tries to activate the orienting system. The inputs to the orienting system are multiplied by a parameter ρ, which is a sensitivity or gain parameter. That’s vigilance.

39:55 And as these features get [instated], the inhibitory signals from the attentional system to the orienting system try to shut off the orienting system. And if the bottom-up excitation to the orienting system is less than the inhibition, the orienting system stays quiet. So the system can resonate and learn. But if inhibition isn't strong enough, the orienting system gets activated. You get reset and search for new categories. This is a very simple computation, because you have an orienting system that's complementary to the attentional system.

40:30 Well, how do you change vigilance based on predictive success? For this, we have to go from unsupervised to supervised ART models. So we'll have an unsupervised ART_a model, an unsupervised ART_b model, linked together by a learned associative map as occurs in Fuzzy ARTMAP. And the key point is a bottom-up input to ART_a can create an output from ART_b because bottom-up and top-down connections occur at all of these levels.

41:05 So in this way, you can learn many-to-one and one-to-many maps. One example of a many-to-one map is, let's say you're trying to categorize visually processed letters A, which come in multiple fonts. You'll learn various visual categories and they're based on visual similarity. At the same time, you're learning auditory categories for saying A and then the associative map can map all of these visual categories of different As to saying A. But it could have been here that these inputs were symptoms, tests, and treatments in a medical database prediction example and you're predicting length of stay in the hospital.

41:55 The possibilities here are endless and there've been many applications. Or let's say you're trying to figure out what this image is and you've learned to say that's a dog. But today you say it's “Rover” and that causes a mismatch, which drives the search to focus attention on the particular combination of features in this dog that will identify it as “Rover.”

42:26 That leads to learning of a visual category of “Rover,” an auditory category for the name “Rover,” and associative map between them, and you can now simultaneously store expert knowledge about that image. Well, how do you conjointly minimize predictive error and maximize generalization so that you minimize using memory resources? Let me read you an answer and then show what it means in images.

42:59 Match tracking realizes a minimax learning principle, namely, given a predictive error, vigilance increases just enough to trigger search and thus sacrifices the minimum generalization to correct the error. So let's say you've made a prediction. That must mean that vigilance is less than the analog match between bottom up and top down. But let's say now you have a mismatch.

43:28 Well, that'll lead to a match-tracking signal that pumps vigilance up till it's just above the analog match, just big enough to drive the search. So you've given up the minimum amount of generalization to correct the error. Well, are ART mechanisms like vigilance control realized in laminar cortical and thalamic circuits? The answer is yes. My Ph.D. student Max Versace and I showed this by developing the Synchronous Matching ART or SMART model (Grossberg & Versace, 2008), which introduces a lot more neurophysiological and anatomical verisimilitude into the model, including spiking dynamics and laminar cortical circuits, interacting with specific and thalamic nuclei.

44:19 This is another example of laminar computing. And here's a schematic of the model. You see all the cortical layers with identified cells, a hierarchy of cortical regions interacting with specific thalamic nuclei and nonspecific thalamic nuclei. A ton of anatomical data got functionally explained in this way and many other data as well. For example, we showed if you have a good enough match between bottom-up and top-down signals, you're going to get fast gamma oscillations during attention.

44:56 There was quite a bit of data about that already, but we also showed if you have a big enough mismatch, you'll get slower data oscillations. That wasn't well known, but since that time there have been experiments in at least four labs in three different parts of the brain confirming that prediction. Most important, vigilance control was shown to be regulated by mismatch-mediated acetylcholine release.

45:26   So a big enough mismatch in the nonspecific thalamic nucleus activates nucleus basalis of Meynert that releases acetylcholine in layer V cells across the cortex, reducing afterhyperpolarization currents and causing vigilance to go up. And I also showed that breakdown in acetylcholine modulation can help to explain the symptoms of multiple mental disorders. During memory consolidation, we know there's a dynamic phase of memory consolidation, while the input exemplar still drives memory search and before direct access occurs.

46:11 But what if the orienting systems cut out? What if you have a lesion in the hippocampus? Well, then, as occurs in medial temporal amnesia, you get unlimited anterograde amnesia because you can’t search for new categories. You get limited retrograde amnesia because you can’t have direct access to previously learned categories. This is a failure of consolidation, which is mediated by the orienting system.

46:40 So you get defective novelty reactions because that is also mediated by the orienting system. And memory consolidation and novelty detection are mediated by the same structures for the same reason. It's normal priming because priming occurs within the attentional system. Learning of the first item dominates. You can get some learning, but you can't then search. And there's an impaired ability to attend to relevant dimensions of stimuli, again because you can't search.

47:16 So now where does inferotemporal cortex fit in within the larger brain? I introduced the predictive ART, or pART model in order to show how the prefrontal cortex, among other things, learns to control all higher-order intelligence. You can find that in a 2018 paper on my web page, I also published it open access (Grossberg, 2018). And in this macrocircuit, these green areas of prefrontal cortex control processes like working memory, learning plans, prediction, optimized action.

48:01 These regions in red control processes like reinforcement learning, emotion, motivation, adaptively-timed learning. The category learning I've talked about in IT is just in those two regions. All these processes control visual perception and there are detailed models of all of these regions and their interactions now. And each brain region, in nature and in predictive ART, carries out a different function, contrasting really dramatically with the homogeneous organization of the typical deep learning network.

48:44 So I’ve told you just a little bit about some aspects of cognition and why they're explainable, but if you put in all the biological models of perceptual cognition, emotion, and action, they're all explainable. And then you can assert how perceptual and cognitive processes use ART-like excitatory matching and match-based learning to create self-stabilizing attentive and conscious representations of objects and events that embody increasing expertise about the world.

49:21 Moreover, complementary spatial and motor processes, that I couldn't mention at all, use inhibitory matching and mismatch-based learning to continually update spatial and motor representations to compensate for bodily changes throughout life. Taken together, they provide a self-stabilizing perceptual and cognitive front end for conscious awareness and knowledge acquisition, which can intelligently manipulate the more labile spatial and motor processes that enable changing bodies to act effectively on a changing world.

50:02 And when you put them all together, they provide a blueprint for designing autonomous adaptive algorithms and mobile robots with behaviors humans can understand and control, because they're both explainable and reliable. See my web page sites.bu.edu/steveg for these models (*Stephen Grossberg*, n.d.). And with that, I'd like to thank you very much for your attention. Everything I talked about and much more is in my book *Conscious MIND Resonant BRAIN: How Each Brain Makes a Mind*.

50:49 For those who don't know, it's self-contained and non-technical. It's written in a conversational style, so that people who know nothing about the mind or the brain can enjoy reading it. And I have friends who are a rabbi, a minister, a painter, a gallery owner, a lawyer, a social worker who've all been enjoying reading it. Also it's a big book.

51:20 It's almost 800 double-column pages with over 600 color figures. So everything is illustrated. But instead of costing $150, it costs $35 for the hard copy and only $17 for the Kindle. Because I spent a lot of my own money, so that people who are interested in the topic can read it.

51:58 And one other comment, if people do have questions or comments about my lecture or anything they're reading in the book, my email is just Steve (S-T-E-V-E) at BU (Boston University) .edu (steve@bu.edu), and I'll be happy to try to reply, so…

[sp: Rahmjoo]
52:13 Thank you! Well, some researchers and explainable AI people like Leonida Gianfagna and Antonio de Cecco (Gianfagna & Di Cecco, n.d.) demand that any explainable AI should at the very least meet these four criteria: to be fair (not biased in one way or another), to be accountable or reliable, to be secure against malicious hacker attacks and not to be fooled easily, and also to be transparent.

52:41 Now you explained how Adaptive Resonance Theory or ART… And by the way, I got to say, I love your creative and clever use of acronyms for your models. My favorite one is SOVEREIGN Model: Self-Organizing, Vision, Expectation, Recognition, Emotion, Intelligent, Goal-oriented Navigation, if I'm correct. Amazing! Anyway, you explained how ART can address and overcome the issues of accountability, security, and transparency of current Deep Learning approaches.

53:13 But it seems that this fairness issue, a.k.a. the problem of algorithmic bias, has also been of growing concern lately, especially since it's regarded by some researchers like Antonio Badia as a practically intractable problem. So I wanted to ask, in what ways do you think ART can contribute to the ongoing quest for mitigating this problem?

[sp: Grossberg]
53:39 Well, when Ali sent me this question, I said, well, first I'd like you to send me a definition of algorithmic bias that will clarify what you have in mind so that I know what I'm trying to respond to. And you wrote me that you borrowed the term from Badia’s book, *The Information Manifold (Badia, 2019)*. And you sent me a quote from page 247 that I will quote in part before I respond to that background information.

54:15 So “There are two main reasons an algorithmic approach to decision making may result in unfair outcomes, either at the individual or group level. One is that data used is biased, and another is that the algorithm analyzes the data in such a way that it yields biased results… [T]he basic point to remember is that algorithms are designed to achieve a certain goal… not created "naturally" by evolution or accident.

54:46 Thus, most algorithms are written to detect certain patterns of interest for a particular objective, not just any pattern… To be able to pick out some patterns and disregard others, programmers build a model of the data by listing expectations about what data should be like in order to qualify as relevant to the problem." Well, as I'll explain below, self-organizing learning classification and prediction models like Adaptive Resonance Theory or ART overcome all the problems.

55:24 It's a general purpose device. But why don't I try to answer that as part of my replies to Ali's subsequent questions?

[sp: Rahmjoo]
55:38 Okay. Thank you so much. Now, as you also mentioned in your lecture, in 1988 you pointed out 17 issues with back propagation in one of your most famous and highly cited papers on nonlinear neural networks (Grossberg, 1988). So it's been 34 years now. Now, do you see any fundamental Copernican change of perspective happening in Deep Learning research? Or [do] we still keep adding epicycles upon epicycles to our Ptolemaic model?

[sp: Grossberg]
56:13 Well, you've sort of anticipated what I'm going to say, and as I said in my lecture, various investigators that I mentioned, Clune, Kirkpatrick, and Velez all have recently attempted to modify Deep Learning to overcome some of those problems. But as Ali just mentioned, my lecture noted that at least to my mind, they're like epicycles that are added to a kind of Ptolemaic model of the solar system to overcome some of its problems.

56:50 But as we all know, the Ptolemaic model ultimately crashed, because it was both qualitatively and quantitative wrong and they could only be solved by throwing out the Ptolemaic model and replacing it with the Copernican model that became the basis for modern astronomy and astrophysics. So ART overcomes foundational Deep Learning problems that can't be solved using epicycles, and it's already done it.

57:24 As I noted in my lecture, Deep Learning is untrustworthy because it's not explainable and it's unreliable because it can experience catastrophic forgetting. And that happens for a basic reason: Deep Learning, just like backprop, which is its learning engine, is just the feed-forward adaptive filter.

57:57 So as you noted in your question, I described these two problems in addition of 15 others in my oft-cited article that I published in 1988 in the first issue of *Neural Networks (Neural Networks, n.d.)*, and I also showed that ART had already solved the problems in 1976. What I find sad is that backpropagation and Deep Learning architects like Geoff Hinton, who knows all of this background, never mention this history and keep talking about making Deep Learning explain the brain.

58:38 But it can't explain the brain, because its foundation is contradicted by basic psychological and neural data.

[sp: Rahmjoo]
58:49 Yes. Great.

[sp: Grossberg]
58:50 Someone should discuss this problem in the Deep Learning community. I like comparative discussion and criticism, but I don't like solipsism in science.

[sp: Rahmjoo]
59:05 Great. Thank you. Now on slide number 50 of your presentation, you pointed out that ART is inconsistent with models where top-down match is suppressive, such as Bayesian explaining away. A similar view is evident on page 195 of *Conscious MIND Resonant BRAIN*, to which you also add “one of many serious problems of the Bayesian models is that fully suppressive matching circuits cannot solve the stability-plasticity dilemma.”

59:38 Now, would you care to further elaborate on this point?

[sp: Grossberg]
59:42 Copious psychological, anatomical and neurophysiological evidence show that top-down expectations obey the ART matching rule. These expectations are matched against bottom-up input patterns. And as the lecture briefly noted, the ART matching rule is defined by a modulatory on-center, off-surround network, and the modulatory on-center is excitatory.

1:00:28 However, acting by itself, it can't fully excite its target cells. It can prime them, sensitize or modulate them to be ready to fire vigorously when matched bottom-up inputs arrive, and when there is a good enough match between the bottom-up input and an active top-down expectation, that's reading out a circuit that obeys the ART matching rule. That's when you get what I noted in my lecture, what I call a feature-category resonance, because it develops between the matched or attended features and the recognition category that they activate.

1:01:17 And it's this resonance that synchronizes and gain-amplifies the matched features while suppressing the mismatched feature. And that sustained resonance is important because it is sustained long enough to drive learning in the more slowly varying adaptive weights of the active bottom-up filter and learned top-down expectation. And because resonance triggers learning that I call the theory Adaptive Resonance Theory. And the ART matching rule avoids catastrophic forgetting as I briefly mentioned in the lecture, because it suppresses irrelevant features, using its off-surround while it's amplifying and focusing attention on the critical features that regulate both bottom-up and top-down learning as well as successful predictions.

1:02:22 Because they're relevant, they've been selected by previous learning experiences which discover this set of features that are predictive or causal in a given situation. And along the way, not only does the ART matching rule achieve causality and predictions—although as the world changes, you have to update your causal explanations—it also solves the stability-plasticity dilemma. In brief, purely suppressive matching can't do any of this.

1:03:01 It shuts off the expected data and so it can't focus attention or learn about it. And there is fully suppressive matching in spatial and motor learning, but that isn't learning to be expert about the world. I can explain that more if you want to know, but that's also in my book and these two kinds of learning, the excitatory match-based learning and the inhibitory mismatch learning are computationally complementary.

1:03:41 It's another example of complementary computing and the match-based learning goes on in the ventral or “what” cortical stream and the mismatch learning goes on in the “where” or dorsal cortical stream: the “what” stream for perception and categorization or prediction, the “where” stream for the spatial representation and action. And then you need “what” to “where” and “where” to “what” interactions so that you can reach for and otherwise engage through approach and what have you: look at, reach for, approach to things that you've recognized.

[sp: Rahmjoo]
1:04:26 Thank you. Now, following from the previous question and considering that the Free Energy Principle and Active Inference framework as works in progress are related to predictive coding and Bayesian brain hypothesis, what is your view on the extent of compatibility between ART and Active Inference? Because despite some

prima facie similarities between the two, do you see them as fundamentally incompatible or irreconcilable?

1:04:55 And how could this issue be rigorously evaluated and positively resolved in terms of reconciliation or integration of ART and Active Inference or otherwise? Because you see, to add some more context here, Smith et al. in their recent paper, “An Active Inference Approach to Modeling Structure Learning,” (Smith et al., 2020) have stated that although they have not explicitly incorporated ART’s top-down attentional and feedback mechanisms, there are mechanisms within their Active Inference-based model which they believe are quite similar to top-down and bottom-up feedback exchange in ART.

1:05:36 So there seems to be some degree of disagreement about the compatibility between the two frameworks.

[sp: Grossberg]
1:05:44 Well, let me try to respond to the two parts of your question separately. So I'm not going to try to talk about Smith et al. for a moment. Let's talk about free energy. And I like getting definitions on the table, because it's really so frustrating to try to remember what something is when someone's talking about it. So I go to Wikipedia.

1:06:12 Wikipedia writes in part that the Free Energy Principle asserts “that systems minimize the free energy function of their internal state, which entail beliefs about hidden states in their environment. The implicit minimization of free energy is formally related to variational Bayesian methods and was originally introduced by Karl Friston as an explanation for embodied perception neuroscience, where it's also known as Active Inference.

1:06:46 We all know Karl is a brilliant and very insightful man. “The Free Energy Principle describes the behavior of a given system by modeling it through a Markov blanket that tries to minimize the difference between their model of the world and their sense and associative perception. This difference can be described as surprise and it's minimized by continuous correction of the world model of the system.”

1:07:20 One more part of the quote: “The Free Energy Principle has been criticized for being very difficult to understand, even for experts, and the mathematical consistency of a theory may have been questioned by recent studies. Discussions of the principle have also been criticized for invoking metaphysical assumptions far removed from a testable scientific prediction, making the principle unfalsifiable. And in a 2018 interview, Friston acknowledged that the Free Energy Principle is not properly falsifiable.”

1:08:01 So that's Friston himself. So, my main concern with the Free Energy Principle, just like any theory about how brain makes a mind, is how much data can it explain in a principled and unifying way? That's what we do in science. We develop theories to explain and predict data. And in the case of the Free Energy Principle, from what I can see here, there is essentially no data.

1:08:33 And you can correct me if I'm wrong. It therefore cannot be evaluated as a physical theory at all. And there's a basic reason for this problem. Our brains are designed to autonomously learn in real time in response to a changing or non-stationary world that's filled with unexpected events. Like today, we're experiencing an unexpected event that I didn't know till recently that I'd be enjoying your company today.

1:09:07 Optimization principles were designed to cope with stationary dynamics whose rules and probabilities do not change through time. So it's not possible to “minimize the difference between their model of the world and their sense in associated perception” because there is no predefined model of the world, which is always changing in unexpected ways. So you need a theory about how the world changes.

1:09:40 Surprise occurs in ART when there's a big enough mismatch between an input pattern and the currently active top-down expectation of a category that it's activating. This mismatch activates the ART orienting system, that I briefly discussed in my book, which interacts with the attentional system where the category learning does occur. And as I illustrated in our discussions of search and vigilance, that it drives hypothesis testing or memory search to discover a better match or to begin to learn a new category.

1:10:19 So ART will discover a better match in the case where the system was attending to some other familiar features when the new input occurs, but the features in new input have previously been categorized. That's why they're familiar. And then the orienting system very quickly shifts attention to the matching category, and you resonate on and you recognize it—consciously often. ART begins to learn a new category when the input represents a truly unfamiliar and novel situation. Now, as to Bayesian methods in science: hey, I'm a mathematician, how can I not love Bayes, right?

1:11:04 But the beauty of Bayes is its simplicity. You just write the probability of two events A and B in two different ways. The probability of B given A times the probability of A, the probability of A given B times the probability of B, set them equal, because they're identical, divide by, let's say, probability A and then optimize. That's Bayes. And it's a useful statistical method and should continue to be used in statistics.

1:11:40 But it's just the formal identity wherein lies its power. It says nothing about any physical reality, whether in physics, chemistry or biology. The Bayes' rule itself tells us nothing about physical reality and contains no heuristics to discover anything about physical reality. For that you need to develop models driven by a profound analysis of large databases. So it turns out that biological models like ART do not incorporate the Bayes' rule.

1:12:17 However, ART does routinely choose the best or optimal categories that represent the data best. So you don't need Bayes to achieve optimality. Also, Bayes works best in a stationary world with stationary probabilities and ART is designed to learn about a non-stationary world. So, you know, one can discuss this till the cows come home. It's good for what it was designed for. And some of the neuroscientists who try to apply Bayes are wonderful experimentalists, but they know no math and no theory.

1:13:00 And you know, it's the temptation of a free lunch. There it is waiting to be applied. There is no free lunch in science.

[sp: Rahmjoo]
1:13:19 Thank you. Now, as a final point of comparison, what are the…

[sp: Grossberg]
1:13:23 Oh, Smith! Smith! I didn't reply to Smith.

[sp: Rahmjoo]
1:13:26 Yes! Yes! Thank you!

[sp: Grossberg]
1:13:28 Okay. You quoted a sentence of Smith. But before that sentence Smith et al. (Smith et al., 2020)wrote: "It is also worth highlighting that, as our model is intended primarily as a proof of concept and a demonstration of an available model expansion/reduction approach that can be used within active inference research, it does not explicitly incorporate some aspects—such as top-down attention—that are of clear importance to cognitive learning processes, and that have been implemented in previous models.

1:14:11 For example, the adaptive resonance theory (ART) model

(Grossberg, 1987) was designed to incorporate top-down attentional mechanisms and feedback mechanisms to address a fundamental knowledge acquisition problem—the temporal instability of previously learned information that can occur when a system also remains sufficiently plastic to learn new (and potentially overlapping) information.

1:14:38 While our simulations do not explicitly incorporate these additional complexities, there are clear analogs to the top-down and bottom-up feedback exchange in ART within our model (e.g., the prediction and prediction-error signaling within the neural process theory associated with active inference). ART addresses the temporal instability problem primarily through mechanisms that learn top-down expectancies that guide attention and match them with bottom-up input patterns—which is quite similar to the prior expectations and likelihood mappings used within active inference."

1:15:18 But as I've already noted, "the prior expectations and likelihood matching mappings within adaptive inference" do not have any of the key learning, attention and memory stability properties of the ART matching rule. The ART matching rule is a unique solution to that problem and its variations. It's been supported by psychological, anatomical, physiological and biophysical data. It also occurs in many species.

1:15:54 Nobuo Suga, for example, shows it occurs in bats (Gao & Suga, 1998). It occurs in ferrets, you know. So also, I think it's important to note that when learning begins in an ART model, it doesn't need prior expectations or likelihoods. In fact, typically the initial bottom-up weights are chosen to be random, if you don't know what you're going to be experiencing.

1:16:24 And the initial top-down expectations are chosen to be large, so that whatever category happens to be learned when it reads out its top-down expectation, it can match whatever features activated that category. So they all start large and they prune to match the critical features that happen to be learned in that category. So there are no built-in models. ART discovers its own models.

1:16:58 I should also emphasize that active inference is also not explainable. ART is explainable because a currently active critical feature activity pattern, namely the features to which attention is paid, controls all learning and prediction by the model and, in principle, can be measured by neurophysiological experiments. A model without cell activities or short-term memory traces that can represent the critical feature pattern can't be explainable.

1:17:35 So I think there are qualitative differences. I don't say people shouldn't use active inference. It may be incredibly useful and powerful in technological applications, but when one is doing, you know, brain science, psychology, it just doesn't match the foundational data. It just doesn't. Not a personal thing.

[sp: Rahmjoo]
1:18:08 Thanks. And I guess you somehow already answered part of this question, but what are the possible ways in which ART's approach to explainable AI, which, if I'm not mistaken, can be described as a model-dependent, intrinsically explainable approach, can inform active inference's approach and cross-fertilize with it, which is based on abductive reasoning through constructing generative models, for example, as sketched out in Parr and Pezzulo's "Understanding, Explanation, and Active Inference" paper (Parr & Pezzulo, 2021).

[sp: Grossberg]
1:18:47 Well, first, I don't think ART is model dependent. As I just noted, one begins, typically, to learn in ART with random initial bottom-up weights and uniformly distributed top-down initial adaptive weights, so you can match any category that you happen to learn. But the authors you quoted write in part that active inference… and here I want to quote them, so I can respond in a little more detail.

1:19:25 Active inference "implies a deep generative model that includes a model of the world, used to infer policies, and a higher-level model that attempts to predict which policies will be selected based upon a space of hypothetical (i.e., counterfactual) explanations—and which can subsequently be used to provide (retrospective) explanations about the policies pursued." So, again, ART works without a generative model of the world or any predefined policies. Of course, one is trying to discover what changing world it happens to be in, and then nobody knows what it is a priori.

1:20:14 And in general, an ART classifier responds to a front-end of pre-processes that process perceptual data from one or another sense, notably vision and audition, where we get most of our information about the world. And that's why a classifier, like ART, begins its work in the brain, in the temporal cortex, where it receives a highly pre-processed perceptual representation.

1:20:51 So, decades of work went into understanding how our brains consciously see and hear. And in the case of vision, ART classifies perceptual boundaries and surfaces that require multiple stages of processing, because, as I mentioned briefly, they were the outcome of what I call hierarchical resolution of uncertainty. You need multiple stages to define a perceptual boundary and surface. One reason being, because our sensory organs register such noisy and incomplete data, like, you may know that our photosensitive retina has a huge blind spot where you can't register any visual signal.

1:21:46 The blindspot is as big as the fovea where all of our high-resolution vision occurs. So it's not a little thing. And moreover, veins come out of the fovea and occlude the retina in multiple places. And you can’t register visual signals on the veins either. So, the signal you're getting is very incomplete and it takes multiple processing stages to overcome those uncertainties.

1:22:16 And my colleagues and I have worked for decades to explain how that happens. Maybe I'll stop there for that.

[sp: Rahmjoo]
1:22:28 Thank you so much. Now, this next question is of a personal interest to me, because currently I’m working on modeling some probabilistic aspects of affective response to music and your most recent paper “Toward Understanding the Brain Dynamics of Music” (Grossberg, 2022) immensely helped me gain a better understanding of entrainment. As you pointed out in the supplementary notes for this paper, violation of prior learned expectations is instrumental in inducing a wide range of affective responses in musical and non-musical situations.

1:23:07 Some psychologists, such as Patrik Juslin, have distinguished between perception and arousal of emotions…

[sp: Grossberg]
1:23:12 I’m sorry to interrupt you, but you left out a question. Is it the lack of time where you just skip to the next written question? I wanted to say quite a bit about it.

[sp: Friedman]
1:23:25 Which question?

[sp: Grossberg]
1:23:26 “How do you see the future of ART in neuro-inspired AI?”

[sp: Rahmjoo]
1:23:30 I think that will be our last question.

[sp: Grossberg]
1:23:33 Okay. So we’ll come back to that.

[sp: Rahmjoo]
1:23:35 Yes.

[sp:Friedman]
1:23:36 Yes.

[sp: Rahmjoo]
1:23:36 Thank you.

[sp: Grossberg]
1:23:36 Okay. Because that's an important question for me.

[sp: Rahmjoo]
1:23:39 Yes.

[sp: Grossberg]
1:23:40 Okay. So sorry to interrupt. I just wanted to be sure.

[sp: Rahmjoo]
1:23:43 No problem. Thank you. Now, yes. Well, some psychologists such as Patrick Juslin (Juslin, 2019), have distinguished between the perception and the arousal of emotions in the context of musical experience. And also several studies, such as the works of Juslin and Gabrielson (Juslin et al., 2008, 2011) from the Psychology Department of Uppsala University, have shown that despite music's ability to communicate a wide range of positively and negatively valenced emotions, it somehow evokes mostly positively valenced emotions (Gabrielsson, 2011).

1:24:20 For instance, we can easily perceive rage or anger in music without necessarily getting angry. On the other hand, we're more likely to actually feel elevated and happy after listening to happy music. And also the evidence shows that this disparity between perception and evocation of emotion is probably even more significant in musical experiences than any other non-musical experiences. So how can this difference in diversity between perceived and aroused or evoked musical emotions be accounted for within ART framework?

1:25:01 Can it possibly be regarded as another kind of broken symmetry, as you mention on page 621 of your book, but specific to musical affects?

[sp: Grossberg]
1:25:15 So, as you've noted, I haven't studied these issues in the context of music,

but I'll try to venture some general comments. I should first note that the LAMINART model which is a development of ART to show how and why all neocortical circuits that support perception and cognition typically share a canonical six-layer circuit... My colleagues and I have modeled how, just as in our brains, variations of this canonical laminar circuit can support all perceptual and cognitive processes.

1:26:04 So there's a major generalization of ART, and we've done it for vision, speech, and cognitive working memory, and planning in particular. So, the main point is that the laminar circuitry is basically in all perceptual and cognitive areas, vision, audition, et cetera, et cetera. That's how one can create a context for discussing music. And in fact, my work on music applied such discoveries. I was able to put together discoveries that had been made based on what I believe were different, evolutionary pressures on the organization of our brains.

1:26:56 But that evolution also discovered, and I try to sketch how, if you put some of them together in a certain way, then capacity for learning and consciously performing music could arise. So now how about arousal? Well, it's essential, of course, to all awareness and consciousness. Neocortex needs to be adequately aroused for waking consciousness to occur at all. [Arousal] also plays a major role in the processing of emotions and is very relevant to musical issues.

1:27:41 Of course, my gated dipole model explains how opponent processes, opposites, are organized in all parts of our brain: perceptual, cognitive, motor, affective. In particular, emotions are organized in pairs in such an emotional dipole. And one reason is because emotions need to compete with each other, such as fear versus relief. For example, in post-traumatic stress disorder therapy, a therapist may try to help a patient to think about positive experiences that generate relief in order to inhibit the chronic fear that's so destabilizing during PTSD.

1:28:38 So their opposites are competing. And another property that arousal enables is that this sudden offset of an emotion like here, let's say, during escape behavior, let's say, you know, a favorite example of mine is, you know, some cruel experimentalist puts the pigeon in a Skinner box, the floor is electrified, the pigeon is feeling pain and fear, it's dashing frantically around, trying to keep its feet off the floor.

1:29:13 It bangs into a red buzzer, the buzzer shuts the shock off, and the animal experiences a wave of relief or positive motivation for learning the escape response. So this rebound from fear to relief, that can be associated with actions that lead to escape and can motivate escape in the future, is energized by arousal in the gated dipole. You shut off the external cue of shock, but the arousal is tonically on, or has sustained activity in both the fear and relief channels.

1:29:58 So because the arousal is sustained or tonic through time and equally activates the fear and relief channels when fear suddenly decreases, then arousal and the relief channel wins the competition, and can thereby call what I call an antagonistic rebound from fear to relief that activates the relief channel and thereby providing motivation for escape, whether reactive or learned through escape experiences.

1:30:33 And I also proved, which is related to some degree to music, that the non-occurrence of an expected event can by itself cause a burst of arousal and thus an antagonistic rebound and can flip emotions from positive to negative in so doing. And I always love that discovery, because I especially love discoveries where the occurrence of nothing has profound effects on future behavior.

1:31:09 So it's the non-occurrence of the expectation because of this mismatch that can flip emotions. Now how this influences the perception of music needs more work. It's… as I mentioned my paper, I haven't tried to study that. My paper on music, I feel, is like, you know, a drop in the bucket. And, hopefully, if I don't get around to it, someone else will.

1:31:43 But the above examples show that arousal and emotion are not the same thing, because arousal can support all emotions: fear, relief, hunger, satiety, whatever. The ones that win the competition are then able to support compatible behaviors by motivating them. And I've also explained that the level of arousal must be chosen within an intermediate range to support normal behaviors.

1:32:15 It's the kind of golden mean. There's an inverted-U on the effects of having arousal from too little or too much. And if you have too little arousal, you have an underaroused syndrome which can support symptoms like autism. And overarousal can support symptoms of a disease like schizophrenia. That’s only one of many factors in these diseases. But I'm happy to say that subsequent clinical data supported those predictions that these two mental disorders are at opposite ends of the arousal inverted-U.

1:33:04 So I think… and you know, this is very speculative, because I've never really seriously studied it and I try not to speculate, but what the hell! The kind of arousal that music activates is generally positive, just like the arousal that activates exploratory behaviors is positive. It’s somehow linked. You know, music is a sonic adventure, if you like. There's no aversive cue when listening to music, except perhaps music that’s played so loud as to cause a headache or ear damage or even a seizure in susceptible individuals.

1:33:49 So there's no particular reason why it shouldn't be positive. So now which question do you want to ask me Ali?

[sp: Rahmjoo]
1:34:04 I think we're left with just two other questions if you're not too tired.

[sp: Grossberg]
1:34:11 One starts with “as a final point of comparison…” another starts with “on a more philosophical note…”

[sp: Rahmjoo]
1:34:17 Yes, we’re onto the last two questions.

[sp: Grossberg]
1:34:23 So do you want to do “how do you see the future of ART and brain-inspired AI…” first?

[sp: Rahmjoo]
1:34:29 And before that… Yes, before that I just wanted to ask you about your view about artificial consciousness. Do you see…

[sp: Grossberg]
1:34:41 I really need to first answer the second question.

[sp: Rahmjoo]
1:34:45 Okay. As you wish. Yes. Okay. Yeah. And how do you see… Yes, it's quite fine. And how do you see the future of ART and brain-inspired AI research in general? In your view, what research areas ought to gain more attention than they do today?

[sp: Grossberg]
1:35:10 So I'll give you quite a general answer, but it implies what I think about this. So first, with a caveat, I like to say I couldn't predict the present, so I can't predict the future. That being said, I believe that all engineering technology and AI will increasingly embody autonomous adaptive intelligence in the coming century. And we can already see its beginnings in autonomous automobiles and airplanes and increasingly autonomous controllers on the factory floor.

1:35:52 And many people have written about it. And I think ART will play a central role in this, as well as other models that are summarized in my book. And that's because already in 1980, I published a thought experiment in the journal *Psychological Review (Grossberg, 1982)*, which was then and still probably remains the leading theory journal in psychology. And you may recall that Einstein derived both special relativity theory and general relativity from thought experiments.

1:36:31 And let me just clarify it by my thought experiment, wherein they derive their enormous power. So my thought experiment was about how any system can autonomously correct predictive errors in a changing world. And the hypotheses upon which the thought experiment were derived were just a few facts that are familiar to us all from daily life. And they are familiar, because they represent ubiquitous environmental pressures on the evolution of our brains over the millennia.

1:37:13 And when they act together, I suggest ART is the unique outcome. That's a huge claim. And I turn to the power of the thought experiment, not to any personal ego trip with that belief. In particular, nowhere in the thought experiment are the words "mind" or "brain" mentioned. So if you accept that these facts about the world exist, which we all do, and that they're always operating on us, then you have to accept the outcome if you believe in the scientific method and logic.

1:37:55 So ART is a universal solution to the problem of autonomous error correction in a changing world. Say it in another way, if you can't find a mistake in the thought experiment, then I think you either have to believe in ART-like dynamics, maybe expressed in LAMINART or your favorite ART variant, or you have to give up your belief in logic and the scientific method.

1:38:26 It doesn't imply that ART can't be further developed. I expect a large number of scientists and technologists to be busy developing ART-like architectures long after I'm gone. So maybe now we can go to your philosophical note with that background.

[sp: Rahmjoo]
1:38:46 Yes. Thank you so much. Now, yes, as I mentioned earlier, I just wanted to ask about your view on artificial consciousness. Do you see consciousness as artificially producible or engineerable (as some researchers like Mark Solms believe (Solms, 2021))? Is there a fundamental distinction between a biologically conscious agent and an artificial agent with a fully-simulated computational model of consciousness?

1:39:19 I know it's a big, big question, but I couldn't resist asking your opinion as an authority on consciousness modeling.

[sp: Grossberg]
1:39:27 I'm happy to give it a shot. So as I just noted, my work on ART suggests itself the universal problem about how we can learn to correct predictive errors in a changing world. My work also shows in its analysis of hierarchical resolution of uncertainty (remember, like, how you go from somewhat noisy retina to the surface representation that can control looking and reaching) how evolution may have been driven to discover conscious states.

1:40:09 So this was a surprise to me too. Conscious states were needed in order to choose that processing level, the level that computes the sufficiently complete context-sensitive and stable representation—in the case of vision, the surface representation—with which to successfully plan and act to realize valued goals. So let me make it clear. So you start with a noisy retina and you have to go up all of these stages until you get a sufficiently complete surface and boundary representation that you can use to regulate successful action.

1:40:55 And if you used one of the earlier stages, it would lead to incorrect actions which would kill you off by Darwinian selection. So how the hell do you know where the stage is, where you can compute this sufficiently complete and context-sensitive and stable one? And I propose… In vision, I predicted that the choice is embodied in what I call a surface-shroud resonance between prestriate visual cortical area V4 and the next processing stage, posterior parietal cortex or PPC.

1:41:38 So it's in V4 [where] you get this really good surface representation. And then a resonance between the surface and spatial attention which fits the surface. That spatial attention in PPC is called a shroud (Christopher Tyler gave it that name (Tyler & Kontsevich, 1995)). A surface-shroud resonance allows you to pay conscious spatial attention to the surface that you're going to use to control looking and reaching behaviors.

1:42:17 So it's a way of ensuring you have a good enough representation to control action. So the shroud is computed in [the] posterior parietal cortex, which is part of the dorsal or “where” cortical stream. And the shroud modulates invariant category learning in the ventral or “what” cortical stream. I can't go into that right now, but my book discusses it. The category learning itself in the “what” cortical stream, as I indicated, is supported by a feature-category resonance.

1:43:00 And so the surface-shroud resonance is modulating invariant category learning in the feature-category resonances. Surface-shroud resonance also supports conscious seeing, the feature-category resonance is supporting conscious recognition, and when they synchronize across streams on a familiar object, that's when you consciously see something that you know about. Okay, so conscious states hereby arise due to learning requirements.

1:43:42 This is sort of fell-out-of-the-wash of how you do invariant category learning, and learning, in particular, without catastrophic forgetting. It's regulating feature-category resonances. So given that the above solution is computationally universal in the sense I sketched, the self-organizing machine that embodies them should be able to support internal representations whose parametric properties mimic conscious states. So whether such a machine can experience conscious qualia remains as much of a mystery for machines as for humans.

1:44:33 And that's because no computational theory, which after all, is just a set of equations, can do more than imitate the dynamics of our brains, perhaps with great precision. I don't have a clue why the representations that my colleagues and I have worked so hard to explain huge amounts of psychophysical data about seeing: texture, shading, 3D form, just go through the list. Why do they support qualia?

1:45:12 I don't know. Ask God, or whatever God you choose to believe in in the 21st century.

[sp: Rahmjoo]
1:45:25 Thank you so much, Professor. I think we have a couple of questions in the chat, but we are actually approaching our two-hour limit. I don't know, Daniel, if it's a good place to stop or whatever you say.

[sp: Friedman]
1:45:43 I think that's a great place to stop. You've given us a lot to think about and digest. And I hope that these words are taken well and paid attention to; might create some categories, activate some categories. But Professor Grossberg, thanks again for this amazing livestream. We really appreciate it.

[sp: Grossberg]
1:46:04 Well, I appreciate it and I'm depending on younger people like yourself to do just what you said, Daniel. I'm not going to be around that much longer. So I hope you have, whether with my work directly or related work, you have a very fulfilling intellectual adventure. I know I've been on a wild ride since I was 17, and that's 65 years of discovery. I loved every minute of it.

Works Cited.

Amari, S.-I. (1972). Characteristics of Random Nets of Analog Neuron-Like Elements. *IEEE Transactions on Systems, Man, and Cybernetics*, *SMC-2*(5), 643–657. https://doi.org/10.1109/TSMC.1972.4309193

Badia, A. (2019). *The Information Manifold: Why Computers Can't Solve Algorithmic Bias and Fake News*. MIT Press. https://play.google.com/store/books/details?id=Ncq2DwAAQBAJ

Banquet, J. P., & Grossberg, S. (1987). Probing cognitive processes through the structure of event-related potentials during learning: an experimental and theoretical analysis. *Applied Optics*, *26*(23), 4931–4946. https://doi.org/10.1364/AO.26.004931

Brincat, S. L., & Miller, E. K. (2015). Frequency-specific hippocampal-prefrontal interactions during associative learning. *Nature Neuroscience*, *18*(4), 576–581. https://doi.org/10.1038/nn.3954

Brito da Silva, L. E., Elnabarawy, I., & Wunsch, D. C., 2nd. (2019). A survey of adaptive resonance theory neural network models for engineering applications. *Neural Networks: The Official Journal of the International Neural Network Society*, *120*, 167–203. https://doi.org/10.1016/j.neunet.2019.09.012

Carpenter, G. A. (1989). Neural network models for pattern recognition and associative memory. *Neural Networks: The Official Journal of the International Neural Network Society*, *2*(4), 243–257. https://doi.org/10.1016/0893-6080(89)90035-X

Carpenter, G. A., Gopal, S., Macomber, S., Martens, S., & Woodcock, C. E. (1999). A

Neural Network Method for Mixture Estimation for Vegetation Mapping. *Remote Sensing of Environment*, *70*(2), 138–152. https://doi.org/10.1016/S0034-4257(99)00027-9

Carpenter, G. A., & Grossberg, S. (1987). A massively parallel architecture for a self-organizing neural pattern recognition machine. *Computer Vision, Graphics, and Image Processing*, *37*(1), 54–115. https://doi.org/10.1016/S0734-189X(87)80014-2

Carpenter, G. A., & Grossberg, S. (1988). The ART of adaptive pattern recognition by a self-organizing neural network. *Computer*, *21*(3), 77–88. https://doi.org/10.1109/2.33

CNS Tech Lab. (n.d.). Retrieved December 12, 2022, from http://techlab.bu.edu/

Explainable Artificial Intelligence and Neuroscience: Cross-disciplinary Perspectives. (n.d.). *Frontiers in Neurorobotics*. https://www.frontiersin.org/research-topics/11553/explainable-artificial-intelligence-and-neuroscience-cross-disciplinary-perspectives

Felleman, D. J., & Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex* , *1*(1), 1–47. https://doi.org/10.1093/cercor/1.1.1-a

Gabrielsson, A. (2011). *Strong Experiences with Music: Music is much more than just music*. https://doi.org/10.1093/acprof:oso/9780199695225.001.0001

Gao, E., & Suga, N. (1998). Experience-dependent corticofugal adjustment of midbrain frequency map in bat auditory system. *Proceedings of the National Academy of Sciences of the United States of America*, *95*(21), 12663–12670. https://doi.org/10.1073/pnas.95.21.12663

Gianfagna, L., & Di Cecco, A. (n.d.). *Explainable AI with Python*. Springer International Publishing. https://doi.org/10.1007/978-3-030-68640-6

Grossberg, S. (1976). Adaptive pattern classification and universal recoding: II. Feedback, expectation, olfaction, illusions. *Biological Cybernetics*, *23*(4), 187–202. https://doi.org/10.1007/BF00340335

Grossberg, S. (1982). How Does a Brain Build a Cognitive Code? In S. Grossberg (Ed.), *Studies of Mind and Brain: Neural Principles of Learning, Perception, Development, Cognition, and Motor Control* (pp. 1–52). Springer Netherlands. https://doi.org/10.1007/978-94-009-7758-7_1

Grossberg, S. (1988). Nonlinear neural networks: Principles, mechanisms, and architectures. *Neural Networks: The Official Journal of the International Neural Network Society*, *1*(1), 17–61. https://doi.org/10.1016/0893-6080(88)90021-4

Grossberg, S. (1999). How does the cerebral cortex work? Learning, attention, and grouping by the laminar circuits of visual cortex. *Spatial Vision*, *12*(2), 163–185. https://doi.org/10.1163/156856899x00102

Grossberg, S. (2018). Desirability, availability, credit assignment, category learning, and attention: Cognitive-emotional and working memory dynamics of orbitofrontal, ventrolateral, and dorsolateral prefrontal cortices. *Brain and Neuroscience Advances*, *2*, 2398212818772179. https://doi.org/10.1177/2398212818772179

Grossberg, S. (2020). A Path Toward Explainable AI and Autonomous Adaptive Intelligence: Deep Learning, Adaptive Resonance, and Models of Perception, Emotion, and Action. *Frontiers in Neurorobotics*, *14*, 36. https://doi.org/10.3389/fnbot.2020.00036

Grossberg, S. (2021). *Conscious Mind, Resonant Brain: How Each Brain Makes a Mind*. Oxford University Press. https://play.google.com/store/books/details?id=8hQuEAAAQBAJ

Grossberg, S. (2022). Toward Understanding the Brain Dynamics of Music: Learning and Conscious Performance of Lyrics and Melodies With Variable Rhythms and Beats. *Frontiers in Systems Neuroscience*, *16*, 766239. https://doi.org/10.3389/fnsys.2022.766239

Grossberg, S., & Versace, M. (2008). Spikes, synchrony, and attentive learning by laminar thalamocortical circuits. *Brain Research*, *1218*, 278–312. https://doi.org/10.1016/j.brainres.2008.04.024

Juslin, P. N. (2019). *Musical Emotions Explained: Unlocking the Secrets of Musical Affect*. https://doi.org/10.1093/oso/9780198753421.001.0001

Juslin, P. N., Liljeström, S., Laukka, P., Västfjäll, D., & Lundqvist, L.-O. (2011). Emotional reactions to music in a nationally representative sample of Swedish adults: Prevalence and causal influences. *Musicae Scientiae: The Journal of the European Society for the Cognitive Sciences of Music*, *15*(2), 174–207. https://doi.org/10.1177/1029864911401169

Juslin, P. N., Liljeström, S., Västfjäll, D., Barradas, G., & Silva, A. (2008). An experience sampling study of emotional reactions to music: listener, music, and situation. *Emotion* , *8*(5), 668–683. https://doi.org/10.1037/a0013505

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., & Hadsell, R. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(13), 3521–3526. https://doi.org/10.1073/pnas.1611835114

LeVine, S. (2017, December 15). *Artificial intelligence pioneer says we need to start over*. https://www.axios.com/2017/12/15/artificial-intelligence-pioneer-says-we-need-to-start-over-1513305524

Miller, E. K., Li, L., & Desimone, R. (1991). A neural mechanism for working and recognition memory in inferior temporal cortex. *Science*, *254*(5036), 1377–1379. https://doi.org/10.1126/science.1962197

Neural Networks. (n.d.). First Issue of Journal, Neural Networks. Retrieved December 12, 2022, from https://www.sciencedirect.com/journal/neural-networks/vol/1/issue/1

Neural Networks: special Issue in Honor of the 80th Birthday of Stephen Grossberg. (n.d.). Retrieved December 12, 2022, from https://www.sciencedirect.com/journal/neural-networks/vol/120/suppl/C

Parker, D. B., & Sloan School of Management. (1985). *Learning logic : casting the cortex of the human brain in silicon*. Alfred P. Sloan School of Management, Massachusetts Institute of Technology. https://www.worldcat.org/title/learning-logic-casting-the-cortex-of-the-human-brain-in-silicon/oclc/17489314

Parr, T., & Pezzulo, G. (2021). Understanding, Explanation, and Active Inference. *Frontiers in Systems Neuroscience*, *15*, 772641. https://doi.org/10.3389/fnsys.2021.772641

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*(6), 386–408. https://doi.org/10.1037/h0042519

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533–536. https://doi.org/10.1038/323533a0

Schmidhuber, J. (2020). *Critique of Honda Prize for Dr. Hinton*. https://people.idsia.ch/~juergen/critique-honda-prize-hinton.html

Smith, R., Schwartenbeck, P., Parr, T., & Friston, K. J. (2020). An Active Inference Approach to Modeling Structure Learning: Concept Learning as an Example Case. In *Front. Comput. Neurosci.* (p. 633677). https://doi.org/10.3389/fncom.2020.00041

Solms, M. (2021). *The hidden spring: A journey to the source of consciousness*. Profile books. https://profilebooks.com/wp-content/uploads/wpallimport/files/PDFs/9781788162838_preview.pdf

Spitzer, H., Desimone, R., & Moran, J. (1988). Increased attention enhances both behavioral and neuronal performance. *Science*, *240*(4850), 338–340. https://doi.org/10.1126/science.3353728

Stephen Grossberg. (n.d.). Retrieved December 12, 2022, from https://sites.bu.edu/steveg/

Tyler, C. W., & Kontsevich, L. L. (1995). Mechanisms of stereoscopic processing: stereoattention and surface perception in depth reconstruction. *Perception*, *24*(2), 127–153. https://doi.org/10.1068/p240127

Werbos, P. J. (1974). *Beyond Regression: New Tools for Prediction and Analysis in the*

Behavioral Science. Thesis (Ph. D.). Appl. Math. Harvard University. http://dx.doi.org/

Wunsch, D. C., II. (2019). Admiring the Great Mountain: A Celebration Special Issue in Honor of Stephen Grossberg’s 80th Birthday. *Neural Networks: The Official Journal of the International Neural Network Society*, *120*, 1–4. https://doi.org/10.1016/j.neunet.2019.09.015