Prolegomena 24 (1) 2025: 155-179
doi: https://doi.org/10.26362/20250107

Book Discussion: Timothy Williamson, *Overfitting and Heuristics in Philosophy* (Oxford: Oxford University Press, 2024)

Q&A on March 5th, 2025[1]

Participants:
TW Timothy Williamson (University of Oxford)
BB Boran Berčić (University of Rijeka)
KS Ksenija Savčić (University of Rijeka)
FČ Filip Čeč (University of Rijeka)
AJ Andrej Jandrić (University of Belgrade)
VB Vito Balorda (University of Rijeka)
MR Matija Rajter (University of Rijeka)

BB: Okay. I think we can start. Maybe I can break the ice. So, let's start. First thing. You believe that hyperintensional properties or relations are representational, that they are not part of the world, but rather part of the way that we represent the world.

TW: Well, of course, representations are themselves part of the world. But I think, for example, if one's talking about something that is not just part of language or thought, then, there can be, as it were, real world intensionality, but not real world hyperintensionality. For example, if you have necessarily equivalent predicates, then they express the same property and or relation.

BB: I was thinking about whether we can find some counterexamples to the thesis. I think that I have one *prima facie* convincing example: uphill and downhill. So, well, x is uphill if and only if x is downhill. That

[1] This online meeting was a Workshop of the research project *Metaphilosophy*, financed by the Croatian Science Foundation (IP-2022-10-2550). The meeting was held on March 5th, 2025. It started at 12:00 and ended at 14:00. You can find a video recording of the conversation at: https://www.youtube.com/watch?v=6WR5UQYkwMw

obviously has to do with the direction of our movements, but it seems that they are objective properties in the world. Going uphill with the car consumes more fuel. Going downhill, you wear up your brakes. And so, uphill and downhill seem to be parts of the world.

TW: Yeah. If we take your example of the car going uphill or going downhill, those are definitely different things, and that's not just some linguistic representation. But when you say the car is 'going downhill', you're using 'downhill' as an adverb. Typically, the semantics of adverbs might be that they express properties of events or something like that, as in a Davidsonian treatment. The events we're interested in here are events of motion. One property of motion is being downhill, which means that you're higher up at earlier times than at later times during the motion. The opposite, being uphill, is the other way around. So, when you think of 'uphill' and 'downhill' as expressing properties of motions or events of movement, they're simply incompatible properties. A movement that is downhill isn't uphill, and vice versa. 'Uphill' and 'downhill' understood that way would in fact be extensionally distinct.

BB: Yeah. Thanks. Ksenija. Unfortunately, she has to leave earlier, so we will start with her. Ksenija.

KS: Thank you very much. Well, my first question has to do with scientific theories and with the degenerativeness of some theories that you have discussed. You discussed specifically justified true belief, but when applied to scientific theories, the position is less clear. You obviously follow an intuition that a good theory should be – well, maybe not simple, but you tend towards elegant explanations and elegant theories. So, as a counterexample to justify true belief, you take theories of semantics that developed in the 20th century, and you argue that those are good developments, that they are fruitful, yet the further developments within the theory of justifying true belief are degenerative. Now, I see your line of argument, but when we speak of scientific theories, such as particle physics theory, they are not necessarily elegant. In fact, there are many theories that are extremely complex. And do you think that there is a level of complexity such that, in itself, it should push us against the theory, or would you agree that we abandon a theory when we gather enough data to support the development of a new theory?

TW: I don't think it can all be data-driven. Sometimes, theories do get complicated. But the key thing is that every extra layer of complication is a cost and we should only be willing to make a theory more compli-

cated when there are really powerful reasons for doing so. This is quite close to the idea that there's something bad about ad hoc theories. In science, often, the criterion of simplicity is doing a lot of work behind the scenes in determining which theories scientists will even take seriously. For example, if you had the theory that, let's say, general relativity is correct except for events happening in Oxford on Christmas day 1983, and for that place and time Newtonian physics works, that's simply a ridiculous theory that no scientist would ever want to take seriously. It's ridiculous because it's got this pointless complexity to it. So, it's being eliminated on those grounds before even getting any kind of serious consideration, which it obviously doesn't deserve. In terms of theories that have a more realistic chance of being taken seriously by scientists, this idea of overfitting is not a philosopher's invention; it comes from scientific practice. Overfitting is a pathology which scientists recognize, because of things like the introduction of new parameters into a theory in a very liberal kind of way just in order to fit the data. So, the problem with these overfitted theories is not that they don't fit the current data, but there's a lot of reason to expect that they won't fit new data because that's typically what happens with such theories. But they're being rejected not on the grounds of actual failure of fit with specific evidence that we possess, but rather just with the expectation that if you do your theorizing in this kind of way, even though, in a sense, it's motivated by the pressure to fit the data, in the long run, it doesn't actually produce very good results in that respect. But this is a projection to what we can expect from future data, because of the overfitting that's going on in the theory construction. So, it has nothing to do with a lack of fit with the current data. It's rather that we expect on more theoretical grounds that there'll be a lack of fit with future data.

KS: Well, thank you. I have my second question, which is in line with the previous one, which was actually written, so I will just read it. And it has a lot to do with what you just said. I agree, of course, that none of our theories should start with 'this theory is probably going to work for everything except in Oxford on a Christmas day'. But, well, here is my question, and I will just read it. So, I take Kuhn's work and his explanation of how science progresses through paradigms, where paradigms are universally accepted models that define research problems and solutions. Before a paradigm is established, of course, our inquiry is primarily guided by curiosity, common sense, and practical concerns. But once accepted, paradigms drive researchers towards highly specialized, very precise

inquiries. And Levy – both Levy and Kuhn – actually use the words 'obscure' and 'counterintuitive' to describe those investigations. No one starts with the supposition that the theory will work in all cases except in Oxford on Christmas day. But sometimes, we just arrive at such conclusions during the inquiry time. Post-paradigm science advances rapidly precisely because researchers focus on predefined problems rather than questioning fundamentals and rather than relying on common sense, and they treat research as puzzle solving – that's from Levy 2003. The progress, of course, comes with that cost you mentioned, and I would even go as far as to say that overfitting is somehow an inevitable byproduct of doing science in that way, yet that is, as far as I know, the only way that we manage to do science successfully. Paradigms in science shift when new data accumulates, and that forces reconsideration. I took the example of Lynn Margulis' endosymbiotic theory, which she developed in 1967. The theory was not only dismissed initially. It was rejected 15 times by different journals, and it was ridiculed, only to become a cornerstone of evolutionary biology in the 1980s. So scientific paradigms, yes, of course, they grow, they become fruitful, and then they come to overfitting and are eventually overturned by new data. So my question is essentially – how do you see the analogy for this process in philosophy?

TW: Well, first of all, in your description of Kuhn's account, one thing that Kuhn emphasizes is that every paradigm is faced with various anomalies, as he calls them. These anomalies often, though not always, consist of data about some particular kind of case, which is actually very hard to reconcile with the paradigm. Kuhn's description of what goes on is that, in effect, people working on the paradigm work with the hope that, sooner or later, these anomalies will be resolved. What they don't do is revise the paradigm in order to achieve some kind of cheap reconciliation with the anomalies. They leave them as anomalies and hope for the best. Kuhn argues that, in effect, this is actually quite an efficient and rational way of proceeding because it enables the full explanatory power of the paradigm to be explored so that it won't be dismissed prematurely. On his account, it's not really the new data that forces the paradigm to be rejected. Rather, what you have are attempts to explain the new data within the old paradigm, which don't work out. Then, you have a new paradigm, which is a new theory that can explain these data far better than the old paradigm did. It will take time. People won't immediately dump the old theory and rush to the new paradigm. They will do their best to understand the new data in a more conservative way. But if eventu-

ally those attempts don't work out and the new theory has a very elegant explanation for what's going on, then people will gradually switch to it. It's not a completely data-driven process. He does not picture scientists as engaging in anything like overfitting. I think that kind of picture of theoretical activity is quite appropriate to philosophy as well. We don't have big data sets of quantitative measurements or whatever it is. But we still have lots of facts, described at a lower level, that we need to account for with the theory. Some of these may take the form in philosophy of thought experiments, with a compelling verdict on a given thought experiment. So philosophy proceeds in a way which, at a very big picture level, is really not so different from what Kuhn is describing. Although, of course, I think that Kuhn does not describe scientists as engaging in overfitting. Whereas, I think, philosophers are guiltier than a lot of the scientists in quickly going to overfitting. They need some consciousness-raising about the problems such a methodology causes.

KS: Yes. I do apologize. I did not attribute that sentence to Kuhn when I said that it seems to me that paradigm-driven science necessarily leads to overfitting. But, yes, in science, new data, at least in some cases, manages to exclude alternative explanations, which I don't think in philosophy, is quite possible in that particular way.

TW: Yes, but I think part of what Kuhn is arguing is that this exclusion of other theories is not as simple as you might think if you thought of it in the way a crude Popperian falsificationist would think of it. He's arguing that, on the one hand, there's always a possibility of having doubts about whether these data are actually correct or whether something went wrong in the data gathering process. But, he's also, and maybe more importantly, allowing that there may be something going on with the data, with the way we're making assumptions which are not intrinsic to the paradigm itself, but assumptions about what's going on in this process, which we make in interpreting the data. Those auxiliary assumptions may later be rejected. It may turn out that the theory can accommodate the data after all, so it's not a simple matter of just excluding certain theories because they're refuted by the data. It's a much more complicated process than that. It's also part of this view that you don't get a shift away from the old paradigm until there's some new paradigm that's doing better. I think it would very rarely be the case that the new paradigm scientists switch to is in any danger of overfitting. It's not that you're never allowed to introduce some new variable. Occasionally, that is a good thing to do. But natural and social scientists are rightly very reluctant to do it. I

don't think that the kind of process of scientific revolutions that Kuhn is describing leads to anything like overfitting. Often the new paradigms find new ways of being quite simple. For example, take the Newtonian revolution; Newton managed to give a unified theory of motion that applied to both terrestrial and celestial motion. That kind of unification is actually a simplifying process. It wasn't that he had to make things all much more complicated. If you take Kepler's astronomy as an example, and the move away from geocentric to heliocentric astronomical theories, in the long run, it led to great increases in simplicity because they could get rid of all the epicycles that people had been relying on. That was one of the reasons why the Ptolemaic astronomy was so unsatisfactory, in the way epicycles kept having to be added. That's a classic example of overfitting.

BB: Filip, one question, please. We'll go in circles until we drop dead.

FČ: Okay. First of all, thanks for having us. I was wondering about the individuation of heuristics. How do we individuate them in philosophy and metaphysics? For example, is the mutual inconsistency between the libertarian and compatibilist meaning of 'could have done otherwise' a heuristic or a clash of heuristics, much like the mutual inconsistency of opposing counterfactuals? Could it be that our notion of a material object – the so-called 'object talk'– is a heuristic as well? It functions properly with chairs and tables but not as well when dealing with energy particles and galaxies. Reliable in most cases but not all? So could these be heuristics? Related to this question, I have a further one that follows this line of reasoning. It seems that we could say that many of our philosophical debates are actually just some kind of clash of heuristics and perhaps even illusionary problems. However, wouldn't such a solution be too radical that it would eliminate a great part of philosophy? The essential question, then, is how can we determine whether we are dealing with a genuine philosophical problem or merely with a clash of heuristics?

TW: Of course, this issue about how you individuate heuristics arises in psychology as well. In fact, psychologists were talking about heuristics before philosophers were. I don't think there is any very neat answer to that, any more than there are neat answers to the question of how you individuate habits, which is not totally unrelated to heuristics because we're talking about, in effect, habits of thought. You contrasted genuine philosophical problems with a mere clash of heuristics: part of what

I'm arguing in the book is that it's heuristics, not necessarily a clash of heuristics, that are responsible for a lot of philosophical problems. These heuristics are the ones that are philosophically most interesting, and maybe even humanly universal or, at least, they're very widespread in certain cultures and perhaps over a wide range of cultures. These are ways of thinking, of reasoning that are, in some ways, very efficient and that people rely on, but they're not 100% reliable. Then philosophical problems come because ingenious thinkers find ways of exploiting the heuristics to get us into contradictions. You can do that because the heuristics are not completely reliable. Even if they are in principle inconsistent, they may still be reliable 95% of the time or whatever. The heuristics I'm talking about were not generally invented by philosophers. They are ways of thinking that are tempting to pretty much all human beings or maybe to human beings in a lot of cultures like ours or whatever. The trouble is that the fact that you're using a heuristic doesn't automatically enable you to know that what you're using is just a heuristic. It may actually feel like a very compelling way of thinking. Once we have understood what these ways of thinking are, we can often see that they're not going to be 100% reliable. I don't think one can give a cheap solution to a philosophical problem by saying, well, this is just a heuristic. If it's very specific to a problem like free will, that's likely to be something philosophers can quite easily recognize as a distinctive theoretical move that you can reject. The deeper philosophical problems come from heuristics much more deeply ingrained in human thinking. They're ingrained because they're really quite efficient most of the time. It can take a lot of work to become aware of them. But also, once you do formulate them, you can see that, although these are not going to be 100% reliable, they are going to be reliable most of the time. That's why we've evolved to use these ways of thinking because of their comparative efficiency. It's not that you can solve any old philosophical problem by saying, oh, we must be relying on a heuristic here. You have to say what the heuristic is and you have to make it plausible that human beings would rely on such a heuristic. It has to be something that's useful a lot of the time. There's quite a high bar to postulating a heuristic in that sense. You have to do some of that work, which I've done in the book, with the heuristics that I've been postulating. I've done some work to show why it's not at all surprising that we'd use a heuristic of this kind and that it's a very general heuristic. It's not something that is only going to be manifest in philosophy itself.

FČ: Thanks. If I may ask a further related question. It seems to me that a vast part of philosophy, if we are relying on finding out heuristics, would be concerned with the program of disentangling problems by showing that they are not about the world but that they are, rather, our way of thinking of the world. Wouldn't that be perhaps a bit problematic for some problems that we have to say that they are world-dependent and not mind-dependent?

TW: Which ones? Can you give an example of what you mean?

FČ: I'm thinking about the free will debate because that's what I'm working on. For example, the experimental philosophy program was used to see if someone could disentangle the debate in some way, but it failed. And then again, here, when we talk about the phrase 'could have done otherwise', it seems to function, depending on the context, as a way of expressing a compatibilist view in one case and a libertarian view in another – treated one way in one context, and used differently in another. Isn't it just a heuristic? If it is, especially because it seems that we also have an overfitting problem there. Because, when you look at the debate, when you read about it, then you see that things get extremely complicated. Things are added on, and in all these cases, all these imaginary examples have to be fitted in in a single notion of what it means that one could have done otherwise. So, it seems that we have overfitting and that perhaps a reply could be: okay, we're dealing with two heuristics. One is grounded in a compatibilistic notion of 'could have done otherwise', and the other is in the libertarian notion of 'could have done otherwise'. And there the clash occurs, and then, actually, we do not have a problem of free will. We have a problem with how we think about free will.

TW: If we're talking about a philosophical paradox, then a part of the solution is going to be to show what was wrong with the ways of thinking that led into this paradox. That doesn't mean that the phenomena we're talking about are not out there in the world. Human beings are acting, and, at least for a given sense of the term 'free will', they either have free will or don't, and that's a worldly fact. For example, you mentioned the phrase 'could have done otherwise'. That phrase uses 'could', the past tense of 'can', which of course, is a typical modal verb. There's a lot to be said about just how we use it to talk of ability and possibility. 'Can' stands for a kind of objective, although fairly restricted, modality. When you talk about what people can do or what something can do, you're talking

about objective possibilities, not simply about ignorance of something. I'm not going to offer my diagnosis of the free will problem, but it would definitely be relevant to understand more about the general use of modals such as 'can' and 'able to' to take that away from the particular philosophical discussion of free will into a context where one's trying to understand, in a very general way, how we talk about what people could have done. Could he have reached for the bottle of wine on the shelf or whatever? That's a more appropriate setting for understanding the use of talk about what someone could have done rather than just trying to produce ad hoc philosophical theories about that one particular phrase, which is just one particular application of a much wider way of thinking about objective modalities. It's very likely that, in applying such terms, we use some kind of heuristic, which would then be something we could investigate.

BB: Well, talking about heuristics, there is a part about the sorites paradoxes at the beginning of the book. So, the heuristic is supposed to be that the small difference doesn't matter. (**TW:** Yes.) Well, on the one hand, it looks perfectly plausible. However, on the other hand, when you are teaching these paradoxes, or when you are talking with people on the street, and when you make the point of the sorites paradox, then many people naturally and spontaneously move from this either-or approach to some kind of gradual approach: If you move one grain of sand, is it still a heap? Well, yes. It is 0.999999 heap. If you remove another? Oh, yes, it is 0.999998 of the heat. Or they say that it is 0.999999 true, or something like that. So, people from this either-or approach very spontaneously and naturally easily switch to this, let's say, gradual approach. And the fact that it is so does not fit with the idea that little difference doesn't matter. Is it really heuristics like the Müller-Lyer illusion? We can't help it in the case of the Müller-Lyer illusion. We still see the difference. We can't help ourselves. However, in the case of this little difference, it doesn't matter heuristics; people spontaneously and easily give up that approach.

TW: We can't help it in the case of the Müller-Lyer illusion because the relevant heuristic is one built into a perceptual system, so it's not under conscious control, whereas the heuristics more relevant to philosophical problems are of a more general cognitive kind. They're at work in our ordinary reasoning about things. That's something under much more conscious control, and we do have a kind of flexibility. Of course, we're very familiar with gradual processes. But invoking numbers like

0.999999 and so on is more culturally specific because, obviously, people who haven't been trained in at least some high school mathematics are not going to say that, but even people who haven't had that training can understand that it's a gradual process. People, of course, can see that there's a problem fairly quickly. Once you alert them to the existence of a sorites paradox, they'll see that there is a process of gradual change. They'll flounder around, trying to give alternative descriptions that are less problematic. If you don't focus on the question of whether it *is* a heap, you can just say things like: 'Well, it's gradually becoming less of a heap', and so on. But we know from the debate on vagueness that those kinds of point don't automatically deliver any solution to the sorites paradox. The difficulty is that people will regard it as silly to say that you can have a heap where taking one grain away will make it not a heap. The reason why they resist that is not just because it's a gradual process. If you talk about a gradual process in more precise terms, you don't get into that issue. The underlying 'small differences don't matter' heuristic operates under the radar, and it gets people into a lot of trouble, as you can see. Of course, people can sense that something is going wrong here. Something I find very striking in a lot of philosophical discussions of it is that people say things like: 'It's built into the concept of a heap that if n grains make a heap, then n minus one grains make a heap'. They regard that as, in some way, a conceptual truth about the concept of a heap. That strikes me as a manifestation of the way in which our philosophical toolkit is impoverished if we don't have the category of a heuristic that we can use to see that there's an alternative way of thinking about what these tolerance principles are, on which they're not analytic, which philosophers have not been trained to regard as an option. They're just heuristics. The problem is that philosophers haven't even had this category of heuristics to think of as an alternative way of understanding what these principles actually are.

BB: I think that maybe a better illustration for heuristics would be the gambler's paradox. So, you're throwing a dice. Throw, throw, throw, and you believe that every time you throw it, the probability that you will get the desired number is higher and higher and higher. However, it's not. Every time, it is the same. However, there is something psychologically compelling about it. So, this heuristic, probably falsely applied, does resemble Müller-Lyer 's illusion. You can't help yourself; you do have that feeling somehow, somewhere. So, this may be a better psychological illustration of heuristics.

TW: You'd have to say what the heuristic was because humans are not likely to have a general cognitive heuristic especially designed for things like long sequences of coin tosses or whatever because that's such a specific phenomenon. It's not unlikely that some kind of heuristic is at work there, but one would have to be quite careful about what the heuristic was to make it psychologically plausible to postulate. That case is one where we can quite easily see when you think about the fact that the tosses are independent and so on, that the probabilities do not change in the way the gambler assumes. It's actually not so like the Müller-Lyer. With the Müller-Lyer, one line goes on looking longer than the other even when you know perfectly well that they're the same length. I don't find with coin tossing that my expectations go along with the gambler's habit of expecting that it becomes more and more likely that the coin will come up heads if it's been coming up tails. Although that's quite widespread, it's not very difficult to get rid of that way of thinking. People who've been a little bit educated in probability theory just don't have any lingering temptation to think that way.

BB: Thank you. Andrej.

AJ: First, I want to say that I find your book very rich and interesting. Your arguments against hyperintensionalism, which has become the mainstream view, are very convincing, especially those in chapter 3, where you argue that various versions of hyperintensional semantics violate the compositionality constraint. But then again, I have a strong feeling that there are cases which call for hyperintensional analysis. An obvious case is mathematics. I can't help but feel that there are true mathematical sentences that express different propositions, e.g., a sentence about the sum of two numbers expresses a different proposition than a sentence about the value of sine function for some argument. In the last chapter of the book, you suggest that all true mathematical statements express the same proposition, but only under different guises; these guises have cognitive significance, but they do not enter into the meaning. This seems to reduce the whole of mathematics to a single proposition. So, when we are doing mathematics, we are transforming one guise of this proposition into some of its other guises. But is that all that we are doing? On the other hand, when we are doing physics, we are trying to establish as many different true propositions as possible. This creates a sharp divide between mathematics and logic on the one side and physics on the other, which is, to some extent, reminiscent of the old divide made by Wittgenstein and logical positivists. Of course, they claimed that mathematics is not

at all about the world. You can say that the sole mathematical proposition is about the world, but still, it is a single proposition. It creates the impression that, in doing mathematics, we are investigating notational variants (of a single claim) rather than discovering truths about the world – I find this problematic because it makes mathematics look like a less worthy object of intellectual pursuit than physics.

TW: I completely share your feeling that there's not such a great divide between mathematics and physics and that, in some sense, they're both equally concerned with reality. One way of getting an entering wedge here is to think about the role of mathematical derivations, which, of course, are used in physics as well as in mathematics. When we're making predictions from a theory, we're deriving consequences from it, usually plus auxiliary assumptions. We're doing mathematical derivations in both. Mathematical derivations are not fully formal in either physics or mathematics most of the time. There is a formal aspect of them, which means that they can, in principle, be checked by a computer. That kind of issue about *form* is explicitly at the level of linguistic form. As a very simple case, if we're using, let's say, modus ponens or disjunctive syllogism, whether something is an instance of that depends on its linguistic form, so we can't completely abstract away to a level of content. That applies to these derivations whether they're done in physics or mathematics: a correct mathematical derivation is sensitive to linguistic form. Another clue that we're not dealing with something purely at the level of content in the case of derivations is that in a typical mathematical derivation, you're using variables, you're writing an equation with variables in it. Those variables simply have not been assigned a specific value because this is all implicitly general. An equation with variables doesn't have a well-defined content because there are no specific values of the variables to give the content. But at the same time, it's also not meant to be just a universal generalization because it's all dependent on assumptions. People have tried to overcome this by weird postulations of arbitrary objects and that kind of thing, but it isn't a very plausible account of what's going on in proof. So, in mathematics but also in those parts of physics where things are being done mathematically, there are quite a lot of indicators that we're not really working just at the level of some kind of pure content; the linguistic form in the specific form it takes with mathematical notation is actually quite relevant. To explain a bit more, I think this use of the term 'content' is misleading because it's still supposed to be governed by some kind of cognitive constraints, which,

as I argue in the book, raise all sorts of problems. If we are working in a purely extensional context, which, in effect, we do in mathematics most of the time because we're not concerned with modal issues there, then the equivalent of these course-grain propositions are simply truth values. If you think of propositions as something quite like truth values, only adapted to a modal setting, that may give a better sense of what they're doing. The other thing I would say is that if you just look at the level of propositions, of course, mathematics can look very trivial. But as soon as you think about properties and relations, the semantics of this intensional kind is not at all trivializing because the mathematical properties are genuinely possessed by some mathematical objects and not by others. There's a mathematical relation between a set and its members, which relates in some cases and doesn't relate in others. If you think about the subject matter of mathematics, not just as a bunch of propositions but in terms of the properties and relations we're studying, those are highly nontrivial and just as differentiated as we need them to be. What may be going wrong here is trying to think of the subject matter in terms of propositions rather than in terms of properties and relations or things of that kind. Once you think in those terms that mathematics is the study of certain kinds of properties and relations and physics of other related ones, there just isn't the same kind of difference between them. That enables one to see that these are both disciplines which are fully concerned with exploring reality and not just playing trivial verbal games.

AJ: Thank you. I could ask another question, but I will save it for the next round.

TW: Sorry, my answers to the questions are a bit long.

BB: Vito.

VB: It seems that in your book, particularly in Chapter 1, you cautioned against premature rejection of philosophical theories based on isolated counterexamples, as a Popperian falsificationist framework would suggest. (**TW:** Yes.) It seems to me that your view resonates with the Kuhnian notion of paradigm shifts in science, where a dominant paradigm is typically replaced by a new one only after accumulating substantial anomalies or counterevidence. Given this analogy, and considering the ongoing debate that you quite extensively went through in your book, namely between intensionalism and hyperintensionalism, could we interpret these as competing philosophical paradigms? If so, does your argument suggest that the 'victory' of one over the other will depend not

on a single decisive counterexample but on the gradual accumulation of evidence that better fits one framework, as is the case with Kuhn's scientific paradigm shifts?

TW: My answer to that is very definite: yes. A lot of the arguments given for hyperintensionalism did make the mistake of thinking that all you really needed was one clear counterexample, and that would be enough to refute intensionalism in general. That's just a mistake about philosophical methodology and also a failure to recognize how what feels like very clear judgments about examples can be generated by heuristics that are not 100% reliable. That's what I argue is going on. The other aspect of this is, again, in Kuhnian terms, that proponents of the hyperintensional paradigm are trying to accumulate explanatory successes and so on. But, in my view, it's not going well because what we see is an accumulation of a lot of very complicated and different hyperintensional models. None of that is actually becoming the standard hyperintensional approach. The explanatory successes of hyperintensional semantics are very limited and it does look quite like overfitting. In the book, I describe the so-called hyperintensional revolution as an attempted coup. It's an attempted coup that has been going on for about thirty years now. It's not established with complete dominance or anything like it, but, of course, it has been very popular.

VB: Yeah. Thanks.

BB: Matija.

MR: Hi there. So, my question goes somewhere along the following lines. In the chapter 'Overfitting and Degrees of Freedom' you write the following: 'A different way to assess the plausibility of JTB is by noting that knowledge is a central focus for our ordinary thought and talk about cognitive matters: is justified true belief a good candidate to play that role?' (69–70). It is claimed here that the term or concept of knowledge has a specific role or function it fulfills in our everyday lives. We should, therefore, evaluate knowledge as JTB on the basis of such a function. If I understood you correctly, you concluded that JTB does not fulfill this function in a satisfactory manner, which is one of the reasons why we should drop the JTB analysis of knowledge. Such a strategy reminds me of the recent methodology of conceptual engineering, the practice of assessing and improving our concepts. Authors working within this framework argue that if a concept fails to fulfill its function in a satisfactory manner, we should revise or replace it. However, authors working

on conceptual engineering also claim that concepts are not the only objects of our evaluation; we can evaluate the functions they are intended to fulfil as well. Therefore, a friend of the JTB analysis of knowledge working within the framework of conceptual engineering might claim that, although our concept of knowledge currently fulfills a function x, it should actually fulfill another function y. It is this function y that is best satisfied by the JTB analysis of knowledge. Do you find this line of argumentation convincing?

TW: The kind of thing you're suggesting is not totally different from a suggestion made by Brian Weatherson in a paper about twenty years ago about counterexamples and philosophical intuitions. He said that maybe JTB cuts at underlying joints better than the concept of knowledge does. Of course, it's not very clear what this function of knowledge should be that JTB would serve. In some papers I refer to in the book, I have done some work on exploring formal models of JTB to see how JTB works formally, and it doesn't really work very well at all. It gives you something very inelegant that is, I would say, of no particular use to anybody. One general point about the function of knowledge is that we use the category in mindreading, in the psychological sense. We understand other people in part by attributing knowledge or ignorance to them. It turns out that the psychological evidence is that using knowledge in that way is much easier than using belief, not just for humans but for other animals as well, let alone talking about justified true belief. Children grasp what knowledge is long before they grasp what belief is. A lot of nonhuman animals have some kind of basic distinction between what other animals know and what they don't know. They don't have a notion of belief, and they certainly don't have a notion of justified true belief. In that sense, the category of knowledge is much more available. It's not very mysterious why that is. When you're talking about what somebody knows, you're talking about how they're related to their environment (roughly speaking). Such relations are comparatively easy to observe. A paradigm is our knowledge of what other people can see: we can see what others can see and what they can't see, and so on. Whereas knowing what they believe is much harder, because what they believe may be detached from their environment in a way that what they know isn't. I say more about this in the section on heuristics for knowledge in the book. When one considers what would be an efficient heuristic for attributing states of knowledge or belief, knowledge is much more straightforward because it can be attributed by a heuristic with the default that the world is open

to view; of course, you have to inhibit that default where there are clear obstacles. By contrast, attributing belief involves something much more heavyweight: attributing clashing perspectives, beliefs that don't fit the world, and so on. For purposes of basic efficiency in understanding others' cognitive states, there's a lot of evidence that knowledge is much better as a starting point than attributing any kind of belief. It's not illegitimate in principle to argue that maybe JTB is a great piece of conceptual engineering, even if it's not the way we naturally think. But there's lots of evidence that it just can't do the job that we do with knowledge, where we need to be pretty good at ascribing knowledge. If you had a bunch of people who went around just trying to attribute justified true belief to each other, they would get into a mess. This is very often the case with proposals for conceptual engineering: they're just totally psychologically unrealistic. What's being proposed couldn't be an easy way for humans to think. When you see elaborate, complicated redefinitions of ordinary words, they're ones ordinary people would simply be unable to apply. It would be similar to the way ordinary people have a lot of difficulty applying very complex legislation to everyday cases. If we followed all these recommendations for conceptual engineering, people would have to spend all their time on the phone to their local philosopher, asking how this conceptually reengineered definition applied to a given case.

MR: Okay. Great. Thanks. This was really informative. Thank you. Okay.

BB: We can proceed to the second round or circle. I have a question related to this issue. You have an argument. I find it very interesting, but I don't know what to think about it. It's a general argument against conceptual analysis. The argument is that normal, average people use concepts like knowledge, causation, action, and so on, and philosophers have complicated analyses of that. So, your argument is that the results of the analysis are not what people have in their minds when they are using these concepts. These results are artificial and unrealistic. People who use these concepts are competent speakers with normal cognitive capacities, and that is simply not what they have in their minds. Okay? But let's take a look at the analogy with numbers or addition. We go to the market, we buy potato, tomato, whatever, and we add four plus three and so. However, addition has properties, mathematical properties, it is associative, commutative, distributive, and so on. But that is not what the average man on the market has in mind when he's adding four and three. Nevertheless, it seems completely legitimate that mathematicians

are analyzing the addition. So, by analogy, it may be completely legitimate to claim that epistemologists are analyzing knowledge with safety conditions, sensitivity condition, justified true belief, causation, whatever.

TW: One has to distinguish between conceptual analysis and other forms of analysis. What mathematicians do is study addition. They're not studying laypeople's concept of addition. They're just trying to produce a good theory of addition. Depending on the exact mathematical setting, one can provide some kind of mathematical definition, maybe in set theoretic terms, of addition. That's a perfectly good form of theorizing. But it's not an attempt to explain the ordinary concept of addition. If you apply that distinction to the case of knowledge, somebody can say: 'I'm not interested in the ordinary concept of knowledge, I'm just interested in what knowledge is'. This might be a form of analysis like analyzing water as basically H_2O, which, of course, isn't to do with how ordinary people thought about it but just to do with what water is. And then they can say: 'I just want to understand what knowledge is'. Fair enough. People then give that as their revised view of what all these proposed accounts of knowledge are, JTB and all the more complicated ones. If you want to, you can think of these not as proposals for conceptual analysis, but just as proposals for what knowledge really is. They're incredibly unilluminating ones, and there seem to be counterexamples to them. These sorts of analysis have little explanatory value. There's no special reason in advance to think that you could characterize knowledge in terms of belief and other factors. Somebody can try if they want to, but there's no special reason for expecting there to be such an analysis. In the case of mathematics, what vindicates the projects of giving a mathematical theory of addition, or whatever, is just that they have fantastic explanatory power. They unify thinking about addition, make it systematic, and reduce it to a small number of basic principles in a theoretically extremely satisfying way. Whereas these attempts to characterize knowledge in terms of belief have no comparable success. It's not just that they all seem subject to counterexamples, but none of them has managed to explain anything very much. So, they just don't have the kind of credit that the mathematical proposals have.

BB: Andrej.

AJ: This is the continuation of my previous question. It seems to me that the main argument against hyperintensionalism in your book is that various versions of hyperintensional semantics are not compositional and

that we should not abandon compositional semantics just to be able to accommodate several counterexamples in metaphysics. That would be too high a price to pay, especially when we can explain away the counterexamples by appealing to heuristics. The main role for heuristics is to explain our mistaken impression that there are intensionally indistinguishable sentences which have a different truth value. But then, what if compositional hyperintensional semantics were constructed? Would you retract your verdict on the convincingness of these characteristic examples in metaphysics, e.g. the ones derived from Fine or Aristotle? What would your diagnosis be in such a situation?

TW: The appeal to compositionality isn't the only thing doing the work there. For example, some versions of truthmaker semantics are compositional if you allow them to have both verification and falsification conditions. I argue that, although they are technically compositional, they are not really in the spirit of compositionality because there isn't a proper compositional explanation of negation. That's concealed by the way the whole theory is based on attributing independent truthmaker and falsitymaker conditions. Even impossible world semantics can be made technically compositional. 'Compositionality' is not a super-precise term because often there turn out to be trivializing ways of technically satisfying the letter of compositionality. Some of those have been used to defend semantics with impossible worlds. In a way, compositionality as a barrier to overfitting has been one of the important constraints in the development of formal semantics. It's been fruitful because people haven't just regarded it as a purely technical constraint, they've looked for proposals that respect the spirit of compositionality. This has been a bar to overfitting. If somebody came along with some new form of hyperintensional semantics that was really very elegant and simple but also properly constrained and that had hyperintensionality as a byproduct, then one should take it seriously. In the case of semantics that maps each sentence to a Russellian proposition, a complex of the properties, relations, objects, and so on the sentence is about, that doesn't have to be totally non-compositional, although it turns out to be very difficult to make it fully compositional and consistent. It could be that somebody will come along with a satisfying theory, but I don't see much sign of one. After all, people have been actively investigating this for the last thirty years. It's unlikely that there'll be some really smooth or simple way of doing it. That would have been found by now, given how many smart people have been looking for it. Of course, the kind of assessments I'm

making have to do with the current state of the field and what evidence we now have. Nobody can rule out for certain that things will still be looking the same way in twenty years' time.

AJ: Yes, thank you.

BB: Filip.

FČ: Oh, thanks. Well, I have a question that is perhaps silly, but I'll ask it nonetheless. I'm bothered by the counterpossibles. It seems to me that if we accept the truth of counterpossibles, then anything goes, and we can have anything. However, if we were to accept them, then we would be also losing something – then we would lose *reductio ad absurdum*. In *reductio ad absurdum*, we accept a claim by rejecting its opposite, as the opposite leads to a contradiction. However, if we accept counterpossibles – in which we treat the antecedent as true even though it is impossible – aren't we, in fact, accepting and building on something that we just typically do not endorse, which we usually see as an endpoint of discussions, as an endpoint of the *reductio ad absurdum* argumentation. So, could this be a way to put a stop to the talk about counterpossibles, or am I just being silly and naive?

TW: *Reductio ad absurdum* is a very good form of proof, widely used in mathematics. It's more than a heuristic: it's a valid principle of logic. In understanding how it works, one has to make the distinction between what's going on epistemically and what's going on at the level of the underlying content. When you make a hypothesis for an argument by *reductio ad absurdum*, say the hypothesis that there are only finitely many prime numbers, that hypothesis is implicitly absurd but not explicitly absurd. From the point of view of the proof as something that brings knowledge, we start off in a position where, initially, we don't know whether the hypothesis is the case, and we derive a contradiction from it. Now, we know that it's not the case. The proof can be summed up in the counterfactual that if there were a largest prime p, then factorial p plus one would be both prime and not prime, or something like that. That's quite a natural way of summarizing the proof. Of course, once we've done the proof, we know that there couldn't be a largest prime. In a sense, once we know that its antecedent is impossible, the counterfactual, though still true, becomes uninteresting. But it has still played a legitimate epistemological role in enabling us to come to know that there is no largest prime number. That's what it's signaling when we summarize the argument. In principle, this distinction between the semantic and

metaphysical level on one hand and on the other the epistemological level is widely accepted in the literature on conditionals. For example, David Lewis makes quite a bit of it in in his book on counterfactuals.

FČ: Okay. Thanks.

BB: Just a small anecdote, a true anecdote about counterpossibles. I live close to the forest, and I asked hunters about the counterpossible that if pigs were flying, we would need a different kind of ammunition to hunt them. And hunters all agreed that it is true. It's a true statement, and they started to explain me the calibers, the diameters of the shells, and so on. So, okay. Yes. Vito.

VB: Thanks. The second question I want to ask is linked to your last chapter and the linguistic guises that you're referring to. If I interpreted your Chapter Five right, you argued that cognitive significance does not supervene on semantic properties, as illustrated by the 'furze' and 'gorse' example. (**TW:** Yes.) You emphasized that linguistic guises, that is, the specific forms in which content is expressed, play a crucial role in cognitive significance. That, in turn, allows for distinct cognitive relations to necessarily equivalent propositions. (**TW:** Yes.) You advocated for a radical separation of content and cognitive significance to avoid distorting semantic frameworks. (**TW:** Yes.) Given this, how does your framework account for the cognitive value of learning new linguistic guises for previously understood concepts, particularly in domains such as metaphysics, logic, and mathematics, where necessary equivalents are prevalent? Specifically, how does the recognition of linguistic guises as 'what we think with,' rather than 'what we think,' enable us to understand the acquisition of new cognitive relations to old truths without conflating these relations with changes in semantic content?

TW: Let's suppose that 'Water is H_2O' is correct, so by the necessity of identity, it's necessary that water is H_2O. So now we've got this new guise under which we can think about water as 'H_2O'. This new guise is connected to the whole of chemical theory; it connects water with hydrogen and oxygen. So, via this new guise for thinking about the same thing, it's hooked up to very powerful theories that enable us to understand things about water. Even in the case of Dr Jekyll and Mr Hyde, when you learn 'Dr Jekyll is Mr Hyde', that casts a completely new light on this character who seemed very respectable, well-behaved, and so on. Although these new guises for individuals, objects, and propositions are just guises for the very same things, they often have new cognitive con-

nections. What you're getting when you gain a new guise for something is a whole lot of further cognitive connections you may not have had before. What I'm resisting is the idea that it must be possible to find a level of content that registers all these cognitive differences. What we have to understand is that in epistemology and cognition, form as well as content plays a really important role.

VB: Thanks.

BB: Matija.

MR: Okay. So, my next question is specifically about the persistence heuristic. I think we already mentioned it in this discussion. (**TW:** Yes.) So the question is as follows. When we apply the persistence heuristic to vague terms or concepts, it leads us to sorites paradoxes because vague terms do not have clear boundaries or defeaters. However, it is at least *prima facie* plausible to claim that we could continue using the heuristic in these cases if we would ameliorate these vague terms. I would claim that, when we talk about the persistence heuristic, we actually want to *maximize* the use of such a heuristic so we can use our cognition more efficiently while at the same time *minimizing* the danger that vague or otherwise inconsistent terms pose to us. Therefore, we can think of conceptual analysis/conceptual engineering as complementing the use of the persistence heuristic. So, I'm curious about what your response would be. Is there any such connection between a kind of conceptual analysis or conceptual engineering and our use of particular heuristics?

TW: If you're thinking that we could make these vague terms more precise, then the problem is that, in most situations, the cost of doing so is to make them incapable of serving their normal function. For example, in the case of 'heap', suppose you could give some more precise definition roughly, equivalent to the original, but you'd have to say something about how many grains are needed and how some have to be stably resting on others, for it to be a heap, and so on. You could say something about, roughly speaking, the statics and dynamics of heaps. But that wouldn't be a way of thinking about heaps we could ordinarily use. When we're talking about heaps, we want to be able to recognize a heap when we see one; not invariably, but most of the time, we can recognize whether something is a heap or not just by looking at it. Having a precise definition wouldn't be useful. In legal contexts, it's very important to have terms that will not give rise to lots of borderline cases. Sometimes, for example, if it's a question of who counts as an adult, for the purposes of

some piece of legislation about voting, or sex, or whatever it is, we can artificially introduce a cutoff point for 'adult': as soon as you become 18, or whatever it is. Those kinds of localized and precisified versions are useful for very specific purposes. But for most of the purposes for which we use vague terms, having a formal definition would be useless: either we would just ignore the formal definition in the way we used the term, or it would be so clumsy and difficult to use that we would have to stop using the term for most purposes.

MR: Okay. Thanks a lot. Thank you.

BB: So, it's already 01:43, 01:45. So I don't know if it's better...

TW: Well, if we go on to 02:00 your time, which is 01:00 my time, then that's fine by me.

BB: Good. Thank you. Well, you have an argument about the hyperintensional explanation that the sentence that snow is white is true because snow is white. It seems that you want to explain it away by appealing to the pragmatics of explanation. It seems that you want to say that this *because* is not part of the world. It's not objectively out there. However, we had the feeling that no matter what amount of pragmatics or explanations you do, this *is because* it is still there; it's objectively out there, and you cannot remove it.

TW: Can you say anything about what this feeling is based on?

BB: Oh, it's something like a classical rationalist *a priori* intuition. You're looking at this sentence that the snow is white is true because the snow is white. And you can't help yourself. It's just like that, like demonstration in geometry.

TW: I'm not saying that 'because' statements have to be false. All I'm suggesting is that when we're deciding which 'because' statements to accept, the heuristic we use has to do with what is explanatory helpful. These judgments will not always be correct, as judgments about the truth value of a 'because' claim, if the 'because' claim really has the standard semantics. A lot of this doesn't even depend on whether the semantics is intensional or hyperintensional. Here's an example. Somebody asks: 'Why is furze just as widespread as gorse'? You tell them: 'That's because furze is gorse'. That's a perfectly good explanation but the content of what you said, that furze is gorse, is exactly the same as if you'd said 'furze is furze', on most semantic theories. That doesn't depend on intensional-

ism. It's also true on many standard hyperintensional theories. These are theories on which content is worldly. Then an informative sentence, like 'furze is gorse', and an uninformative sentence, like 'furze is furze' have the same content because they're both worldly in the same way, talking about the same thing twice over. So they won't differ in content. The point is that the differences in how helpful an explanation is are sensitive to differences in form where there's no difference in content. Those differences play a large role in our judgments about which 'because' claims to accept and which to reject. The problem is that our assessments of these 'because' claims are partly about the world but also very sensitive to cognitive aspects of this situation, such as what is epistemically helpful for somebody who's asking a given question.

BB: Thank you. Andrej.

AJ: Here is a follow-up to what Filip mentioned earlier. I am not sure how we identify heuristics. Is it transparent to us which heuristic we are using?

TW: No. It's not.

AJ: So, we can go wrong in identifying the heuristic that we are using. The heuristics that you describe in your book – the supposition heuristic, the 'why?' heuristic, the belief ascription heuristic, etc. – are quite diverse. It seems like the 'why?' heuristic, which Boran mentioned in the last example, covers several different heuristics. The 'why?' heuristic results from projecting the pragmatics of posing questions and giving answers onto the semantics of 'because' statements. However, we can project different aspects of such a dialogical setting. Would we then get a different heuristic for each projected aspect? And is ascribing heuristics liable to sceptical challenges? There could be several mutually overlapping heuristics, such that each of them fits all the data about our actual behaviour, so how can we tell which heuristic we are using?

TW: Well, these are high-level psychological hypotheses about what is going on. They are susceptible to experimental investigation. I haven't been doing that kind of investigation because I'm more concerned to get these hypotheses on the table and theoretically worked out than to do all the experimental work. These questions will arise for psychology as well about how psychologists individuate heuristics. In the specific case that you mentioned of the explanatory 'why heuristic', the basic idea is to assess 'because' claims in terms of how helpful we would find the

corresponding answer to a question if we were in the position of wanting or needing to ask it. Probably, a whole bunch of different pragmatic and epistemic factors will play into how helpful an answer is to a question and will affect what judgments we make about 'because' statements. This needn't mean that different heuristics are involved; it just means that different factors influence how helpful an explanation is. Explaining is a quite complex cognitive act with different aspects. As far as I can see it, it's fundamentally the same heuristic, but applying it will evoke different sorts of considerations in different cases. If somebody wants to make the case that there are several different, mutually independent heuristics, and you can easily have some of them without others, they can go ahead and try to make that case. On the basis of the evidence and cases I've seen, it seems adequate to treat this as a single heuristic.

AJ: Thank you.

BB: We have a few more minutes. Does anybody have an urge to ask or comment on something? Oh, if not, then I'll misuse my position. About supervenience. It seems, if I get you right, that you are rejecting the notion, or the concept, or the tool of supervenience because it's supposed to be asymmetrical, but, actually, analysis shows that it is not asymmetrical. Analysis shows that it is symmetrical. I think that there is something wrong with that analysis. Well, the concept of supervenience was initially meant as asymmetrical. If you cannot change A properties without changing B properties, that means that A properties depend on B properties, not the other way around. Now, my feeling is, some vague and general feeling, that argumentation goes this way: A supervenes on B, but then we introduce a border case where B also supervenes on A. Aha, you see, the relation is symmetric, it's not asymmetric. But this holds for many other relations, say, subset. Initially and intuitively, the subset is an asymmetrical relation. However, we can introduce the assumption that the set can be a subset of itself, and then it's over. Then, the relation of the subset becomes a symmetrical relation. And the same holds for grounding.

TW: I don't think 'symmetrical' is the right word. I think you just mean that it's not asymmetric. If it were a symmetric relation, that would mean that whenever x is a subset of y, then y is a subset of x. But nobody thinks that. We're just talking about the difference between asymmetric relations and ones which are not asymmetric. In the case of supervenience, it was given a modal definition: you can't change A without changing

B, roughly speaking. The trouble with that modal definition is that it doesn't give you asymmetry because asymmetry implies irreflexivity, that nothing supervenes on itself, which is what people had in mind by dependency. The relation between A and B when you can't change A without changing B is trivially reflexive because you can't change A without changing A. The modal definition doesn't quite have the formal properties you wanted it to have. Of course, the initial move is just to define an asymmetric notion of dependence via supervenience by saying that it's a case where X supervenes on Y, but Y does not supervene on X, which will be automatically asymmetric and irreflexive. But, as I mention in the book, some examples have that kind of asymmetric supervenience, but we still judge that they aren't cases of dependence of the kind that we wanted. What was going on was that people wanted to capture dependence, they gave a modal definitions of supervenience and then asymmetric supervenience, which captured many of the cases, but not all of them. My own view is that the reason they didn't catch all of them was because these judgments of dependency, in ways very close to what I was saying about 'because' judgments, are very sensitive to explanatory considerations. For that reason, our judgments about them are unlikely to be fully consistent with each other. Asymmetric supervenience may be the best one can get in this area. Of course, there's a massive literature on postulating new relations of grounding and so on, to try to make sense of all of that. I suspect it's a quicksand.

BB: Good. Thank you. Well, it's already 02:00. Tim, thank you very much for your time and your energy spent on us.

TW: Well, thank you all for all the interesting questions.