An Algorithm for Thought (v1.0) to Help Solve Problems of the Anthropocene

Carl O. Pabo, Ph.D.

Posted November 2022

www.carlpabo.com

Table of Contents

Executive Summary

Society desperately needs to find more effective ways to address the critical problems of the Anthropocene, problems like global warming and environmental degradation. Challenges like these are so complex that it's easy to get confused, or discouraged, when trying to develop potential solutions. New methods of thought are needed to work effectively amidst this complexity, and at Humanity 2050 we are testing a strategy that has two basic steps:

1) We use "special focus teams" to ensure that team members have the time and attention needed for careful thought (Pabo, 2021; and p. 5 of current document).

2) An "algorithm for thought" (offered here) provides a way of transforming the overall problem of planning (as when developing a plan for climate engineering) into a series of tens of thousands of smaller, more manageable little puzzles. (In this way, our algorithm is somewhat like a computer program that works by breaking a larger problem into a long series of smaller, more manageable steps.)

The algorithm described in this white paper emerges only after a careful consideration of the power and limits of human thought. Since one must understand a tool before one knows how to use it most wisely, analysis here builds upon a new model of thought developed by Humanity 2050 founder Dr. Carl O. Pabo. This model of thought begins by noting that ideas can only exist as specialized structures or specialized physical events in a physical world. Ideas must be embodied in brains, in sound or electromagnetic waves, or in computers and books. And, when assembling or explaining a plan, ideas must (physically) move from one mind to another via processes that are limited by the constraints of working memory.

Working memory also plays an important role in the process of thought, and limits of working memory capacity (Cowan, 2005) end up affecting the way in which the human mind can develop new ideas. These limits become painfully clear when trying to develop or explain a plan of great complexity, such as how we might enact a carbon tax or a climate engineering program. Long-term memory allows for more complex ideas, yet human cognition also has profound limits here. New ideas — with appropriate, nuanced complexity — develop only through a slow process of deep and deliberate thought (with ideas emerging and developing over time as the connections in the brain gradually change).

Our new algorithm for thought shows how problems of planning for the human future can be approached more effectively once they are seen as problems of "constraint satisfaction." Society needs plans that are consistent with many different constraints — constraints based on our knowledge about physics, chemistry, biology, and human behavior; constraints on the budget, on the desired outcomes, on the acceptable level of risk, on the economic and social implications, and so forth. With so many constraints, this overall problem (when planning how to address problems of the Anthropocene) becomes far too complicated for the human mind to see everything at once, so we have set up an algorithm — working via iterative optimization — that breaks the overall problem into pieces small enough so as to avoid overwhelming human cognitive capacity.

This new algorithm also helps ensure that each team member can contribute efficiently and effectively to the team effort. It should allow these special focus teams to work carefully enough so as to develop plans that are clear, actionable, acceptable to society, and powerful enough to help solve these problems of the Anthropocene.

There is nothing easy here: Implementing this approach will require immense dedication and effort from team members (who tackle problems at least as hard as those faced by Charles Darwin when he struggled to understand the origin of species). And this method will only work effectively if each member of the team develops — then constantly revises and extends — a large, neurophysiologically encoded corpus of information and ideas that can be used as a backdrop when considering each aspect of the problem.

We'll be testing this strategy at Humanity 2050 with our work on climate engineering, yet this new algorithm for thought can be applied to many other complex challenges of the Anthropocene. It should help humans survive in a world where complexity otherwise threatens to overwhelm the capacity of the human mind.

Note: This algorithm — as offered here in version 1.0 — cannot reduce every step in the analysis to a simple numerical basis (as is possible in the classic "traveling salesman problem"), and thus will not work with the kind of mathematical precision that a computer scientist might expect after studying algorithms for combinatorial optimization (Papadimitriou and Steiglitz, 1998). Yet the algorithm sets out the full challenge of effective planning in a clear, organized way. It should help ensure a level of care that is missing from most discussions about how to address challenges of the Anthropocene, and the algorithm itself will be amenable to continued processes of optimization and improvement over the years.

Introduction

I. How Special Focus Teams Will Benefit from This Algorithm for Thought

Society is facing a crisis of complexity as it tries to find effective ways to address the challenges of the Anthropocene. Our work at Humanity 2050 aims to facilitate work amidst this complexity, helping society develop better action plans to address key challenges of the human future, and we work in a way that combines two novel strategies:

> 1) We use special focus teams (which were described in depth in a previous white paper and are summarized below).
>
> 2) We show that the problem of planning really is a problem of "constraint satisfaction" (with a list of criteria that must be satisfied so as to optimize expected utility of the resulting plans), and we then offer a new "algorithm for thought" that gives a more systematic way of solving this kind of problem.

Our previous white paper ("Special Focus Teams" to Help Solve the Problems of the Anthropocene) describes our strategy of using new, carefully organized teams to help society manage this complexity. These small teams play a critical role in our work at Humanity 2050. Each team will be dedicated to developing an action plan to address one of the complex challenges of the Anthropocene, and we're now testing this approach with a team that is exploring prospects for climate engineering. This very talented team will build on the work of others — analyzing, integrating, and extending ideas about climate engineering so as to help evaluate the challenges and risks that will arise if society tries such a program (e.g., spraying aerosols in the stratosphere to reflect some of the incoming sunlight and thus to help cool the Earth). More generally, such teams should be able to think carefully enough, and work hard enough, so as to develop powerful, acceptable, actionable plans to address many of the complex challenges of the Anthropocene.

As emphasized in our previous paper, team members will need to have the kind of intelligence, time, and focus necessary for careful thought. And there should be nothing surprising about this need for new ways of focusing thought: Global challenges of the Anthropocene involve a mind-boggling complexity. Planning requires that we try to look decades ahead and foresee what will happen in a rapidly changing world with billions of people and billions of computers. Planning also is hard because we must consider a pan-disciplinary range of issues — including relevant aspects of science, technology, economics, local and national politics, international relations, morality, law,

misinformation and disinformation, cybercrime, fraud, corruption, and warfare. (It makes no sense to have plans based on wishful thinking about how people "should" be. We need plans that can work with people as they are.)

When approaching these challenges, special focus teams will benefit from having our new algorithm for thought (summarized in Section VII below). Our algorithm will help improve the overall efficacy of work at Humanity 2050 and — likewise — will help other groups assemble better action-oriented plans for addressing the complex challenges of the Anthropocene.

II. How Algorithms for Thought Can Help People Think More Clearly

Our development of this algorithm for thought builds on the observation that people can learn rules that affect the way in which subsequent ideas are processed. We see this, for example, with the kind of algorithm that a grade school teacher offers students when teaching them how to do long addition. The teacher does not explain the neurophysiological constraints (involving working memory limits) that make it hard for students to track more than a few digits and a few operations at a time, but students are taught how to line up the numbers and start by adding digits in the right-hand column, how to carry numbers as necessary, etc. Students are — in essence — able to learn an algorithm that facilitates subsequent patterns of thought, as they move on to subtraction, multiplication, etc. And more complex examples of such stepwise patterns of reasoning are seen in classroom applications of Euclidean geometry and in many other areas of mathematics. Such algorithms offer well-tested, well-proven methods that allow students to get the correct answers more quickly and more reliably than otherwise possible.

Although rarely taught in the same way in a formal classroom setting, there also are well-accepted rules about the best ways to study. Students are advised, for example, to minimize distractions when studying, keep up to date with assignments, get enough sleep, and avoid the need for "cramming" the night before a test. Again, such rules are rarely explained in neurophysiological terms, but these strategies ensure that, day after day, important new ideas from the lectures, readings, and problem sets are reliably and stably transferred to long-term memory. They help ensure that underlying patterns of neural/synaptic activity can be upgraded so as to represent these new ideas.

Humanity 2050 now leverages these human abilities to "think about thought" and "learn how to learn" as we offer new strategies to help deal with the challenges of the Anthropocene.

Part I — A New Model of Thought

III. Key Features of a New Model of Thought

Our description of this new algorithm for thought proceeds in two stages. Part I of this manuscript offers a new model of thought.[1] This model uses a few simple diagrams to explain key neurophysiological steps and limits involved in the process of thought, and Part I helps us understand how the complexity of the challenges of the Anthropocene force the mind to work at the very limits of human cognitive capacity. It explains why society will need systematic approaches to make the best use of human thought, and to avoid pitfalls and errors that are possible amidst the mind-boggling complexity of the modern world.

Part II offers our new algorithm for thought and shows how this algorithm lets the mind be used in a more efficient, effective way when planning how to address the challenges of the Anthropocene.

* * * * *

This model emerged from a very personal struggle with concerns about human thought and the human future. I resigned my position as a tenured professor of biophysics at MIT because I wanted time to think about some aspects of human thought that did not seem to be adequately addressed by any existing theories or models. It took time — during my first few years of work on this project — to decide how to set up my own new models. Yet one point about ongoing advances in human knowledge always seemed clear: Deep progress — as in physics, biology, and chemistry — usually depends on being able to relate the phenomena of interest to some underlying flow of physical events.

And furthermore: Although we don't always need to think about the world this way,[2] science has shown that everything on Earth occurs via an ongoing flow of atomic/molecular events. (Or, trying to be a bit more precise so as to account for the

[1] Readers who are impatient (who want to see the conclusion before understanding the full neurophysiological context and background) can get a quick overview by jumping ahead to Part II before returning to read the full paper.

[2] For practical purposes, we usually want the simplest description or model that will meet our current needs. This often leads us to gloss over such fine-grained details, but — every time we zoom in and turn up the magnification and look at the world — the atoms and molecules are always there. We then see the water molecules in the winds of a hurricane, the atoms in a bar of gold, the glucose molecules feeding our nerve cells and powering the process of "thought."

role of subatomic particles, photons, other force fields, etc., we might say: Everything on Earth arises via a fine-grained flow of physical events and physical forces.)

The Physical Basis of Thought: Given this underlying, fine-grained physical structure of the world, I set up a model of thought that begins with a very simple, unassailable physical assertion: *Ideas always, and only, arise and exist as some type of specialized structure and/or some specialized type of physical event in a physical world.* Sometimes they are embodied in brains, sometimes in sound waves or electromagnetic waves that move through the air, sometimes in computers, and sometimes in books, as suggested in the figure below:

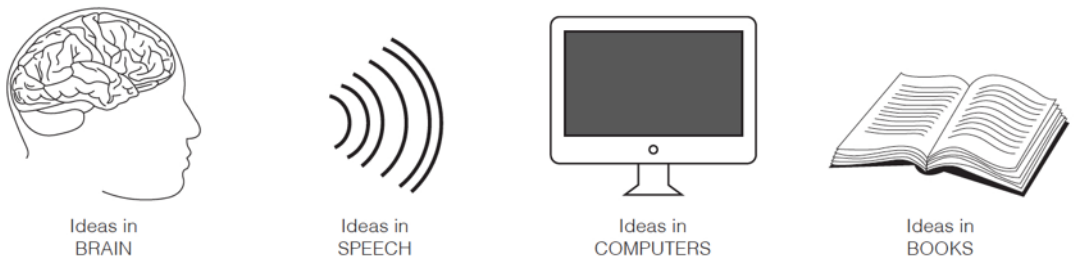

Figure 1

In short: Ideas only exist as embodied in some underlying order, as patterns within the structure of matter or the motion of waves. Unless another copy exists elsewhere, ideas will be lost if one scrambles or destroys connections in the brain (as in traumatic brain injury or dementia), disrupts the circuits in a computer, or burns the pages in a book.

For current purposes, we consider thought as a physical process that creates, alters, or upgrades the types of specialized, highly ordered patterns that arise in settings like those shown in Figure 1. And, in our work at Humanity 2050, we're interested in creating well-ordered, physically embodied patterns (new documents, representing new ideas) that will give society better ways to address these complex challenges.

Note: When we start with a physical perspective as in Figure 1, we are adopting essentially the same kind of initial stance that a neuroscientist might adopt. However, we proceed from there in a radically different way. We are not trying to compete with or replace any detailed models from neuroscience. We want a model of thought that is physically grounded and reliable, yet is simple enough to allow a succinct explanation of the power, the limits, and the mechanistic foundations of human thought.

This physical view (that ideas are always/only embodied in a physical form) naturally leads to fascinating questions about how ideas develop and how ideas move from one place to another. Taking the case of speech, for example, we consider the flow of physical events that is involved as ideas moved from one mind to the next as in Figure 2 below.[3] Even under the best circumstances (with both speaker and listener paying careful attention), we see that abstract ideas typically move via a relatively simple, serially ordered stream of words and symbols,[4] as suggested in the diagram below:

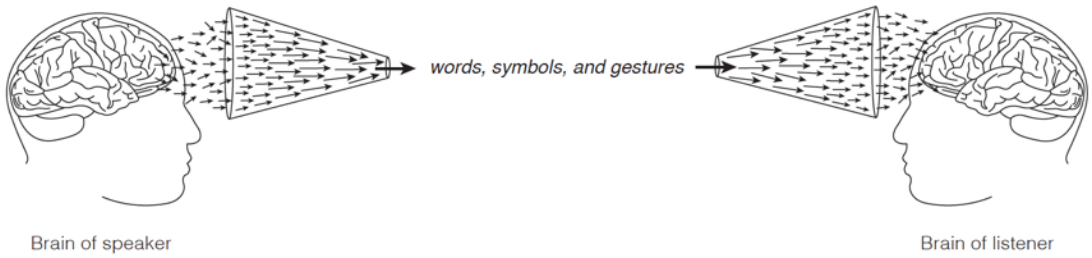

Figure 2

At first glance, this pattern of information transfer seems so obvious, so well known, that one might wonder why it deserves any special discussion, but limits of human working memory capacity put shocking constraints on the complexity of the ideas that can readily move from one mind to the next in the moment of speech.

Working Memory Constraints: When words come out in a linear stream, the listener needs some way to temporarily store or hold the concepts while waiting for the rest of the sentence. This helps ensure that the sentence gets interpreted correctly, revealing the full meaning and implications, but human working memory — this kind of temporary storage facility — has a characteristic, and very limited, storage capacity. Humans are stuck with an information transfer system that 1) relies on existing patterns

[3] Each of the diagrams used to explain our model of thought (Figures 2-7) can be seen as a way of blocking off a region of space-time (here in Figure 2, a region of space containing the two brains and an interval of time in which a sentence is spoken). Science tells us that everything that occurs in these regions of space, over these intervals of time, occurs at an atomic/molecular level. Yet, setting details aside in these black box models lets us 1) maintain a clear physical frame of reference without 2) feeling a need (as a neurobiologist might) to get involved in trying to track all the biophysical details. This makes our model simple enough to keep it "in mind" even when we're trying to understand how to make best use of the mind/brain as we try to find better ways to address the challenges of the Anthropocene.

[4] Gestures, as noted in Figure 2, can also play an important role in human communication, but their role in information transfer becomes relatively less important as ideas become more abstract and more complicated — as in the theories of physics and chemistry or as with detailed plans needed to address the challenges of the Anthropocene.

of neural activity in the brain of the listener/reader (i.e., relies on neural patterns representing some initial set of concepts) and then uses the new input (as from the teacher) to help 2) tie these concepts together in new ways (Cowan, 2005). Working memory can only handle a few basic ideas (a few "chunks") in any moment.

These working memory constraints mean that the speaker — trying to impart some new model or idea to the listener — can only work with a few concepts in any moment (as when the grade school teacher tells the class that "2" plus "3" equals "5," or when the physics professor says that "force" equals "mass" times "acceleration"). In everyday life, we have little reason to think about all the neurophysiological events that must be occurring in the background here (as existing ideas get woven together into larger patterns). Yet, there are important constraints. If we want the listener/reader to be able to make a reliable copy of some new information, we can only connect a few ideas at a time. There's a 4-chunk limit,[5] as suggested in the diagram below, that controls the way in which abstract ideas — like a model, an equation, or a plan — can move from mind to mind.

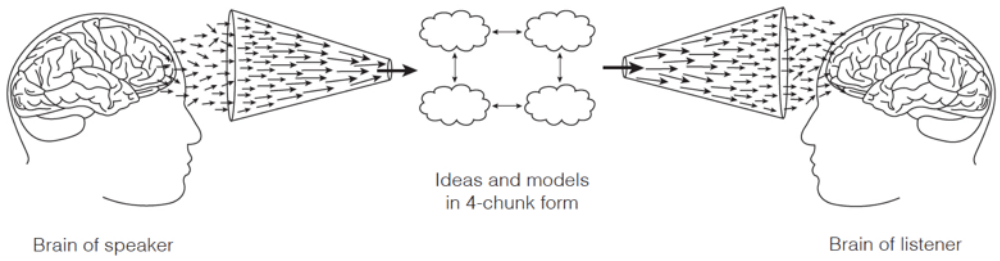

Figure 3

Note: When working memory tasks just involve a linear string of digits, we may be able to remember a seven-digit telephone number as we walk across the room to dial (Miller, 1956). But the challenge is harder when we need to keep track of the relationship between concepts — not just remember a string of numbers — and it thus turns out that people can only handle about four chunks at a time.

[5] Not even neurobiologists can describe all the neurophysiological details involved here, but we can picture a "chunk" as comprising some pattern of neural activity — so firmly established via synaptic connections — that the brain can handle it as if it were a reliable, fixed unit. When learning a new rule, we can only handle a few such chunks at a time, but continued careful study allows the brain to link such pieces together to develop larger, stable patterns of neural activity. Thus, at later stages of study, there will be new chunks representing the fact that "2 + 3 = 5" and the fact that "F = ma". These new chunks will let the student solve new, more complex problems. In general, education does not let the mind somehow "juggle more balls"; rather, it "glues balls together," creating new clusters that can be juggled almost as easily as if each cluster were a single (composite) object.

Note also: There's a rough correspondence between this 4-chunk limit and the amount of information that can usefully be encoded in a single declarative sentence (or a single line of computer code). We break ideas into sentences — and are advised to keep things concise — because 1) well-organized sentences give a way of signaling which concepts are most closely connected and 2) the punctuation gives the reader/listener a chance to pause and absorb one 4-chunk unit before the next such unit is offered.[6]

Given these cognitive constraints of working memory, one can start to see why the human mind has so much trouble developing useful responses to the complex challenges of the Anthropocene. A few 4-chunk frames and a few simple declarative statements may be sufficient as one tries to summarize the problems, and thus, for example, scientists now know that "burning fossil fuels causes global warming."[7]

Yet this kind of simple, 4-chunk level (by itself) will never be sufficient when trying to develop a meaningful way of solving a problem like that of climate change. It may be tempting to respond with some simple idea about a "solution," perhaps saying: "We'll need to tax fossil fuels to bring this problem under control." Yet no one should confuse this broad-brush assertion with a meaningful, actionable plan.[8] No one has yet figured

---

[6] Literature (as with Molly Bloom's soliloquy at the end of James Joyce's *Ulysses*) is full of examples that seem to violate the 4-chunk limit introduced in Figure 3. (Her stream of consciousness soliloquy has about 22,000 words and lacks any normal sentence structure.) However, literature serves a fundamentally different purpose than the type of writing needed to develop and share some clearly defined plans for the human future. Reading Joyce can be a wonderful experience, but it does not ensure (as we need to do as we plan for the human future) that some special pattern of well-ordered conceptual relationships gets transferred from one mind to the next with an (essentially) error-free mode of transmission.

[7] Of course, this kind of simple summary is possible only because of decades of prior research (now summarized in thousands and thousands of pages in reports of the Intergovernmental Panel on Climate Change (IPCC) (e.g., IPCC, 2022). We can have confidence in this conclusion — that burning fossil fuels causes global warming — only because it is consistent with an immense body of underlying data.

[8] Actually, both 1) the analysis of problems of the Anthropocene (as with climate change) and 2) the process of planning some adequate response to these challenges each have stages when immense detail is needed, and each have stages when a simple summary is adequate. Initial studies of climate change started (historically) with some early warnings that "CO_2 may cause serious problems" (Hansen, 1981). This then expanded into a huge body of research and data (as seen in the IPCC reports mentioned in the previous footnote). Only later — after years of careful study — could these observations be compressed to give a simple, reliable summary about the danger of fossil fuels. Likewise, economists may now have some idea about the potential utility of a carbon tax. However, as emphasized in the text, this cannot be accepted and acted upon until the idea is expanded to give a detailed, practical plan. Yet there may be a later stage when compression again becomes possible, when we can look back and say, "yes, the carbon tax helped us solve this problem." In this sense, our current, pressing need to consider all details and all potential ramifications of climate engineering reflects the fact that society is only partway along a path leading from awareness of this potential approach to development of a clear, coherent plan that might later be referenced as if it were a single object.

out how such a tax could be imposed (on a global scale) in a way that's acceptable, readily enforceable, and effective. It's not clear how it actually would work in a world with eight billion people and almost 200 countries, in a world where more than 80% of the primary energy still comes from fossil fuels (IEA, 2021).

I don't want to get sidetracked here by discussing all the issues involved in setting up a carbon tax. This example just highlights the complexity involved when developing and vetting plans to address challenges of the Anthropocene. Any fully developed plan for a carbon tax, or effective plans for climate engineering, will have a level of complexity far beyond anything that can be captured or described at a single, simple 4-chunk level.

This 4-chunk limit — dominating so many of the basic patterns of human thought and human discourse — presents a central challenge as society tries to deal with the complex problems of the Anthropocene. True, we need to use these 4-chunk patterns as we write and as we speak with others, yet any careful analysis of the challenges society now faces (challenges like those of climate engineering) will require hundreds or thousands of such 4-chunk ideas, with each such idea connected to the others in carefully defined ways.

Although not yet emphasized above, this 4-chunk limit of working memory capacity also comes into play as we try to rearrange ideas, or try to develop new ideas, in some active moment of thought.[9] This 4-chunk limit is readily apparent as we use the rules of arithmetic learned in grade school and try to add a tip before leaving the restaurant. It's apparent in the standard structure of a logical syllogism, for example, as we are asked to make an inference from the two premises that "all men are mortal" and that "Socrates is a man." It's readily apparent as one sees the modest complexity involved in each of the "Boolean operations" introduced in George Boole's book *An Investigation of the Laws of Thought*. (And, likewise, the risks from pressing beyond this limit are readily apparent when Lewis Carroll offers logical puzzles more complex than the standard syllogism (Carroll, 1958).[10] As complexity increases, we need some

[9] The cognitive challenges of "thought" and the cognitive challenges of "listening" are not as different as they might first seem. Indeed, every moment in a conversation brings fresh cognitive puzzles. Listening is never some fully passive process. One must be actively engaged in every moment, working to reconstruct (in one's own mind) the meaning and significance of the speaker's words and deciding how to respond.

[10] Consider the substantively greater cognitive challenge when there are three premises (each with several chunks), as when Lewis Carroll asks: What we can logically conclude if told that a) no potatoes of mine, that are new, have been boiled; b) all of my potatoes in this dish are fit to eat; and c) no unboiled potatoes of mine are fit to eat?

careful, stepwise way to proceed — breaking the problem into a series of smaller puzzles so as to have a reliable way of getting the correct answer.)

We now turn our attention to see how a second, very different type of memory comes to bear as the human mind tries to address complex new problems.

Long-term Memory: Long-term memory (if carefully developed and nurtured) can set up far more complex patterns, as it can rely on intricate connections among the brain's approximately 100 billion neurons,[11] and patterns can arise here that are not readily compressed into single 4-chunk frames. Much of the real power of the mind lies here, for: 1) this network *is* "the self" (a point driven home in a painfully obvious way when the previous self of a family member disappears as dementia ravages the brain). However, when someone is healthy, when long-term memory is intact, this network also helps track the interrelationship of myriad different ideas and concepts. It thus 2) helps represent the relevant "background" and "situation," often providing a vital context for the few chunks shared in any moment of speech,[12] and 3) this network of ideas in long-term memory is the main cognitive resource we have as we try to think about complex new problems.

How Learning *Gradually* Gives the Brain New Power: Long-term memory also allows for complex "transitional states" as new ideas gradually develop, and it can be upgraded/updated in ways that allow the brain to acquire capabilities that it did not have at an earlier stage of neural network development. This storage capacity and this plasticity allow for various kinds of deep, slow learning and deep, slow thought. Ideas can develop over time, with connections in the brain gradually changing — month by month, and year by year — as suggested in Figure 4 (which is intended to represent a time series of different neural/synaptic patterns that may emerge as learning proceeds, or as a new plan or new theory gradually develops in the mind of the individual).[13]

---

[11] One of the most frequently cited studies estimated that the "typical" human brain has about 86 billion neurons (Azevedo et al., 2009). Available evidence suggests that the human brain is — for the most part — just a scaled-up version of the brain of other higher primates (Herculano-Houzel, 2012).

[12] And thus, for example, effective communication usually requires that speaker and listener have some shared sense of the relevant situation or context.

[13] This diagram is intended to remind us of the way in which ongoing neural activity — as when studying a new course in college — gradually changes neural circuits so that the student ends up with mastery of the new material (correctly answering questions on the final exam that they would have had no way to answer at the start of the semester). As noted in the text, this kind of process can occur on many different time scales — days, weeks, months, or years — but we'll use this diagram when discussing patterns of thought so "deep" that they require changes in long-term memory before one can master challenges at the next level.

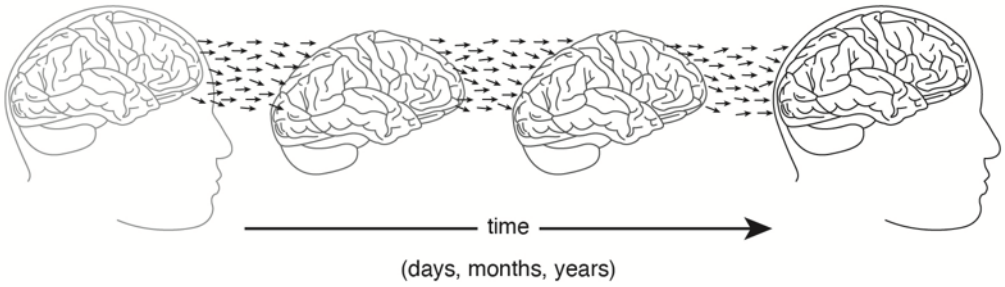

Figure 4

Obviously, these internal changes in the neural/synaptic architecture only occur in the context of all ongoing sources of experience, information, and advice affecting the individual, so perhaps it's better to expand the diagram as in Figure 5 below:

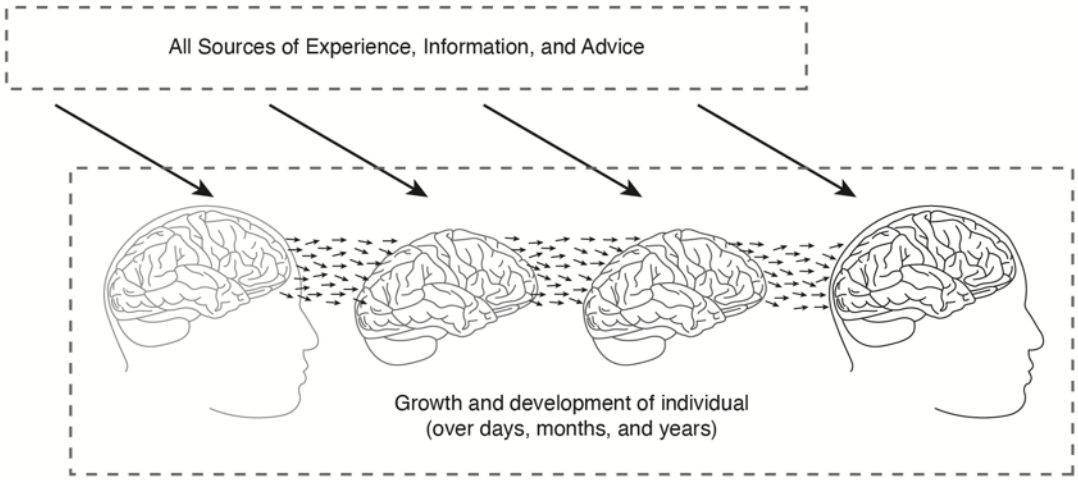

Figure 5

The cumulative benefits of such gradual changes in long-term memory are perhaps most readily evident when comparing students at different stages in the educational process. Thus, for example, neural networks present in the mind of a graduate student studying organic chemistry are capable of representing nuanced ideas about molecular orbital theory and about the way in which the shape of such orbitals may control the relative reactivity of a series of different compounds. But these ideas cannot yet be meaningfully handled — are as yet "unthinkable" — in the context of the brain of a student just beginning their first course in introductory chemistry. Years of study may be needed before the student will have synaptic connections among neurons arranged in a way that will allow them to understand and apply these new ideas.

When considered in the context of standard patterns of educational development — as above — it's easy to see what's happening: education facilitates preliminary changes in neural networks that are later needed to solve much more complicated problems. The mind changes in moving from high school, to college, to graduate school, to postdoctoral research. (Note: in the stance taken here, the full power of the human mind arises at a level of physical complexity and detail that neuroscience itself has not yet been able to capture.) A correspondingly vast set of preliminary changes is needed as a scholar approaches some fundamentally new problem, and is needed now as society struggles to find effective ways to address the challenges of the Anthropocene.

Layer after layer of ideas will need to come together in new ways, and this takes time, continued study, and careful attention. Looking at the lives of great thinkers and great artists makes it clear that critical insights tend to come only after years of prior preparation and observation, with data showing that it usually takes at least ten years of careful, deliberate practice to reach a level of world-class performance in any new area.[14]

The need to work on this time scale (a constraint that seems awkward in an age of computers) is one of the most fundamental limits inherent in the development and use of long-term memory. Those exploring some fundamentally new, uncharted realm often need to work for years — slowly updating neural/synaptic connections — before the mind can think at a level needed to analyze and solve difficult new problems (especially when, as here, analysis of the issues requires a pan-disciplinary perspective, far beyond anything offered by our modern educational system).

Note: Problem solving at this level certainly requires one to "know the facts" (to acquire information as a sufficiently large — and suitably sophisticated — set of 4-chunk patterns). Yet even that stage — by itself — is far from sufficient. There's the slow, difficult challenge of sorting out all the connections and relationships among this initial

[14] One of the early studies of the preparation needed for elite performance (Hayes, 1981) started by picking 76 acknowledged master composers of classical music, then checked to see how long each composer had been working before writing their first (generally acknowledged) masterpiece. Different composers had started writing music at different ages (Mozart was writing symphonies when he was 9), but — in almost every case — the composers had been working for ten years or more before they wrote these masterpieces. (Two were written after nine years; one after eight years; yet none at an earlier stage.)

set of ideas,[15] and there's the related challenge of rejecting purported "facts" that, when examined more closely, seem irrelevant or unreliable. (Taking a few famous examples of ideas that needed be overturned like this: For Copernicus it was the idea that the Earth is the center of the universe; for Einstein the notion that space and time are independent parameters. Similarly, Darwin needed to dismiss the idea that the Earth had been created in 4004 B.C. and needed to question the standard, implicit assumption that all members of a given species were somehow fundamentally the same. He needed to see how they varied, and how these differences might allow for a process of evolution.)

IV. Charles Darwin and the Use of Incremental, Long-term Thought

Charles Darwin is one of my intellectual heroes, so I'll use him to illustrate what happens with this type of deep, slow, incrementally assembled thought — thought so profound as to require a complex, multi-year series of changes in long-term memory before one can answer the question or solve the problem. Figure 6 (below) emphasizes how Darwin collected ideas from every source at every stage of his work, how he constantly tried to weigh and integrate all these ideas as he kept working on the problem of the origin of species. A rich, widely variegated stream of input data was needed, but — inside Darwin's head — everything needed to proceed at an entirely biochemical and biophysical level (as emphasized in Figure 1). Synapse by synapse, year by year, the mind/brain that Darwin had when he boarded the *Beagle* in 1831 (left side of Figure 6) needed to be restructured and rewired to become the mind/brain that could write *The Origin of Species* in 1859.

[15] Teams employing this algorithm will typically need to start their inquiry without knowing all the facts or knowing which ideas will be most important. And this — almost inevitably — will lead to a significant "cognitive overhead" arising at later stages: new ideas, new priorities, and new relationships among ideas may emerge in a way that requires everything else to be double-checked in the context of new constraints. (Classroom learning usually avoids this kind of challenge, since a good teacher and a good textbook can introduce ideas in a logical, well-organized way. The teacher knows what concepts will be needed at later stages, and thus can proceed in a way that avoids any need to radically restructure neural networks partway through the semester.)

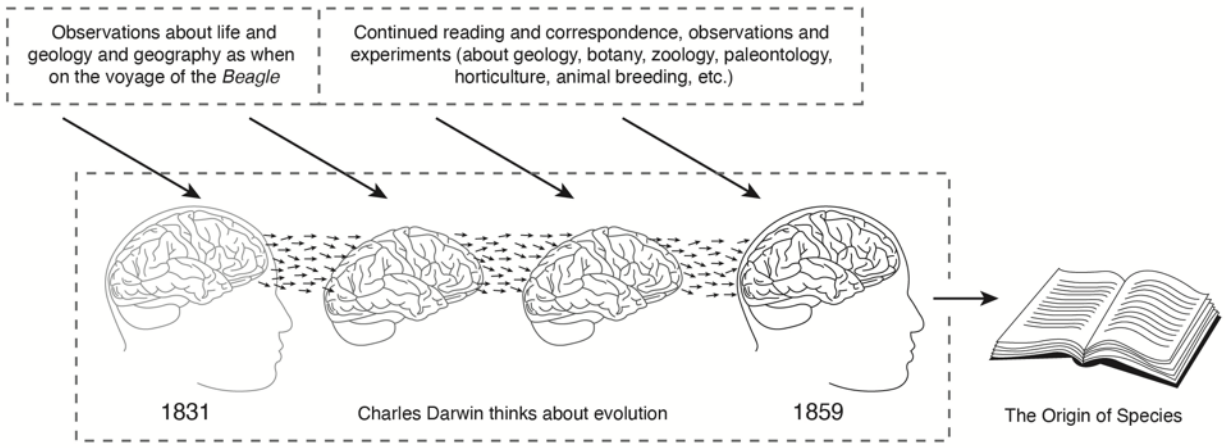

Figure 6

Obviously, we cannot know any of the subtle neurophysiological details involved in the development of Darwin's brain. But we do know (Browne, 1996 and 2002) that, by the time he formulated his theory of evolution, he had immersed himself in ideas about geology, botany, zoology, anatomy, physiology, and plant and animal breeding. He had absorbed and pondered all he had seen and experienced on the voyage of the *Beagle*; he had read Charles Lyell's *Principles of Geology* (showing that our earth was shaped and reshaped by changes occurring over immense spans of time). John Gould, the ornithologist who studied specimens from the Galapagos, had told Darwin that the finches he collected there represented a set of distinct but closely related species, and Darwin had read Thomas Malthus's *Essay on the Principle of Population* (showing that populations tend to grow exponentially until constrained by some kind of competitive pressure).

Undoubtedly, Darwin was a genius, but much of his genius arose from the patient, almost cyclic, recursive way in which he worked during the decades needed to develop his theory. He'd look for weak points in his argument, ask new questions, gather more data, adjust his theory, and repeat the cycle. As he kept working, he took full advantage of the kind of neural/synaptic plasticity that's accessible when someone focuses on a problem in a steady, systematic way over a period of years. He gradually updated his own long-term memory in a way that let him think more clearly about the remaining questions and concerns (and that let him find better ways to explain his theory). He could develop his theory, and could write *The Origin of Species*, only because his brain had time to progress through a series of intermediate neural/synaptic configurations — only because he was able to absorb, integrate, and extend ideas gleaned from a wide variety of different sources. He used a kind of incremental, multi-stage, long-term thought (which includes, but extends far beyond, the types of thought

discussed in recent books by world-famous psychologists like Daniel Kahneman and Steven Pinker).[16] Time and time again — changes in long-term memory (consolidating one level of analysis) were needed to provide the foundation for the next stage of Darwin's work.

This last point is so important when developing ways to address the challenges of the Anthropocene that it's worth pausing to repeat and to summarize: Very simply, we've seen how so much of the neurophysiological life of the individual depends on ideas that are stored in long-term memory. As seen in the educational process (and as seen with the life of Darwin), steady progress over a period of years occurs only because of incremental changes in long-term memory. Neural networks present at a later stage can absorb ideas (and solve problems) that would have been too difficult to understand (or too difficult to solve) at an earlier stage. As neural networks change like this, they can explore and express ideas that quite literally would have been "unthinkable" at an earlier stage in an individual's development.[17] Darwin — obviously — could not simply trade in the brain of 1831 (left side of Figure 6) for the brain of 1859. He needed to earn it, to build it, to create it through active neurophysiological processes — observing, reading, talking, writing, thinking until neural networks in his brain were restructured in a way that let him understand the origin of species and let him explain the answer to others.

V. Why Special Focus Teams Must Learn to Think Like Darwin

In many ways, the intellectual challenge faced by teams working on the problems of the Anthropocene will be analogous to the intellectual challenge that Charles Darwin had faced. Thus, any group exploring prospects for climate engineering, for example, will need access to widely dispersed forms of knowledge. They'll need to consider issues of climate change and global weather patterns, climate engineering proposals, energy use, agriculture, terrestrial and ocean ecosystems, national and international politics, misinformation and disinformation, cybercrime, human history, legal theory and practice, morality/justice, and the global economic system.

---

[16] For example: Kahneman (2011), *Thinking, Fast and Slow*; Kahneman, Sibony, and Sunstein (2021), *Noise: A Flaw in Human Judgment*; Pinker (2018), *Enlightenment Now;* and Pinker (2021), *Rationality.*

[17] There is even an analogous process that occurs inside a computer as it carries out a calculation. Later stages often rely on data structures that did not exist — had not yet been calculated or updated— at an earlier stage in the analysis. In this sense, the final answer is impossible to calculate (is, for the computer, "unthinkable") until the computer has completed all the earlier stages of data processing.

Teams will then need to develop some plausible action plan — a plan showing if/when/how a climate engineering program that uses stratospheric aerosols might best be deployed to help avoid the worst ravages of climate change. As they begin work, team members will need to gather information that currently exists only in a highly dispersed physical form — as present in books, research articles, policy reports, and other minds all over the world. And teams will need to start collecting, integrating, and evaluating potentially relevant ideas without, at first, having any way to know exactly what sources of information or types of information will be most useful, and without any way to foresee the final perspective and conclusions that will emerge from their report. (As work proceeds, the team will need to keep reviewing/reevaluating decisions about the depth of analysis required for each of the topics noted above, and about the best, most reliable resources to use when studying each aspect of the problem.)

The overall process used by such a team might be depicted as in the following diagram (which, for simplicity, just shows one mind rather than the minds of all members of the team):

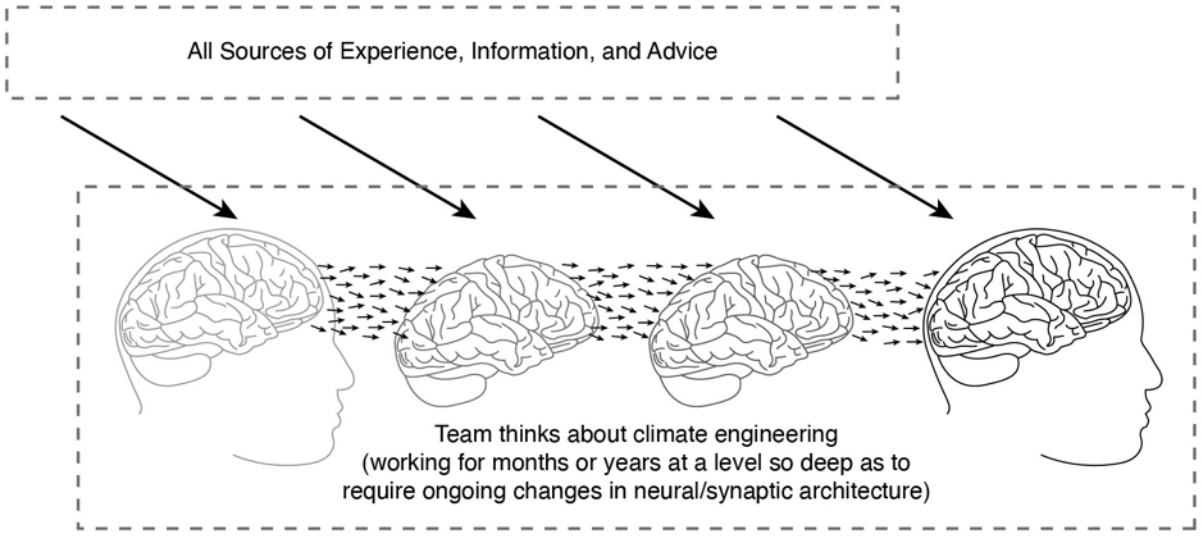

Figure 7

As suggested by the layout of this figure, there is a striking similarity between the challenges of developing a good theory (Figure 6) and the challenges of developing a good plan (Figure 7). Darwin had succeeded because he insisted on a kind of "coherence" between 1) his theory and 2) the "full set of facts" available at the

time.[18] An effective plan must be coherent in somewhat the same way, since — at heart — both a good theory and a good plan must offer, or must describe, some kind of "mechanism" that can work reliably in the physical world.[19] Requiring that plans be as coherent as scientific theories means that there are a large number of different conditions/constraints that need to be satisfied when planning. Yet this is time well-spent, since checking carefully at this stage helps reduce risks that plans will fail when implemented amidst real-world complexity.

Scientific theories always, at least to some extent, remain provisional (Popper, 1959). There always is the risk — at least in principle[20] — that some new discovery will force an existing theory to be modified or discarded. Plans, likewise, can fail if they overlook (or misunderstand the importance of) so much as *one* key variable. Concerns of this form can be seen in terms of a theory of coherence proposed by Paul Thagard, which assumes that the best interpretation or the best hypothesis will be the one that is most consistent with all currently available information (Thagard, 2000). We just extend this idea here to show that it serves equally well when it is used to describe the challenge inherent in developing a plan that must satisfy some large number of constraints.[21]

When actually implemented, a climate engineering program can work as planned only if 1) every key aspect of the final plan is consistent with 2) all relevant aspects of science, technology, economics, local and national politics, international relations, morality, law, misinformation and disinformation, cybercrime, fraud, corruption, and warfare. This complexity makes it easy to see why a team effort is needed when planning, and it highlights the immense cognitive challenge that such teams will face.

---

[18] Darwin's theory was not fully proven, not fully accepted by the scientific community until advances in the 1930s and 1940s revealed the genetic basis for this process of natural selection. Darwin was able to infer the answer from a "subset" of the data (as available to him in the middle of the 19th century), but his theory survives only because it is consistent with *all* the data — both with data Darwin could see in 1859 *and* with all data subsequently acquired by other scientists via other modes of inquiry.

[19] Evolution has worked via a process of natural selection for billions of years; plans for climate engineering may need to work for 100 years or more.

[20] Logically, as emphasized by Popper, there is some chance that any theory (even, hypothetically, a well-established theory like Darwin's theory of evolution) might need to be revised or replaced on the basis of some future observation. However, at a practical level, scientists and funding agencies seem to implicitly assign different levels of "statistical risk" to different scientific theories. As emphasized by Thomas Kuhn (1962), the daily work of science needs to proceed on the assumption that the main ideas (the key paradigms) presented in the textbooks are right.

[21] In our terminology, a theory or plan is "coherent" if it's consistent with all expected/known constraints.

(Although it's difficult to make any direct comparison, the cognitive challenges here seem harder than the challenges that Darwin had faced.)[22]

There will be an immense challenge in attempting to develop a plan that will meet desired goals and also will be consistent with all relevant aspects of life in a complex world.[23] And, given the physical foundations of thought as discussed in Section III, we are led to a kind of corollary of coherence theory that affects every other aspect of our work at Humanity 2050: if incoming ideas are to help when we check for coherence, they must be assimilated, absorbed, and properly represented in the brains of our team members. This is the only way that such ideas (covering a pan-disciplinary range of relevant issues) will get to fully participate, as team members think about the problems of climate engineering. There are no shortcuts for someone who wants to think carefully; there is no ethereal realm that the mind can access in the moment of thought. Notes and draft versions of the plan can help as team members share ideas, but books and reports do us no good when lying unread on the desk: Minds/brains must have successfully assimilated a wide range of relevant facts before they can have any reliable way of checking to see whether the plan fits with the full set of constraints.

* * * * *

We can visualize the challenge facing the special focus teams by imagining that each person sits in front of a huge pile of 4-chunk frames — as if from a box of puzzle pieces that had been dumped out on the table. Team members are then faced with the challenge of sorting through this pile of ideas — rearranging pieces, discarding pieces, constructing new 4-chunk units, etc. — trying to understand key forces/factors at play in the modern world, trying to develop a plan for climate engineering, and trying to ensure that all features of the plan are consistent with all relevant knowledge about the world.

However, this overall task of sorting and arranging these "puzzle pieces of the

---

[22] One might argue that Darwin had an "advantage" — an easier problem — since he happened to focus on a biological process that was so fundamental that the main principle was "overdetermined" by the available evidence. Challenges of the Anthropocene are different. Anyone planning for the human future works in a problem domain that is so challenging that they never can be assured of having some final, "correct," best answer. A team will work to reduce the risk of error and to develop better strategies, but the number of possible plans is so large — and so much of the future remains unknown — that they never will have some rigorous way to prove that their final proposal is best or optimal.

[23] It may be impossible (as mentioned in footnote 22 above) to develop some "perfect plan." The argument here, though, is that having a more systematic approach (as described in this paper) should help us develop better plans than we could without some understanding of human cognitive limits and of this idealized "goal state" of coherence, and without some algorithm to help us work as carefully as possible when developing the plan and checking for coherence.

Anthropocene" is so demanding that it needs to be done in a series of stages or cycles, and we thus developed a new algorithm for thought.

After working on the problem of climate engineering, and after thinking carefully about the best way to organize and manage the kind of cognitive processing needed for this task, we decided that our special focus teams will use a strategy that relies on:

1) developing an explicit list (or set of lists) with all the key constraints and assumptions;

2) repeated cycles of thought and writing — working on a set of draft documents that summarize the best current plan, the remaining challenges and concerns, and the most plausible alternative plans that we have at the moment.

As explained in Part II, there is a cumulative power here, for each such cycle will upgrade both 1) the document itself and 2) the long-term memory of the team members (those sharing the document) in a way that improves the efficacy/depth of the analysis that will be possible on the next cycle. This approach lets us mimic the methods that Darwin had used with his notebooks and draft documents, but we'll have the advantage having multiple "CPUs" since all team members will be focused on one common goal.

Part II — A New Algorithm for Thought About the Anthropocene

In Part I, we introduced our new model of thought, and then explained how the problem of planning can be described in terms of "coherence theory."

In Part II, we now formalize ideas about this need for simultaneous satisfaction of myriad different constraints when developing plans to address the challenges of the Anthropocene, and we offer a new algorithm for thought that gives a systematic way of proceeding when trying to develop a coherent plan.

VI. A Formal Description of the Challenge of Ensuring Coherence

Our approach to the problem of planning begins by developing three lists of constraints that will be relevant when planning how to address some (particular) challenge of the Anthropocene. The first list, entitled PLAN, includes a list of N_{plan}

features, where each feature describes a separate part or aspect of the plan. Likewise, there is a list entitled WORLD, with each of its N_{world} features describing some key (anticipated/expected) aspect of the world of the future in which the plan will be employed. There also is a list entitled OUTCOME, with features here describing ways in which the plan is expected/designed to change the world. Finally, there are a few yes/no "flags" (sets with single elements) needed to ensure that the plan is clearly described, is actionable, and will be acceptable to society.

Using the problem of climate engineering as an example, a few key features in the three main lists (where numbers are added for ease of reference but have no other real significance) might appear as follows:

PLAN
Feature 1: *Political leaders will set the overall targets (the amount of cooling desired), while the scientific and technical team will determine the precise way in which aerosols are distributed and monitored so as to try meeting these goals*.

Feature 2: *The program will need to ramp up slowly (over a period of several years) to allow for careful monitoring of any unexpected/unanticipated side effects of spraying aerosols in the stratosphere*.

Feature 3: *The scientific/technical team must have preliminary plans ready to show how they would respond to a volcanic eruption*.

Note: We give a few examples here, but this list — and the two other lists below — will each contain a dozen or more key elements or features.

WORLD
Feature 1: *Unified world governance is unlikely; the plan needs to work in a world with myriad independent nation states (with some democracies and some autocracies)*.

Feature 2: *The threat of cyberattacks will increase as computers become ever more powerful.*

OUTCOME
Feature 1: *This program can help society avoid some of the most catastrophic effects of climate change*.

Feature 2: *No aspect of this program will reduce society's focus on desperately needed efforts to limit, and then eliminate, CO_2 emissions and start using other energy sources.*

* * * * *

Developing these feature lists (summarizing key aspects of the WORLD, the PLAN, and the OUTCOME) is a critical step in preparing to use our algorithm, and years of careful study and careful thought may be required when developing and refining such lists. We'll devote a future white paper to the challenges that arise at this stage, but we offer a brief summary here:

Developing the List of Features of the WORLD: The intellectual task of understanding the WORLD of the Anthropocene is akin to the task that Darwin had faced in trying to understand the physical, geological, biological WORLD in which evolution had occurred. Years of study (schoolwork, reading, observation, and correspondence) were required before Darwin could develop some reliable sense of the key features that he would need to consider when developing his theory of evolution.

The task involved in trying to highlight key features of the WORLD of the Anthropocene is even more daunting. So much is happening simultaneously, so much is changing, that it's hard to know what trends are reliable and what trends are the most significant. And yet there's no point in trying to develop a plan unless one knows something about the WORLD in which the plan will be deployed.

Thus, a special focus team, or any group attempting this level of planning, must 1) read and consider ideas from myriad different prognostications, projections, and trends, and then must 2) try to develop some integrated overview (focusing on features that seem most important, most reliable, and most relevant to the plans under consideration).

(Note: We hope, in the future, to bring different groups together to try developing some shared "best estimate" of the key features of the future WORLD, and this should help everyone who is working to ensure a livable human future.)

Developing the List of Features of the PLAN: The intellectual/cognitive challenge involved in developing some draft version of the PLAN will — on the whole — be somewhat less daunting than that involved when trying to develop a list of key features describing the WORLD. Three things help make it somewhat easier to develop the list of features describing the PLAN: 1) The amount of information needed to describe a plan is less than the amount of information needed to describe the world as a whole. 2) It may be possible to start by using selected elements of plans proposed by other groups. And, furthermore: 3) The algorithm itself (as described in section VII and summarized in Figure 8) is designed to help with iterative improvements of the PLAN, so there is no expectation that this first version will somehow be complete, fixed, and final. (Note: With our current algorithm, there's more pressure to optimize the

description of the WORLD at the earliest possible stage, since the algorithm itself will not have any way of double checking the way in which the WORLD is described, and since WORLD provides the context/background in which every aspect of the PLAN and the OUTCOME then get evaluated.)

Developing the List of Features of the OUTCOME: This list gives a summary of the key ways in which the PLAN is expected to change the WORLD, and there's a profound challenge in working out this list: Enactment of any PLAN means that one enters a world of pan-disciplinary complexity that includes all scientific, technical, social, political, economic, moral, legal, and criminal factors. This means, of course, that there is no way to jump from some feature-by-feature description of the PLAN to some feature-by-feature description of the expected OUTCOME. When developing this list, one must try to foresee how events will evolve — how actions may lead to a wide range of reactions and responses (as if a billiard ball is repeatedly bouncing off the cushion and other balls).

Obviously, it's very hard to make this kind of prediction, and it may take months or years of work to try foreseeing the most plausible way in which events may unfold. Using climate engineering once again as an example, we note: Any attempt — in the PLAN — to set up a system for handling damage claims will (over time) lead some lawyers to focus on finding ways to maximize such damage claims. In a similar way, plans to transfer funds for payment of such damage claims will raise further risks. Various forms of graft and bribery are sure to ensue (along, of course, with submission of various exaggerated and falsified claims).

This list of $N_{outcome}$ features will, quite naturally, include the expected, desired benefits of the program, with some clear specification of the spatial and temporal scales on which these are expected to arise. Yet, for clarity and completeness, it also must include a list of problems that seem likely to arise when the plan is deployed. Thus, taking another example from climate engineering, spraying aerosols in the stratosphere may cause some expansion in the size of the ozone hole. We thus need to list any dangerous, undesired outcomes as well as the desired, favorable outcomes that are expected. (If there are too many negative, undesired effects, these terms will be inconsistent with the yes/no "flag" requiring that the program be acceptable to society.)

* * * * *

Once the team has these three lists in place, they can take up the challenge of trying to ensure coherence: This PLAN must be double-checked to ensure that it's consistent

with all known constraints, and the PLAN may (repeatedly) need to be corrected/updated and checked again in a cyclic optimization process.

This cyclic process will be outline in the next section, but the main challenge is one of trying to ensure that every element of every list gets considered in the context of every other list and, indeed, each such list, taken by itself, must be self-consistent. Thus — at a formal level — the full problem of ensuring coherence includes at least $N_{plan} \times N_{world} \times N_{outcome}$ triplet cross-terms that must be considered (where such cross-terms can be visualized as elements of a 3-D matrix), and there also will be some doublet terms, like those needed to ensure that the plan itself is internally consistent.[24] [25]

Thus, we see: This kind of optimization problem — as discussed here with respect to challenges of developing a plan for climate engineering via the use of stratospheric aerosols — is far beyond anything that the human mind can consider in a single pass. As discussed in the next section, and as we will highlight in Figure 8, planning needs to proceed with some type of cyclic (recursive) process. Changing even one aspect — as with decisions about the type of aerosol to be used, or the level of international cooperation that is expected — could have ripple effects that require additional cycles to see whether other changes are needed.[26]

---

[24] The outcome of this planning process — after a careful attempt to ensure coherence — will depend on the full, final list of constraints that is applied when analyzing potential plans. However, since *every* term (in every one of these lists) must be consistent with *every other* term (in every one of these lists), the final answer *will not* depend on the precise way in which different terms are partitioned among the various lists. Thus, one could either 1) have a separate set of "flags" designed to double-check that the final plan is simple enough to understand, acceptable enough to be implemented, and powerful enough to ensure the desired outcome, or one could 2) just include these three key terms amidst the general list of N_{plan} elements.

[25] The process of checking all these terms may already sound tedious and complicated, but — even here — we have temporarily set aside concerns about more complex ("higher order") interactions that may occur among some of the terms. Thus, for example, the plan may be stable with respect to features WORLD (i), WORLD (j), and WORLD (k) when each of these terms is considered individually, but might be unstable due to the way in which these three real-world features interact in some positive feedback loop that amplifies their net effect. Careful planning will need to consider such risks (which can be explicitly acknowledged if the list of WORLD features is expanded by adding, as new elements, a few clusters of terms that are most likely to interact in this way).

[26] In principle, this kind of sensitivity could make it nearly impossible to develop a coherent plan in some finite/tractable number of steps through the cycle shown in Figure 8. There are no guarantees, since — as we start — we have no way to know the complexity of the problem space, yet we need to hope that the search process will gradually converge. As a practical step, it may help if we always keep a careful record of plausible "partial solutions" (even if temporarily setting them aside to try other approaches). And it may help if we pay careful attention to dominant terms (key features of the WORLD, the PLAN, and the OUTCOME) at an early stage in the process of planning. This should increase the odds that later concerns can be accommodated with a fine tuning of a core plan that emerges after a year or two of

Given the complex "computational challenge" of ensuring coherence, and the resultant need for a cyclic process of decisions and revisions, it also becomes clear: choices of values, goals, and priorities (implicit when setting out the initial, desired $N_{outcome}$ terms) need to be made at an early stage in this optimization process. The overall problem simply is not separable in a way that would let these issues/concerns just be added at a later stage since, in principle, the whole optimization process may need to start all over again.

VII. Our New Algorithm for Thought

To tackle this kind of problem, we need an approach that 1) breaks the overall challenge of planning into pieces small enough that they can fit comfortably in the human mind, yet 2) still allows the special focus team to develop a PLAN that satisfies all the relevant constraints. We thus need some way in which a long series of small, "local" decisions (about particular aspects of the plan) have some prospect of leading to an overall solution that will satisfy all desired criteria for coherence.

Once constraints have been set up as described in the previous section, everything then proceeds via an interactive process in which the team checks and re-checks the revised plan to make sure it meets all the constraints, and then (repeatedly) revises the plan as necessary. The main "recursive loop" (at the very heart of this approach) then proceeds as follows:

work. (This core plan would be somewhat analogous to a branching diagram that Darwin had sketched in his notebooks in 1837. He needed to work out myriad other details, but he never needed to go back and change this fundamental assumption that two or more new species could arise by gradual divergence from some shared common ancestor.)

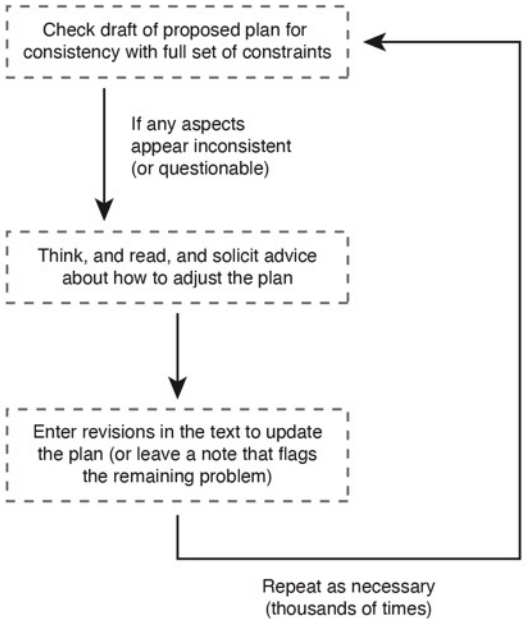

Figure 8

As work on the plan continues, team members must keep checking (top box in Figure 8) for any aspect of the PLAN that still hasn't been specified (at least in outline form), and for any aspect of the PLAN that seems inconsistent with ideas about the WORLD or the desired OUTCOME (or, indeed, with other aspects of the PLAN).[27] [28]

Our algorithm thus proceeds by taking a "global concern" about the overall coherence/consistency of a plan and recasting this in terms of the more manageable, "local" problem of looking for any place where an inconsistency may still exist. When such inconsistencies are noted, team members will then think, read, and solicit advice about how to adjust the plan so as to improve prospects for overall coherence (as in

[27] Teams, on occasion, will need some way of handling deadlocks that arise when team members disagree about whether/how to adjust the plan. Different teams will have different ways of dealing with such deadlocks. They may vote to appoint a "CEO" who makes occasional executive decisions; they may stage a debate among advocates of different options, then vote and work on a version of the plan preferred by the majority of team members; or they may temporally let different subgroups press ahead along different branches of the "search tree."

[28] The whole process gets a bit more complicated than shown in the figure, since there also will be occasions when adjustments are made to the list of features foreseen in this future WORLD or features in the expected/desired OUTCOME. These changes (not directly shown in Figure 8) will "reset" some of the conditions under which subsequent cycles of evaluation/optimization of the PLAN then proceed.

the box in the middle of Figure 8).[29] And, as in the lower box, they will then 1) amend the written plan in the hope of overcoming such problems or at least will 2) leave a note in the text alerting other team members to the problem.

Note: There are perhaps two different intellectual standards (two different "frames") that could be applied when considering this algorithm for thought. At one level, it might be compared with existing methods used when planning how to address challenges of the Anthropocene, and — from this perspective — we think it offers a major advance via the way in which it can help the human mind work effectively amidst the otherwise overwhelming complexity of the Anthropocene. At another level — as compared with numerical methods in computational, combinatorial optimization (Papadimitriou and Steiglitz, 1998) — it's clear that future development/descriptions of our algorithm will need more clearly defined, more formal ways to address issues such as non-satisfiable constraints, statistical/Bayesian thinking, speed of convergence, challenges of avoiding/escaping local optima, etc. (Our special focus teams are, of course, well aware of such challenges and try to address them in ad hoc ways, but we do not yet have any formal way of including all such features in the algorithm outlined in Figure 8.)

* * * * *

Among these remaining challenges, we have no way to know for sure whether (or how quickly) our algorithm will converge on a plan when addressing any particular challenge of the Anthropocene.[30] Like Darwin during his early years of work, teams using this algorithm will just need to do the best they can with whatever information they have at the moment, and then keep pressing ahead and learning more. The human mind is not afforded any omniscient view that might let it — at the beginning — see the final outcome of any particular planning process.[31]

---

[29] When making such changes to the plan, team members may not have any immediate, direct way of knowing whether they are right, and just need to resume by checking again in the next cycle. (They may not know whether these adjustments, avoiding some current inconsistency, actually will end up as part of the final plan. Changing one part of the plan may lead to inconsistencies elsewhere — to problems with other cross-terms that only will be seen as the analysis proceeds.)

[30] We expect, when dealing with problems as complex as those involved in climate engineering, that it may take several years of intense effort to develop a reliable plan. We realize that this seems slow, but we fear that any plans developed more rapidly — as perhaps by other groups that don't use such a systematic approach — are likely to still have serious flaws. (Our algorithm should provide a relatively quick way to check any other such proposals, which should pass the coherence test on the first full cycle if the proposal really offers a coherent plan.)

[31] Given the full-world complexity, given limits in human cognitive capacity and human foresight, it also will be important to continue with annual reviews of any such PLAN (and of expectations about the WORLD and the OUTCOME) even after programs get underway.

We expect, as a general rule, that many iterations will be needed before there is any chance of convergence, and we need to recognize that all early judgments of team members are being made by minds/brains that do not yet have all other ideas (all other puzzle pieces) assembled in a way that would provide a full, well-organized background. Team members also may — at any point — decide that they need to go back and read a series of articles about a topic that may not have been given adequate attention. (Plans, for example, always need to be consistent with expectations about the larger social/political environment — the WORLD — within which they will operate. New information about cybercrime, or about potential legal challenges to a climate engineering program, may force the special focus team to step back and revise some aspects of their view of the WORLD, and this — in turn — may affect subsequent revisions of the PLAN.)[32]

Although not emphasized in our description above, we anticipate that intermediate stages in the planning process — as when proposing ways that society might proceed with climate engineering — will involve development of several variant plans. I.e., we expect that there will be cases in which radically different plans can easily turn out to have similar overall risk/benefit ratios, and it would seem dangerous to drop them at some early stage in the planning process.[33] The final decision — the best proposal that can be offered for society's consideration — will be reliable only if it gets made after there's been time to carefully assemble and weigh all key ideas.

* * * * *

Adoption and use of this algorithm will have several other consequences that are important enough to deserve special mention here:

It may appear, at first, that this algorithm constrains us to just work on one narrow aspect of the plan at a time, but — actually — it places no upper bound on the level of

[32] Thus, expanding on this one example: risks of cyber-sabotage will affect the way in which academic teams (which may be a relatively "soft" target for such attacks) will be able to interact with the operational team that's actually responsible for running the climate engineering program. Just as changing one word on a crossword puzzle alters the range of possible answers for other parts of the puzzle, so addressing this one risk will result in potentially dramatic changes to other aspects of the climate engineering program. In this case, limiting dependence on calculations done by academic groups may reduce one set of risks, but — at the same time — will mean that the whole program becomes less open and transparent, changing things in a way that may make it harder to catch any errors in calculations used by the operational team, and making it much harder to gain and maintain the kind of public support that the program will need.

[33] And, in a similar way, we expect that there will be cases in which a proposed course of action acknowledges serious problems that may occur at a later stage, but mitigates the overall risk by offering several alternative "backup plans" that can be implemented as necessary.

creativity and insight that a team member can bring to bear if they somehow are able to "jump ahead" and fix many things at once. There may be times when some large-scale changes are needed in the proposed PLAN, and it would be fine if someone could, working in accordance with the lower box on Figure 8, just sit down and write out a final, perfect plan that is consistent with the full set of constraints.[34]

It also is important to understand how this approach can be used to supplement and double-check other methods that may be used for planning. We are not aware of any other group that has set out such a careful way of developing and evaluating their plans for dealing with the challenges of the Anthropocene, yet — as mentioned in footnote 30 — the methods offered here could readily be applied to double-check proposals offered by other groups.

Perhaps the most fundamental advantage of this algorithm involves the way in which it helps to ensure that the cognitive challenge inherent in processing the stack of "puzzle pieces" remains at a comfortable (or at least tolerable) level. Each time a few pieces in this puzzle are linked together — each time that some little problem is noted and fixed — can provide a little "reward," a sense of satisfaction, to someone who cares about addressing the challenges of the Anthropocene and who enjoys thinking at this level. (It's a bit like the satisfaction that others may feel with each bit of progress in solving a crossword puzzle or in the assembly of a physical jigsaw puzzle.) Ideally, thought about the challenges of the Anthropocene becomes a kind of flow experience (as in Mihaly Csikszentmihalyi's book *Flow: The Psychology of Optimal Experience*).[35]

VIII. Use of Writing as a Support System for Incremental Long-term Thought

When implementing this algorithm for thought (Figure 8), we anticipate that key transformations will first occur at the neural/synaptic level (in the minds of team members and advisors), but we use writing as a "support system" for this process of

[34] Our method thus works at a level consistent with characteristic constraints on human thought (as discussed in Part I). However, our algorithm never precludes the possibility for some (hypothetical, almost "magical") flash of insight or inspiration that somehow lets the team move forward more rapidly.

[35] Prototypical examples of this kind of flow experience (as discussed by Csikszentmihalyi) involve a surgeon performing an operation or a pianist giving a concert. Each works in a "zone" where challenges at hand are substantive enough so as to demand full attention, yet — in this kind of "flow experience" — problems are rarely so severe as to become unmanageable or overwhelming. The situation may demand the scientist's, or surgeon's, or artist's full attention and full set of skills (usually blocking out concerns about any other aspect of life as various "problems" arise in the moment), yet problems get resolved in a satisfying way as the "performance" continues.

thought.

In essence, we've set things up so that — as cycles continue — our method also allows for an easy interchange between ideas represented in the brain (left side of Figure 1) and ideas represented in written form (as on the right side of Figure 1). This works well since written language has enough flexibility to describe almost any conceivable plan (at least any plan that is clear enough to be shared with others), and since written language presents ideas in an external, tangible form that often is more stable than human memory.

A written text can help capture ideas present at one stage and can provide the starting point (as in the upper box on Figure 8) for any new cycle of thought. Reading some section of text (reading, for example, about some aspect of the proposed plan for climate engineering) takes ideas out of this "freeze-dried" form of thought — as present in the text — and starts to engage the mind so as to continue with the relentless effort to double-check for consistency with all known constraints.

If a team member is sufficiently engaged/immersed in the process of planning, an amazing thing also happens in the background as thought proceeds: Conscious attention to one step/aspect of the plan (as when writing) can activate associative memory networks in the brain,[36] bringing related ideas to mind. This will — almost automatically — stimulate the rest of the brain (the subconscious) to search for conflicts that may arise with any of the items describing i) the rest of the plan, ii) the state of the world, iii) the expected/desired outcome of the program, or with iv) concerns about whether the overall plan is clear, acceptable, and actionable. These neural mechanisms (allowing access to the parallel processing capability of the brain) may not work in a fully effective or fully systematic way, but there is a dramatic improvement in efficiency of the algorithm when the computational subconscious can begin to help like this: Conscious attention to one component of the plan can — simultaneously — allow comparison (at a subconscious level) with many different features of the expected (future) WORLD and of the expected/desired OUTCOME.[37]

This ability — this chance to take advantage of neural mechanisms that can allow for "parallel processing" — only arises/emerges if other aspects — as with all N_{world}

[36] These associative networks tend to facilitate the development/elaboration of other, related ideas, often revealing — in a moment of insight — patterns encoded in the neural networks of the brain that had not yet come to conscious attention.

[37] When thinking at this level, one still is using the same basic algorithm as in Figure 8; one just takes advantage of the parallel processing capability of the brain, simultaneously checking for consistency between 1) one feature of the plan and 2) a whole set of features from WORLD and/or from OUTCOME.

elements of the set entitled WORLD — have been studied so carefully, assimilated so fully, that they are accessible to thought even when conscious attention is focused on considering details of the plan.[38] Preparing for work at this level is not easy. It requires intense effort and focus at all prior stages of the work (so as to have full command of all key features of PLAN, WORLD, and OUTCOME). Yet this kind of attention pays off with a radical improvement in speed/efficiency when a kind of "parallel processing capability" can be achieved.

Since this algorithm (Figure 8) is implemented in a way that relies on writing as a support system, we also insist on having some written output — some edits to the shared text — at the end of each cycle. This forces every member of the team to be as clear and specific as possible after each cycle of thought, and this helps ensure that the algorithm works via a kind of "ratchet mechanism" (with each successive insight getting captured, saved, and shared with other team members). Every cycle of edits is, of course, driven by a desire to ensure that the team has the best possible predictions about the WORLD, and that — within this context — the team has designed a PLAN that has the best chance of achieving the desired OUTCOME. (Edits thus often arise as amendments to the list of features describing this future WORLD, the proposed PLAN, or the expected OUTCOME.)

The process — updating the plan and then fixing ideas in written form — is somewhat akin to what happens, stage by stage, as a rock climber ascends a cliff. Each time that the lead climber moves ten or fifteen feet further up along the wall, he/she pauses and fixes some new anchor, with a piton or a hex nut, that secures the rope at this new, higher position on the rock wall. In a similar way: written language holds the idea in place while preparing for the next move. Each new draft provides a somewhat better way of assembling ideas and of organizing plans than anything the team had at an earlier stage (and drafts from these intermediate stages also provide a convenient way of tracking progress as the team moves forward).

Shocking insights may not occur very often, but there will be myriad little advances every day, with new patterns of neural activity that arise during the intense, focused process of thought. And some of these new ideas will be so fresh, so novel that they are not yet established in a stable biochemical form in the brain of the writer. New ideas may be surprising (even to the person who had the insight!), and writing things

[38] At the end of Part I, we had emphasized that ideas need to be fully assimilated in the mind (instantiated in long-term memory) in order to facilitate careful thought about these challenges of the Anthropocene. Prospects for some kind of "parallel scan" involve even more stringent constraints: If we want ideas to be accessible at this level, they must be absorbed so deeply as to become almost second nature.

down helps ensure that these new ideas don't get lost. Writing thus provides a way of capturing ideas, and this is critical during an interim period when ideas have not yet been consolidated in long-term memory. And, of course, writing also provides an easy, natural way of sharing ideas with other team members, who then — in turn — apply the algorithm as shown in Figure 8. They can respond with further corrections, comments, and advice — leading to further cycles of thought and a new round of revisions and edits.

IX. Further Features to Note When Using this Algorithm

This kind of thought will take time, but this algorithm — with writing as a support system — allows team members to take advantage of all other tools and strategies that are available to people who are engaged in serious thought. Thus:

1) Team members can combine this methodology with almost every other strategy used by serious writers and thinkers. They can seek advice from anyone on the planet; can read any book, research article, or policy report that seems relevant; can switch perspectives whenever they want a new way of looking at a problem. They can pick up a pen and try to make a sketch or flowchart; can see what happens if they try to argue the negative case; can temporarily silence the inner censor and quickly write out a dozen different alternatives. (Most of these alternatives will be rejected upon closer examination, yet this approach can open up the flow of ideas from the subconscious in a way that occasionally leads to fresh, useful insights.)

2) Other forms of written records — outlines, tables, sketches, flow charts, etc. — can be mixed in with the text and shared with the team in a similar way, thus becoming another important aspect of our strategy of using writing as a "support system" for thought. (As when writing the text itself, a whole series of decisions and judgments are involved in setting up a flow chart, and it preserves these decisions in a fixed external form in a way that allows advice, comments, and corrections from others, and that allows for gradual refinement of ideas over time.)

3) Anyone with training in expository writing will have some sense of how to break down the challenges of thinking/editing in ways that help to avoid overwhelming the mind in any given moment. Using outlines and subtitles, and keeping an organized paragraph structure, helps to maintain a mental scaffold around which more detailed ideas can later be organized. One does not try to rewrite individual sentences at the same time that one pauses to consider the overall order of different sections in the paper. A similar ability — readily shifting back and forth among different levels of detail

— is required when using this algorithm for thought (Figure 8) and when updating any white paper that's intended to capture the current best ideas of the team.

4) Looking even further ahead, we note: This way of setting up the planning process — setting up everything in terms of specific lists of constraints and having a well-defined algorithm that one can use when planning — also should be helpful when progress in artificial intelligence reaches a stage at which computers can assist with the recursive, cyclic processes outlined in Figure 8.

X. Avoiding Tasks That Would Interfere with Incremental Long-term Thought

It will be hard enough to solve these complex problems that team members will need to keep a clear focus. Minds of team members must keep developing and adapting in ways that help them better address these global challenges. That is: it's vital that upgrades/updates to long-term memory occur in a way that helps team members as they use this new algorithm for thought. And, here, we must understand that taking on other ancillary tasks will not just waste time. It may interfere with the process of thought in a more direct way, since it introduces a risk of training the brain to work in ways that are at cross purposes with the real demands of developing the plan. Putting this in physiological terms, team members need to be careful: there's always the risk that the "wrong synapses" will get strengthened (when performing another task) and that ideas about the plan will later get judged via criteria that were relevant to that other task, but not actually relevant to the plan the team is trying to develop.

For example:

1) Any attempts to publish in academic journals will tend to force team members to adopt the kind of narrow disciplinary perspective that we've been working so hard to avoid (Pabo, 2021). It may force team members to adopt certain standard patterns of thought, or may constrain the planning process by limiting them to stances that are deemed politically correct in the current cultural environment. We risk constricting our search if we limit ourselves to ideas that fit with any conventional, disciplinary patterns of thought. Teams need the freedom to work outside the box and to explore approaches that may — at first — appear unconventional or unpopular.

2) Any broader, public discussion of early draft versions developed by a special focus team leaves a risk that the team will get stuck with a fixed viewpoint or stance (defending these early ideas in a way that both wastes time and that risks rigidifying their own thinking, rather than allowing them to easily move on with further changes

and refinement of this plan). And — at this early stage — there's nothing to be gained by engaging in debate with members of a broader public who have not yet thought about issues of climate engineering in any meaningful depth (or who have not tried to think carefully about key features of the world of the future in which the program will need to operate). Teams will, of course, need to engage in public debate as they finalize their plans, but it would be a distraction for team members to start engaging in broader debate before completing their own analysis.

3) Team members also may be distracted from the task at hand if they start setting up workshops designed to show how these ideas about thought and planning can be applied in other settings (outside of those involved in addressing the challenges of the Anthropocene). This algorithm for thought offers important, broadly applicable new ideas that could be used by individuals, companies, and government, yet it would be a distraction to try "selling" these ideas about thought to anyone else at the moment. Our teams at Humanity 2050 must focus first on applying these ideas to help address the pressing challenges of the human future.

XI. Summary

In our work at Humanity 2050, we focus on developing plans that will help society address the complex challenges of the Anthropocene. We try to devise plans that are 1) clear, 2) actionable, 3) acceptable to society, and 4) powerful enough to help solve these problems.

Obviously, it would be nice if there were some easier way of developing such plans, some method that did not require the level of thought demanded here. Yet, given the limits of human thought, we are suspicious of any planning process that doesn't proceed as methodically as this algorithm for thought.

The overall development of our ideas at Humanity 2050 begins with the belief that there's no chance of effectively addressing problems of the Anthropocene unless we first pause to understand how the modern world is affected by a crisis of complexity (Pabo, 2020). We then proceed in a way that begins by analyzing these challenges to thought (Part I of this current manuscript) and by then offering a new algorithm for thought and planning (Part II). We realize, of course, that there is an immense amount of hard work left to do (when addressing any particular challenge of the Anthropocene), but our method offers a useful start — a tool that will help at every subsequent stage of this work.

As explained in this paper, the challenge of planning (in our first test case, the challenge of trying to develop an effective plan for climate engineering) can be described as a problem of trying to ensure coherence or consistency among a very large number of constraints. Every element of every list must be consistent with every element of every other list. There will be so many cross-terms that there is no way for the human mind to somehow consider them all at once.

We thus needed, and thus developed, an algorithm for thought that can work amidst this complexity. Careful analysis is needed to make meaningful predictions about the (future) world in which the plan will be employed, and we set up the algorithm in a way that avoids overwhelming the mind in any given moment. Our strategy is a higher-order analog of the type of algorithm that a grade school teacher offers to students who are learning principles of long arithmetic. Our strategy — like those offered in grade school — proceeds in a way that keeps the overall goal in mind yet breaks the problem down in a way that gives a steady stream of manageable sub-problems. Obviously, risks of human judgment and problems resulting from the limits of human cognitive capacity still remain, but our algorithm is clear enough that it can be implemented in a very direct way by special focus teams and will help them develop better ways of addressing the challenges of the Anthropocene.

This approach — using our new algorithm for thought — has several key advantages: 1) It avoids any situation in which team members might otherwise be overwhelmed or paralyzed by the complexity of the planning process, and yet it ensures that the team eventually considers all relevant terms. 2) It engages everyone on the team in a fully active process of thought. The overall knowledge base available to the team keeps increasing even when inconsistencies are detected, even when the plan needs to be revised. 3) It provides the mind of each team member with a steady stream of little problems and challenges, and it allows for meaningful progress as puzzles get solved and the analysis proceeds.

This algorithm will help with every aspect of our work at Humanity 2050 and — since it will allow plans to be developed more carefully — this algorithm will help improve prospects of a livable future for our children and grandchildren.

Bibliography

Azevedo, F. A., Carvalho, L. R., Grinberg, L. T., Farfel, J. M., Ferretti, R. E., Leite, R. E., Jacob Filho, W., Lent, R., & Herculano-Houzel, S. (2009). Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain. *The Journal of Comparative Neurology 513*(5), 532-541, doi:10.1002/cne.21974.

Boole, G. (1954). *An investigation of the laws of thought: On which are founded the mathematical theories of logic and probabilities*. Dover Publications. (Original work published 1854)

Browne, J. (1996). *Charles Darwin: Voyaging.* Volume 1. Princeton University Press.

Browne, J. (2002). *Charles Darwin: The power of place.* Volume 2. Princeton University Press.

Carroll, L. (1958). *Symbolic logic and the game of logic*. Dover Publications. (Original work published 1886)

Cowan, N. (2005). *Working memory capacity.* Psychology Press.

Csikszentmihalyi, M. (2008). *Flow: The psychology of optimal experience*. HarperCollins.

Darwin, C. (1958). *The origin of species.* Mentor. (Original work published in 1859)

Dimnet, E. (1956). *The art of thinking.* Fawcett.

Hansen, J., Johnson, D., Lacis, A., Lebedeff, S., Lee, P., Rind, D., & Russell, G. (1981). Climate impact of increasing atmospheric carbon dioxide. *Science 213,* 957-966, doi:10.1126/science.213.4511.957.

Herculano-Houzel, S. (2012). The remarkable, yet not extraordinary, human brain as a scaled-up primate brain and its associated cost. *Proceedings of the National Academy of Sciences*, *109*(Supplement 1), 10661-10668, doi:10.1073/pnas.1201895109.

Hayes, J. R. (1981). *The complete problem solver.* Franklin Institute Press.

IEA. (2021). *World energy balances: Overview.* https://www.iea.org/reports/world-energy-balances-overview

IPCC. (2022). *Climate change 2022: Mitigation of climate change. Contribution of Working Group III to the Sixth Assessment Report of the Intergovernmental Panel on*

Climate Change. Cambridge University Press. doi:10.1017/9781009157926

Kahneman, D. (2011). *Thinking, fast and slow.* Farrar, Straus and Giroux.

Kahneman, D., Sibony, O., & Sunstein, C.R. (2021). *Noise: A flaw in human judgment.* Little, Brown Spark.

Kuhn, T. (1962). *The structure of scientific revolutions*. University of Chicago Press.

Lyell, C. (1997). *Principles of geology.* Penguin Books.

Malthus, T. (1993). *An essay on the principle of population*, ed. Geoffrey Gilbert. Oxford University Press. (Original work published 1798)

Miller, G. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review 63*(2), 81-97.

Pabo, C. (2018). *Mind in the 21st century: Human thought for a human future.* Unpublished manuscript.

Pabo, C. (2020, January 10). *Civilization and the complexity trap.* https://medium.com/@humanity2050org/civilization-and-the-complexity-trap-162c64eecb81

Pabo, C. (2021). *"Special focus teams" to help solve the problems of the Anthropocene.* https://humanity2050.org/wp-content/uploads/2021/12/White-Paper_Special-Focus-Teams.pdf

Papadimitriou, C. H., & Steiglitz, K. (1998). *Combinatorial optimization: Algorithms and complexity*. Dover.

Pinker, S. (2018). *Enlightenment now: The case for reason, science, humanism, and progress.* Viking.

Pinker, S. (2021). *Rationality: What it is, why it seems scarce, why it matters.* Viking.

Popper, K. (1959). *The logic of scientific discovery.* Hutchinson & Co.

Tetlock, P. E., & Gardner, D. (2015). *Superforecasting: The art and science of prediction.* Random House.

Thagard, P. (2000). *Coherence in thought and action.* MIT Press.

About the Author

Dr. Carl O. Pabo's research in biophysics made him a world leader on the structure and design of DNA-binding proteins. However, impelled by deeper questions about the meaning and limits of human knowledge, and prospects for the human future, he resigned his academic appointments to work on "theories of thought."

After years of work, Dr. Pabo developed a new model of thought (some aspects of which appear in this manuscript). This model — unlike models from the cognitive neurosciences — is designed to be simple enough so that it becomes possible to 1) track and think about processes of thought at the same time that 2) one thinks about real world problems. As with the algorithm developed here, this model allows for new ways of thinking, new ways of addressing our global problems, and it offers new hope for the human future.

In 2017, Dr. Pabo founded Humanity 2050, a nonprofit institute dedicated to ensuring a livable human future in the world of 2050 and beyond. Work at Humanity 2050 focuses on practical applications and real-world progress, using these new strategies for thought to help develop acceptable, actionable plans to address complex challenges of the Anthropocene.

Carl Pabo received his B.S. from Yale in 1974 (*summa cum laude*) and his Ph.D. from Harvard in 1980. He has been a Professor at the Johns Hopkins University School of Medicine (1982-1991) and MIT (1991-2001), and an Investigator with the Howard Hughes Medical Institute (1986- 2001). While beginning work on these new theories of thought, Dr. Pabo had a Guggenheim Fellowship and appointments as a Visiting Professor at Caltech, Stanford, Berkeley, and the Harvard Medical School (2004-2007). In 2017, he returned to Caltech as a Visiting Professor to teach a course on "The World in 2050." He is a member of the National Academy of Sciences and the American Academy of Arts and Sciences.

Acknowledgments

The development of the strategy proposed here, and the preparation of this manuscript would not have been possible without amazing help from other members — past and present — of our team at Humanity 2050 and from a set of outside readers. Critical contributions were made by Roger Brent, Eric Pabo, Ken Patterson, Liz Savage, Matt Thomson, and Bruce Tidor, with wonderful help from Megan Acio in preparing the figures.