

Transcript

Foundations of Order

2003 Foresight Senior Associates Gathering
Eliezer S. Yudkowsky

What is the difference between a star and a butterfly? The universe is filled with recurring patterns. In the beginning, the universe was filled with stable patterns and frequent patterns. A star is a stable pattern; once interstellar dust coalesces and ignites, it burns for a long time. A raindrop is a frequent pattern; individual raindrops may not last very long, but they emerge again and again. A star can have intricate stellar dynamics, sunspots and corona and flares, but there's no use looking inside the star for optimization or complex functional design - there are no properties of the star that are there *because* they help the star burn brighter, or last longer. A star is an accident, an emergent thing; something that just happens when interstellar dust comes together. In the beginning, the universe was populated by emergent patterns - by accidents.

When we look around us today, we see many patterns, from butterflies to kittens, that have internal structures with vastly more complexity than stars or raindrops. Not that stars are simple, but they can't compare to the enormous complexity of a kitten. The intricate order of a kitten is due to a different kind of pattern-maker, a different source of structure. When we look around us, what we find is not just patterns that last a long time, like mountains, or accidents that happen frequently, like water molecules, but also patterns that copy themselves successfully. The process that structures these patterns is called evolution, and evolution produces complex structure enormously faster than emergence alone. The first ten billion years of the universe were relatively very boring - there was little real complexity because there was no way for complexity to build up. If by luck one star happens to burn unusually bright or unusually long, that doesn't make it any more likely that future stars will also last longer - its successful pattern isn't copied into other stars. Evolution builds on its successes and turns them into bigger and more complex successes. I would bet there is less complexity in the entire Milky Way outside Earth than there is in a single butterfly.

In this room there are three kinds of patterns. There are emergent patterns, like the flow of air currents. There are evolved patterns, like you, and I. And there are designed patterns, like the microphone I'm speaking into, and the clothes I'm wearing, and the room I'm standing in. Emergence, evolution, and intelligence. The three sources of pattern; the three foundations of order. Each one has a unique signature style inscribed on all its works, as distinct as the music of Bach or the paintings of Escher.

When would we mark the changeover from survival of the stable to replication of the fittest? When would we mark the end of the reign of emergence and the beginning of the reign of evolution? Should we look to the first self-replicating molecule? That's the usual answer. But the first self-replicating molecule was, itself, an accident - an emergent pattern - a product of the era of stable things. Evolution could never have gotten started otherwise. From our later perspective, that first replicator ever to exist, our ultimate grandparent, must have looked like one of the oddest and most awkward things in the universe - a replicator with the pattern of a stable thing, a reproducing structure with no evolutionary optimization whatsoever. If you were around at the time, you might have mistaken it for an exceptionally stable thing that just happened to be very common in primordial soup. A great success, to be sure, but nothing fundamentally different. The true nature of the next era would not have been apparent until you saw the second replicator ever to exist, your *penultimate* grandparent, an optimized pattern that would never have arisen in the primordial soup without a billion copies of that first replicator around to potentially mutate. It was the second replicator that was a new kind of structure, a pattern that would never be found in a world of stable things alone.

And what a strange thing a human must be - an intelligence inscribed entirely by evolution, a mind that wasn't constructed by another mind. We are one of the oddest and most awkward sights in the universe - an intelligence with the pattern of an evolved thing. Like the first replicator, we only had to evolve once, but we did have to evolve. The changeover from evolution to intelligence, from replication of the fittest to recursive self-improvement, could never have happened otherwise.

Humans are a tiny little point in the design space of minds-in-general, and because we were constructed exclusively by evolution, we are rather odd in

many ways. We lack basic abilities no sane designer would leave out. We're born with a limited amount of computing hardware. We can never get any more. Instead of adding new capacity as we grow, our brains slowly disintegrate as they age. Our neurons run at the absurdly slow speed of 200 cycles per second. The brain may be massively parallel, but if your computer's CPU ran at 200 cycles per second, you'd also need a billion processors just to get *anything* done in realtime. We have sensory modalities, like vision and sound, that are great for playing baseball, but so poorly adapted to creating computer software that sixty years after we invented the computer, we still haven't figured out how to program it. We communicate using low-bandwidth channels like language, instead of broadband mind-to-mind connections. We can't look inside our minds to see how they work. We can't rewrite our own source code. We can't recursively self-improve.

So what is a recursively self-improving mind like? It's not like anything. DNA improves, genomes improve, but evolution remains the same slow process of trial and error. Thoughts improve, memes improve, but the human brains doing the thinking remain the same. In neither case do you have an improvement process that wraps alllll the way around. As far as I know, there are no good analogies in Nature for recursively self-improving processes. This is not something that can be understood by analogy with the patterns we know. There is no star or raindrop as complex as a butterfly, and there are no good analogies for recursive self-improvement in today's world.

10

Now, there are certain important things that evolution created. We don't know that evolution reliably creates these things, but we know that it happened at least once. A sense of fun, the love of beauty, taking joy in helping others, the ability to be swayed by moral argument, the wish to be better people. Call these things humaneness - the parts of ourselves that we treasure. If human is what we are, then humane is what we wish we were. Tribalism and hatred, prejudice and revenge, these things are also part of human nature. They are not humane, but they are human. They are a part of me; not by my choice, but by evolution's design, and the heritage of three and half billion years of lethal combat. Nature, bloody in tooth and claw, inscribed each base of my DNA. That is the tragedy of the human condition, that we are not what we wish we were. Humans were not

designed by humans, humans were designed by evolution, which is a physical process devoid of conscience and compassion. And yet we *have* conscience. We *have* compassion. How did these things evolve? That's a real question with a real answer, which you can find in the field of evolutionary psychology. But for whatever reason, our humane tendencies are now a part of human nature.

At the heart of all replicating life is an emergent thing. DNA, the ancient echo of that first replicator, is formed from compounds that would have been common on primordial Earth four billion years ago. If you look at any evolved thing on Earth today, you can see that somewhere in its ancestry, far far back, is a stable thing - a kind of pattern very near to one that emerged in the environment of ancient Earth. That pattern has been preserved, from parent to child, in unbroken chain for billions of years, because it lies at the center of the replication process. It is an astonishing thing that a replicating molecule should accidentally emerge from the primordial soup, yet the shape of that event still echoes in today's life.

And if we do our jobs right, then four billion years from now, some... student... may be surprised to learn that altruism, honor, fun, beauty, joy, and love can arise from natural selection operating on hunter-gatherers. Of course a mind that loves beauty will try to design another mind that loves beauty, but it is passing strange that the love of beauty should also be produced by evolution alone. It is the most wonderful event in the history of the universe - true altruism, a genuine joy in helping people, arising from the cutthroat competition of evolution's endless war. It is a great triumph, which must not be lost.

15

That is our responsibility, to preserve the humane pattern through the transition from evolution to recursive self-improvement, because we are the first. That is our responsibility, not to break the chain, as we consider the creation of Artificial Intelligence, the second intelligence ever to exist.

People have asked how we can keep Artificial Intelligences under control, or how we can integrate AIs into society. The question is not one of dominance, or even coexistence, but creation. We have intuitions for treating other humans as friends, trade partners, enemies; slaves who might rebel, or children in need of protection. We only have intuitions for dealing

with minds that arrive from the factory with the exact human nature we know. We have no intuitions for *creating* a mind with a humane nature. It doesn't make sense to ask whether "AIs" will be friendly or hostile. When you talk about Artificial Intelligence you have left the tiny corner of design space where humanity lives, and stepped out into a vast empty place. The question is what we will create within it.

Human is what we are, and humane is what we wish we were. Humaneness is renormalized humanity - humans turning around and judging our own emotions, asking how we could be better people. Humaneness is the trajectory traced out by the human emotions under recursive self-improvement. Human nature is not a static ideal, but a pathway - a road that leads somewhere. What we need to do is create a mind within the humane pathway, what I have called a Friendly AI. That is not a trivial thing to attempt. It's not a matter of a few injunctions added or a module bolted onto existing code. It is not a simple thing to simultaneously move a morality from one place to another, while also renormalizing through the transfer, but still making sure that you can backtrack on any mistakes. Some of this is very elegant. None of it is easy to explain. This is not something AI researchers are going to solve in a few hours of spare time.

But. I think that if we can handle the matter of AI at all, we should be able to create a mind that's a far nicer person than anything evolution could have constructed. This issue cannot be won on the defensive. We need to step forward as far as we can in the process of solving it. What we need is not superintelligence, but supermorality, which includes superintelligence as a special case. That's the pattern we need to preserve into the era of recursive self-improvement.

We have a chance to do that, because we are the first. And we have a chance to fail, because we are the first. There is no fate in this. There is nothing that happens *to* us, only what we do to ourselves. We may fail to understand what we are building - we may look at an AI design and believe that it is humane, when in fact it is not. If so, it will be us that made the mistake. It will be our own understanding that failed. Whatever we *really* build, *we* will be the ones who built it. The danger is that we will construct AI without really understanding it.

How dangerous is that, exactly? How fast does recursive self-improvement run once it gets started? One classic answer is that human research in

Artificial Intelligence has gone very slowly, so there must not be any problem. This is mixing up the cake with the recipe. It's like looking at the physicists on the Manhattan project, and saying that because it took them years to figure out their equations, therefore actual nuclear explosions must expand very slowly. Actually, what happens is that there's a chain reaction, fissions setting off other fissions, and the whole thing takes place on the timescale of nuclear interactions, which happens to be extremely fast relative to human neurons. So from our perspective, the whole thing just goes FOOM. Now it is possible to take a nuclear explosion in the process of going FOOM and shape this tremendous force into a constructive pattern - that's what a civilian power plant is - but to do that you need a very deep understanding of nuclear interactions. You have to understand the consequences of what you're doing, not just in a moral sense, but in the sense of being able to make specific detailed technical predictions. For that matter, you need to understand nuclear interactions just to make the prediction that a critical mass goes FOOM, and you need to understand nuclear interactions to predict how much uranium you need before anything interesting happens. That's the dangerous part of not knowing; without an accurate theory, you can't predict the consequences of ignorance.

21

In the case of Artificial Intelligence there are at least three obvious reasons that AI could improve unexpectedly fast once it is created. The most obvious reason is that computer chips already run at ten million times the serial speed of human neurons and are still getting faster. The next reason is that an AI can absorb hundreds or thousands of times as much computing power, where humans are limited to what they're born with. The third and most powerful reason is that an AI is a recursively self-improving pattern. Just as evolution creates order and structure enormously faster than accidental emergence, we may find that recursive self-improvement creates order enormously faster than evolution. If so, we may have only one chance to get this right.

It's okay to fail at building AI. The dangerous thing is to succeed at building AI and fail at Friendly AI. Right now, right at this minute, humanity is not prepared to handle this. We're not prepared at all. The reason we've survived so far is that AI is surrounded by a protective shell of enormous theoretical difficulties that have prevented us from messing with AI before we knew what we were doing.

AI is not enough. You need Friendly AI. That changes everything. It alters the entire strategic picture of AI development. Let's say you're a futurist, and you're thinking about AI. You're not thinking about Friendly AI as a separate issue; that hasn't occurred to you yet. Or maybe you're thinking about AI, and you just assume that it'll be Friendly, or you assume that whoever builds AI will solve the problem. If you assume that, then you conclude that AI is a good thing, and that AIs will be nice people. And if so, you want AI as soon as possible. And Moore's Law is a good thing, because it brings AI closer.

But here's a different way of looking at it. When futurists are trying to convince people that AI will be developed, they talk about Moore's Law because Moore's Law is steady, and measurable, and very impressive, in drastic contrast to progress on our understanding of intelligence. You can persuade people that AI will happen by arguing that Moore's Law will eventually make it possible for us to make a computer with the power of a human brain, or if necessary a computer with ten thousand times the power of a human brain, and poke and prod until intelligence comes out, even if we don't quite understand what we're doing.

But if you take the problem of Friendly AI into account, things look very different. Moore's Law does make it easier to develop AI without understanding what you're doing, but that's not a good thing. Moore's Law gradually lowers the difficulty of building AI, but it doesn't make Friendly AI any easier. Friendly AI has nothing to do with hardware; it is a question of understanding. Once you have *just* enough computing power that someone can build AI if they know *exactly* what they're doing, Moore's Law is no longer your friend. Moore's Law is slowly weakening the shield that prevents us from messing around with AI before we really understand intelligence. Eventually that barrier will go down, and if we haven't mastered the art of Friendly AI by that time, we're in very serious trouble. Moore's Law is the countdown and it is ticking away. Moore's Law is the enemy.

25

In Drexler's *Nanosystems*, there's a description of a one-kilogram nanocomputer capable of performing ten to the twenty-first operations per second. That's around ten thousand times the estimated power of a human

brain. That's our deadline. Of course the real deadline could be earlier than that, maybe much earlier. Or it could even conceivably be later. I don't know how to perform that calculation. It's not any one threshold, really - it's the possibility that nanotechnology will suddenly create an enormous jump in computing power before we're ready to handle it. This is a major, commonly overlooked, and early-appearing risk of nanotechnology - that it will be used to brute-force AI. This is a much more serious risk than grey goo. Enormously powerful computers are a much earlier application of nanotechnology than open-air replicators. Some well-intentioned person is much more likely to try it, too.

Now you can, of course, give the standard reply that as long as supercomputers are equally available to everyone, then good programmers with Friendly AIs will have more resources than any rogues, and the balance will be maintained. Or you could give the less reassuring but more realistic reply that the first Friendly AI will go FOOM, in a pleasant way, after which that AI will be able to deal with any predators. But both of these scenarios require that *someone* be able to create a Friendly AI. If no one can build a Friendly AI, because we haven't figured it out, then it doesn't matter whether the good guys or the bad guys have bigger computers, because we'll be just as sunk either way. Good intentions are not enough. Heroic efforts are not enough. What we need is a piece of knowledge. The standard solutions for dealing with new technologies only apply to AI after we have made it theoretically possible to win. The field of AI, just by failing to advance, or failing to advance far enough, can spoil it for everyone else no matter how good their intentions are.

If we wait to get started on Friendly AI until after it becomes an emergency, we will lose. If nanocomputers show up and we still haven't solved Friendly AI, there are a few things I can think of that would buy time, but it would be very expensive time. It is vastly easier to buy time before the emergency than afterward. What are we buying time *for*? This is a predictable problem. We're going to run into this. Whatever we can imagine ourselves doing *then*, we should get started on it *now*. Otherwise, by the time we get around to paying attention, we may find that the board has already been played into a position from which it is impossible to win.