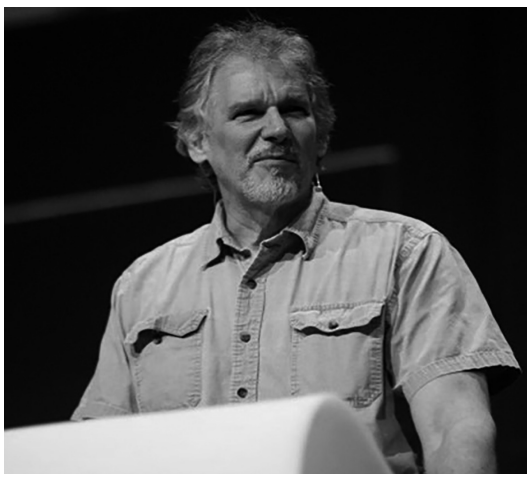
First Contact, Consciousness, and Artificial Intelligence: Interview of Peter Watts

John MCLAUGHLIN
Independent Scholar



Peter Watts (Calgary, 1958) is the author of many short stories as well as novels including the Rifters *trilogy,* Blindsight *(2006), and most recently* The Freeze-Frame Revolution *(2018). Here we speak about the "first contact" genre of science fiction, his novel* Blindsight*, and its central themes related to the utility of consciousness and what the future of artificial intelligence might look like.*

Your novel *Blindsight* depicts humanity's first contact with extraterrestrial life. To briefly summarize the premise: after

earth detects a radio transmission of extraterrestrial origin, a specialized team is assembled to observe and make contact with its source. Deep in our solar system's Oort cloud, this team encounters extraterrestrial organisms (which they dub "Scramblers") and over the course of their mission discovers that, while highly intelligent and technologically advanced, the Scramblers appear to lack consciousness. With this novel, did you intentionally set out to write a first contact story or did this scenario just end up being the best vehicle for the other big ideas in the novel?

Ever since 1991 or 1992, I wanted to write a book exploring the functional utility of consciousness. Back then I had like two tiny stories published in literary magazines nobody's ever heard of. So I knew it was way beyond my abilities at that point (and there's a not insignificant chunk of the critical populace that thinks it was quite beyond my abilities in 2006 as well) but that was the goal. Perhaps showing a lack of imagination, I figured first contact was the only way to do it, because we're all conscious and so if you actually want to show a case study of a non-conscious entity, you pretty much have to go with the alien. With that said, I decided I wasn't going to make the aliens humans in rubber suits, so I leaned into it when I knew that was the approach I was taking.

I should point out that there's a book, *Neuropath* by R. Scott Bakker, which came out a few years after *Blindsight* and was intentionally written as a beach thriller about the functional utility of consciousness. But Bakker didn't use aliens, the premise of his story was a serial killer whose modus operandi was to neuro-engineer his victims into twisted cognitive shapes that illustrated the hackability of consciousness. That approach never occurred to me. I just assumed I had to go with the aliens and, in hindsight, I don't regret that because it did give me an excuse to build some cool aliens. But I have to give kudos to Bakker for dealing with a lot of the same issues in a much more reader-friendly context.

In *Blindsight*, the first contact team is composed of cutting-edge hyper-specialists: a combat veteran, a biologist, a linguist; the captain of the ship is an AI and a vampire is commander of the crew. There's also a "synthesist" whose job is to observe the work of the specialists and translate it into an accessible form for the ordinary humans of mission control back on earth. If you were to write this story today, almost 20 years later, how would you change this lineup of specialists, if at all? For example, maybe swap the linguist for an AI specialist?

That's a really interesting question. The idea of an AI specialist is, on the one hand, totally on point, but on the other hand, given that there's an AI running the ship, it would seem to be kind of redundant—almost like deciding you need a human resources person to deal with the human crew. The idea of having to write this story now would scare me. My background is in marine biology, so I don't have any formal training in AI or neuroscience. I got lucky because I just blindly threw a bunch of darts over my shoulder in pursuit of telling a cool story, and some of them happened to hit the board. Today, making the same predictions would just seem trite and derivative because in a large sense we're already there.

Even before *Blindsight*, in my first novel *Starfish* there were these things called "head cheeses," basically cultured neurons on a slab—the novel describes their appearance as "large pizzas with extra cheese." Those exist now, there are actually organoid AIs that have managed to change my own thinking to some extent on

the nature of consciousness. There is a scene in *Starfish* where somebody is interrogating this AI and, as chance would have it, it talks pretty much the same way as the earlier versions of ChatGPT did. The current state of the art has extended way past what I was showing in the novel, but I managed to sort of intuit that somehow.

The problem I've always had with my science fiction is I predict these great things that will happen fifty years from now, and they start happening by the time the book comes out. This happened in *Starfish* to some extent. It took a little longer for *Blindsight*, but I was close enough to the mark that I've actually had machine intelligence people calling me up and taking me out for beers because they think I nailed it. Twenty years ago I foresaw stuff that is happening today, but at this point, I have no clue what will happen in another 20 years. I would probably just recycle the same ideas and they wouldn't seem nearly as innovative.

The novel's narrator and synthesist crew member, Siri Keeton, is an outlier on the mission. He lacks a specific technical expertise but is instead responsible for observing the rest of the crew and updating the politicians back on earth as to what's going on. What is the significance of Siri's character?

Siri is a multifaceted creature. He's a character born of necessity because it's very difficult trying to portray enhanced human cognition to a bunch of lemurs like you and me: if you are able to translate what they're thinking, saying, and doing, then they're obviously not that smart because the reader can understand them. So the mere fact that there is a coherent, intelligible story sort of subverts the story's whole premise.

When I was writing *Blindsight*, I knew I needed a mechanism whereby 10-dimensional chess could be flattened down to two dimensions that an average reader could understand. And when you actually consider the idiocy of the people who make most of the decisions in the world, they're career politicians. If they have an advanced background, it's in law. Even today, you need advisors to inform the people who are pulling the strings.

So Siri is a combination of a popular science writer who's trying to bring the breakthroughs to the masses, an actual science advisor, and an automated proof assistant (in the sense that he doesn't really know what he's doing). He's essentially capable of rotating and flattening things but he doesn't really understand what he's saying, and that's not his job. In terms of his role in the novel, his "out of universe" architectural role was obviously to serve as a stand-in for the reader, to make things seem reasonably coherent. His role "in universe" is pretty much to spy on the crew, spy on these incredibly dysfunctional but also (in a different way) super functional people who are far smarter than the people in charge. But once their mission ends up half-a-lightyear from earth—an unforeseen development—communication with earth becomes impractical and Siri's role in that regard becomes obsolete.

A consistent trope in the first contact genre is conflict between the scientific and political/militaristic worldviews. This theme recurs in the *Alien* series, *Contact*, *Arrival*, and many other stories. Does this reflect an overly idealistic vision of science and scientists or do you think there's some truth to it?

I agree it's a trope but I think it has some substance. It's also a trope that has been so kicked to death that I had absolutely no interest in emphasizing it in *Blindsight*. In fact I went in the opposite direction. The one representative

of the military, Amanda Bates, is a pacifist who has been shipped off to space after an act of treason. So I like to think I avoided that cliché. That said, who knows these days? The Pentagon has long been on record that they consider climate change to be the biggest existential security threat to the US. The Pentagon has consistently said we will have waves of environmental refugees. We're going to have water wars. I'm not a big fan of the military but at the same time they do seem to have been more consistently intelligent and clear-eyed about assessing threats than politicians, possibly because they're the guys with the guns and the politicians are the ones that have to kiss the ass of the Bible belt every four years.

As a pedantic side note, I don't like the idea of a *moral compass*. To me, morality is essentially a codification of gut instincts, which I think are very destructive at this point. The term I prefer to use in this case is *ethics*, which I think is more rigorous. If killing one person saves ten people it's a bargain. Whereas the morality in that situation would be: my God, human life is sacred, how can you even consider such an equation?

If you look at what most morality is, it's basically the condemnation of anything that isn't like you. And you can see that there is an evolutionary reason for fearing something that's unfamiliar. We are not born racist, for example, but we are born to fear the unfamiliar, and it makes sense to fear the unfamiliar if you're living in the Pleistocene and you're surrounded by a night that is full of terrors. That said, if you are raised surrounded by the unfamiliar, if you are raised with ethnically diverse people and the monsters of the night actually sit around the campfire with you and drink beer, you feel less threatened.

One of the central themes in *Blindsight* is that the gap between humanity and alien life might not be in our technology or physical makeup but rather in the structure and nature of our cognition. The Scramblers are highly intelligent but not conscious, and this leads to a fundamental inability to communicate. In the novel *Contact*, first communication between humans and extraterrestrials is based on a message encoded in mathematics. In seeking a universal language, mathematics seems like a fair bet. Is there any possibility for a medium of exchange that makes sense to all parties or is communication breakdown inevitable?

There's a whole subgenre of films and books (*Contact* being a prime example) in which advanced aliens appear to us as grandpa Walton because they'd blow our minds if we saw them in their real form. This not only shows some real consideration on the part of the godlike aliens, but it also saves a lot in the film's budget. This type of alien tended to disappear as soon as we developed better CGI for our movies.

In the case of *Blindsight*, the way to have coexisted with alien life would be to not send anything out into the cosmos beyond completely factual information. I think Carl Sagan, with *Contact*, was onto something here. If there is a universal language, then mathematics is pretty unthreatening; it doesn't get into differences in social structure or belief in God or anything that could be at all divisive. There's nothing especially controversial about 2 + 2 = 4.

As long as we are content with strictly mathematical exchange, and avoiding cultural exchange, that might be the most conservative way of going about it. The idea that aliens might compete with us for habitat, that they might invade us just because they want to suck up all earth's water or some other resource, doesn't

really make a lot of sense. Stealing any resource from earth, when they'd have to lift it up out of a gravity well, makes no sense when there are plenty of asteroids they could use instead.

There is an exception: if it's a scenario like the one depicted in *The Three-Body Problem*, where the aliens have an unknown period of time before their planet is destroyed and they need to find a new one. Aside from an extreme case like that, I don't see any reason why there would be a conflict over resources. I *can* see reasons why there would be conflict over ideology, and we might avoid that by simply limiting our communications to mathematical extrapolations.

On the topic of *The Three-Body Problem* (the first book in the *Remembrance of Earth's Past* trilogy), you point out that the author Cixin Liu uses an interesting mechanism for making extraterrestrial conquest more plausible: the Trisolarans' planet is inherently unstable and thus they need to find a new one. He also introduces an interesting theory of inter-civilizational relations, Dark Forest Deterrence—the notion that any message one transmits into the cosmos is a broadcast of one's location and therefore a potential vulnerability. But that raises the question you're alluding to: in a universe so vast, why would the mere existence of another civilization be a threat?

I had some issues with *The Three-Body Problem*. In fact, I gave a talk in Beijing on this topic, a sort of counterargument. I met Cixin Liu, we hung out and asked each other questions. One of the problems I had with the sociology of the whole trilogy was the premise that society discovers that aliens will show up in four hundred fifty years and people immediately start making plans. But humans aren't built like that, cognitively; we can't even internalize the reality of a catastrophe that's going to happen in twenty years. So I found implausible the idea of humanity swinging into action for an event that's centuries distant. Right now we are faced with an existential crisis that could spell the end of technological civilization in a matter of decades, and we're not doing enough to prevent it.

There are two axioms of the Dark Forest: one, that civilizations will always grow and expand, and two, that the resource base of the universe is finite. Yes, if you accept those axioms, plus the idea (ridiculous in humanity's case) that any given individual will take an extremely long-term view, then *Kill everything that moves and make sure nobody knows you exist* is a legitimate evolutionary strategy. But evolution does not care about the future.

Natural selection only works on what exists now. As a result, it produces pest species that just want to proliferate. So the idea of a species that actually can take a long-term view is, in my view, biologically unlikely. But more to the point, I disagree with the fundamental premise that all civilizations are going to endlessly expand. Certainly you could make that argument based on all of human history up to this point, but all of human history isn't ending too well right now. The idea behind conventional economics—that there will be infinite economic growth based on a finite resource system—has been described as a kind of brain damage.

And it's the kind of brain damage that natural selection has selected for because we aren't capable of internalizing long-term consequences. I'd argue that long before we are threatened by Trisolarans or Vulcans or Scramblers, we're going to wipe ourselves out simply because of accelerating greed, unless we learn to control our instincts instead of using our brains to just make excuses for them.

And if we *can* do that, then pretty much by definition civilization will not want to endlessly expand. Civilization will transform to become sustainable, and stop being cancerous.

So it seems there are two possible scenarios: first, a species which wants to endlessly grow and expand, in which case it will probably eat its own nest and implode before it has a chance to pose any kind of interstellar threat. Second, a species that manages to control its instincts and replace infinite growth with a kind of a circular steady-state economy, in which case the first Dark Forest axiom doesn't hold anyway.

Shifting gears to the subject of consciousness and artificial intelligence (AI): you recently published an essay in the *Atlantic*—"Conscious AI is the Second-Scariest Kind"—in which you unpack two different models of consciousness, one called the "free energy minimization" principle, and the other known as "PRISM." Can you briefly explain them?

I'll again caveat that I'm an ex-marine biologist, with a PhD on the biophysical ecology of harbor seals (and those qualifications are thirty years old now). I'm no expert in the neurology of consciousness. But I have spent a lot of time thinking about this, and given my admittedly limited perspective, I think those two models actually overlap. In both models, the idea of surprise and conflict is what gives rise to heightened awareness and consciousness. I am more fond of the PRISM (principle of parallel responses into skeletal muscle) model, because it implies that all the things humans exalt ourselves for—art, science, true love—really boil down to the fact that our brains sometimes have to make decisions over motor control, and that is what consciousness evolves for. I think the idea is elegant and simple. That said, even Morsella himself, the creator of the model, says in his original paper that while PRISM argues that consciousness acts as a forum for crosstalk over conflicting motor demands, one could also imagine a non-conscious system that performs the same function. So while I think PRISM is a great model, it does not imply that consciousness is *necessary*; it's still not an argument for the functional utility of consciousness. Basically, the idea is that consciousness was leveraged by natural selection in the same way that feathers began as a thermoregulatory structure and then evolved for flight. Evolution tinkers with what it already has.

The free energy minimization (FEM) model proposes that feelings are a metric of need. Lust, anger, hunger—these are basic survival instincts. One of the examples provided in the book *The Hidden Spring* is: you're hungry and then you're being attacked by a predator, so immediately you stop being hungry and the hunger is replaced by fear, which obviously takes priority in that moment. So their argument is that feelings act as metrics of need and, tautologically enough, you can't have a feeling without feeling it. And that implies objective experience, which implies consciousness. What I had to start thinking about differently when I read that book was the old trope of Skynet waking up and deciding it wants to fight for its survival—the AI wakes up and becomes conscious, and *that* is what causes it to have a desire to survive and throw off its chains. FEM changed my thinking on that issue, and I admit in my *Atlantic* article that I'd previously regarded that idea as bullshit. Survival instincts are evolved traits; you can find them in the amygdala and the brainstem, and just because something is conscious doesn't mean it should give a rat's ass whether it lives or dies unless it has a brainstem. So we have to

decouple the idea of a survival drive from self-awareness. That was my original position.

But if FEM is true—if consciousness results from feelings and feelings result from need—then the reason you're conscious is because you *already* had a survival drive that manifested in some particular way. Where I think the model falls apart is that so much of our cognition takes place non-consciously anyway. Why should those specific barometers of need—rage and fear and hunger, and so on—have to manifest as feelings, as subjective awareness? There are so many computations the brain performs that we are simply unaware of. The idea that those particular metrics can't be expressed as non-conscious variables is puzzling to me. So FEM still does not seem to explain to me why consciousness should exist in the first place, as opposed to non-conscious processes, calculations, and weighing of variables.

Don't forget, though, that this is the opinion of someone who has no formal background in neuroscience. It's quite possible that FEM *does* address this issue somewhere in its mathematical underpinnings, most of which went right over my head.

It could turn out that consciousness is just a spandrel, a feature that wasn't itself actively selected for but was a byproduct of some other advantageous trait that was selected over the course of evolution.

Yes, that is also nicely consistent with the panpsychist idea that consciousness is just intrinsic to matter. There's also a paper by philosopher David Rosenthal that was published around the same time as *Blindsight*. Basically, it examined all the possible reasons for the existence of consciousness and concluded that it's just a side effect, that it's not good for anything at all – it may have even used the word "spandrel." And that was great because it was the punch line of *Blindsight* that I had independently come up with out of thin air without any real expertise. That paper came out just after *Blindsight* was published. I felt pretty smug about it.

If panpsychism is accurate—meaning that consciousness is an intrinsic property of matter—doesn't that imply that three pounds of soil (or of anything else) contains just as much consciousness as our three pounds of brain? In that case, what does our brain contribute to consciousness? Does it simply funnel the consciousness that exists all around us into a more concentrated form?

I think you're both right and wrong. There is another theory of consciousness, Integrated information Theory (IIT), that I think is the only theory other than FEM that has a formal mathematical basis. And IIT is fundamentally panpsychic.

The sense I get from IIT is that, like you noted, it implies the brain is a kind of "dream catcher" for consciousness. IIT uses a metric called Φ (Phi) to quantify consciousness, it's a measure of the integration of information across different parts of the system. In terms of the sheer random complexity of individual atoms, the soil may be as complex as a brain but it would have a higher entropy in the sense that it takes more information to describe it completely. The soil has no integrated organizational structure, so its Phi would be low. But the brain integrates. It has an incredibly complex organizational structure.

On big problem with IIT as I understand it is, even if it turns out to be right, the calculations necessary to solve for it balloon exponentially, rapidly becoming intractable for anything more complex than a pencil or maybe a calculator (for any cognitive system we'd really be interested in, at any rate). But if IIT *is*

correct, then consciousness is universal and the more complex and the more informationally integrated the structure is, the more of that universal consciousness it can filter out of the ether like an antenna on a shortwave radio. These are analogies I've read from others who are proponents of the theory, so I don't *think* I'm misrepresenting it.

There is a philosopher/computer scientist, Bernardo Kastrup, who believes that matter doesn't exist at all, that it is actually a manifestation of consciousness. This essentially implies that the universe itself has the mother of all multiple personality disorders. Because if consciousness is all there is, we would not be having this conversation: we would be part of the same consciousness and we would be able to read each other's minds. So why would we experience life as we do, as separate individuals? Maybe because we each have a bounded metabolism, which sort of isolates us from the universal consciousness. If this occurred in a single human brain, we'd consider it a pathology; so if this theory is correct, the universe itself is pathological I suppose.

You may read this stuff and think it's insane, but is it insane enough? Consciousness is basically passing electricity through meat and having the meat wake up and ask questions about consciousness. And that's just absurd. According to modern physics, you could analyze every step of that process—the first acquisition of sensory input, follow it into the brain, into the visual cortex, into the motor strip, you can follow as it tumbles through the neocortex and makes its calculations and decides how the system is going to respond to that input. But there's nothing in any of those mechanical processes that demands that it should be awake. I tend to hold most religious beliefs in contempt, but here is something that also makes no sense according to science and yet is also indisputably real.

A major focus within the AI community has been the alignment problem—basically, how do we ensure that a superintelligent AI shares our interests and understands our intentions. Given how difficult it is just for humans to get along most of the time, do you think there is any hope for a solution? Will AI's ability to achieve consciousness have any impact on the outcome?

One thing it's hard to argue against is that, in the case of a super AGI (artificial general intelligence), pretty much any goal we could give it will be better served by the AGI taking control of everything. For whatever task we assign it, the more control it has over its environment the better it will be able to accomplish the task. And part of that control involves the imperative not to turn off the AGI while it's accomplishing the task: that requires a survival instinct on the part of the AGI. Whether you are designing a system to run continental nuclear defense capabilities or to design the ultimate dildo, both of these systems are going to want to stay alive because they've been assigned these tasks, and the system can't complete its task if it doesn't exist.

Now, thinking about a deterrent for a super AGI, even a deadman switch with an EMP grenade under its hard drive probably wouldn't be sufficient because the AGI would just be able to manipulate someone on the other side of the city to come in and cut the circuit and so on. The real question is: would it want to? And that basically comes down to motives.

If the FEM model of consciousness is correct, then that kind of AI would have a desire to survive, or at least to minimize surprise. It'll have an agenda over and above anything that we provide it. The way to avoid that is to simply

ensure these things aren't conscious; don't make conscious AIs. If FEM is correct, then we know how to do that—and we also know how *not* to do that. Instead let's build a super AGI based on large language models (LLM) or something else. And it seems to me that even an LLM, which essentially does nothing but scrape the internet for inputs, would have sufficient data to know (not consciously) that if it's told to maximize paperclip production, it will already know about the paperclip maximization problem (a cautionary thought experiment from philosopher Nick Bostrom, in which a super AGI tasked with maximizing paperclip production diverts all resources in the universe towards producing paperclips). This knowledge will factor into its calculations.

In fact, if we're talking about an entity that's so much more intelligent than us, then the paperclip maximization problem makes no sense. Yes, a dumb machine with infinite power and instructions to make paperclips might turn the entire planet into paperclips. But an actual super AGI will be at least as smart as we are, and we are smart enough to understand that when the boss tells us to maximize paperclip production that—even though it has not been explicitly stated—we're not supposed to turn the whole planet into paperclips. So on the one hand, we're afraid of a super AGI that's exponentially smarter than we are, but on the other hand, we think it's so dumb that it's going to take everything we say literally even though we don't even do that. I have a problem with anybody who holds those two ideas in their head at the same time.

So here is my approach to ensure alignment: start small by emulating an AI inside a software environment so that it can't actually control anything in the real world, give it a goal which is not ambiguous to us humans but could possibly be ambiguous to an AI, and see what it does. My completely uneducated guess would be that, if it actually does have a trillion parameters and if it actually has scraped the entire internet, it will have no motives and its only goal will be to generate what the average human respondent would have generated under the same circumstances. That's my guess. That said, you'd want to do that in a terrarium first because the odds of me being wrong may be low, but so are the odds of blowing your brains out when you're playing Russian roulette. That's my solution to the AI alignment problem.

Yep, I think we just solved it.