Psychological Biases in the Era of Nuclear Weapons and AI - 1

[00:00:00] Welcome and Guest Intro

Ryan Vest: Welcome to *Horns of A Dilemma*, the podcast of the *Texas National Security Review*. I'm Ryan Vest, executive editor of *TNSR*, and I'm here with our editor-in-chief, Dr. Sheena Chestnut Greitens. We're pleased to have joining us today, Rose McDermott, author of the article, "The Influence of Psychological Factors in the Search for Strategic Stability."

Rose is the David and Marianna Fisher University Professor of International Relations at Brown University, and a fellow in the American Academy of Arts and Sciences. She works in the area of political psychology. She received her PhD in political science, and MA in experimental social psychology from Stanford University, and has also taught at Cornell and UCSB. She has held fellowships at the Radcliffe Institute for Advanced Study, the Olin Institute for Strategic Studies, and the Women and Public Policy Program, all at Harvard University, and has been a fellow at the Stanford Center for Advanced Studies in the Behavioral Sciences twice.

Rose, welcome to *Horns of a Dilemma*. It's great to have you on the show.

Rose McDermott: Thank you so much for having me. I'm really honored to be here.

Sheena Chestnut Greitens: Well, thanks so much. We really appreciate the chance to talk about your fascinating article today.

[00:01:09] Psychology Meets Deterrence

Sheena Chestnut Greitens: Much of your scholarship is built around this convergence between strategic deterrence and these questions of psychology, which is really, you know, I think a fascinating aspect of great power competition, and of world politics.

So in this article, you start by laying out some of these fundamental assumptions about the nature of human psychology that could be detrimental to our concept of deterrence. And I wondered if you could just get us started in the article, by laying out how your article explains the foundations of this problem.

Rose McDermott: Yeah. What I tried to do with the article is to think about some of these fundamental biases in human psychology. I'm not going to say they're universal in that every single person has them, but when you look at the studies, most people have them, most of the time. And so what that means is that the errors that people have in their judgment, and the kind of decisions they make, tend to have sort of systematic biases in the way that they happen.

And so I was trying to think through the influence of these biases on nuclear deterrence, and on strategic stability, in the area of emerging technologies—you know, especially the effect of things like artificial intelligence, on the way that we think about how safe we are in the world ,and the kinds of problems that could happen, not even intentionally, but just sort of accidentally, through the kind of biases that people have in the way that they make decisions, and therefore think about ways that we might be able to mitigate it, or at least lessen the risk.

[00:02:43] Type 1 vs Type 2 Thinking

Ryan Vest: You referenced Kahneman's nomenclature of Type 1 and Type 2 decision-making strategies, and they're pretty fundamental to this article. I was wondering if you could explain what these are, and why they're important to understand in this context.

Rose McDermott: Yeah, I was really fortunate that Amos Tversky, who did the bulk of his work with Daniel Kahneman, was one of my dissertation advisors. And so I was deeply influenced by that work from very early on in graduate school. And, you know, Amos didn't co-win the Nobel because they don't give it to people who are already dead, but Danny accepted it in both their names because the work was done together.

And the distillation that Danny did in, you know, *Thinking, Fast and Slow* about Type 1 and Type 2—I'm not sure that it's, you know, specifically the way that the empirical foundation of it would work, but it's a really good shorthand to think about it, where Type 1 are these fast, intuitive, emotionally based gut feelings or gut reactions, and that's how you make the decision. Most people make most of their decisions that way.

So when you think about habit–the best predictor of what you ate today, is what you ate yesterday. You know, we fall into these habits, and the reason we do it is because the most expensive part of our body, in terms of energy, and glucose, and everything, is our brain. And so anything we can do that lessens the cognitive load is really great for preserving our ability to do things like, you know, run away from threats, and things like that.

So, when we have these fast, intuitive things, they work well most of the time. It's just that in certain situations that are mostly predictable, they lead to systematic biases, and that's how we can think about the way that it might get us into trouble, with regard to the use of nuclear weapons, or the use of AI.

Type 2 is what we all hated when we were in grad school, and we had to learn calculus, right? Like, it's hard, abstract, boring data that we go through to make hard decisions. You absolutely don't want to make decisions about what to eat or what to wear using Type 2. It's just a complete waste. But when you have a hard decision to make, where you have a multi-attribute choice, right—lots of different choices, really hard decisions, high stakes, time pressure, whatever—that is a more effective way to make an optimal choice, meaning making the right choice.

Where your gut feelings can really guide you in the wrong direction, knowing a lot of data can help you figure out things that are more optimal choices. It's just that people don't like to do it because it's hard, and it's time consuming. I mean, you know, if you give somebody a choice between hanging out with their friends and doing calculus, there are people who would choose calculus, but most people would not, right? That's the kind of Type 1, Type 2 distinction, which are buckets, but they're not like empirical, you know, representations, if that makes sense.

Sheena Chestnut Greitens: That makes a lot of sense, just thinking about everyday life, and the way we go through the myriad decisions that come up for us all, every day.

[00:05:41] Bias One Overconfidence

Sheena Chestnut Greitens: I was really struck, in the article, by the way that you talk how these judgmental biases—and you go through, I think four of them—can affect decision-making, and that they might have particularly pronounced effects in generating behavior that wouldn't be what we'd predict under a standard rational actors, in world politics, model. And that might be particularly pronounced when it comes to things like emerging technology,

nuclear deterrence, and therefore strategic stability, and some of the countries that have nuclear weapons today.

I actually would like to just go through each of these four biases, because you lay them out really nicely in the article. And if I remember right, the first one is overconfidence. And so, you know, why is it that being overconfident is going to be a problem when it comes to strategic stability, and nuclear deterrence?

Rose McDermott: Yeah, I think overconfidence is one of the biggest problems that humans have around all kinds of stuff, where–what it basically means is you're more confident than you are accurate. So, even if you're 90% accurate, if you're a hundred percent confident, you're still overconfident. When you get your confidence and your accuracy aligned, it's called well-calibrated, right?

So people who are really well calibrated, for example, are weather forecasters. Why? Because they make predictions all day every day, and then they get immediate feedback. I say it's going to rain and, gee, it's sunny outside, right? I say it's going to snow, and it's raining. And so you get immediate feedback, and you learn from that feedback in ways that allow you to be better calibrated.

I had a colleague once, who would give an exam where he was trying to teach overconfidence, and you didn't just have to give the answer, but you had to say how confident you were. And if you said you were a hundred percent confident about something you were wrong about, the grade he gave you on your exam was negative infinity.

And he was trying to teach you, that even if you're 99% accurate, if you're a hundred percent confident, you're still not well calibrated. And the way that this gets us into trouble is that when we're overconfident, we don't always, for example, think about all the ways that things could go wrong. And so we don't necessarily do the intricacy of planning that might be required.

I thought about this, you know, recently with the series of mass shootings that have happened, you know, on all kinds of college campuses. And it recently happened at Brown, so it was sort of salient, but it was the second time it happened to me; it happened at UCSB as well. And one of the things that struck me is the plan going in was, we hope it doesn't happen.

Well, that's not really a plan, right? And so you have to think through the ways in which you would respond to various worst case scenarios. And if you're overconfident, you don't necessarily do that planning, because you think you can handle it in the moment, or you think that it won't happen. And that's not

necessarily accurate to what's actually going to happen, and how you could plan in a way that would prevent, you know, a catastrophic error, like the accidental launch of a nuclear weapon or, you know, letting AI get out of control.

I just heard the report the other day on NPR, about how all these chat bots now are really restricting emotional and social development in children and adolescents, because they're so sycophantic, and they tell you how wonderful you are. And I do think that inculcates narcissism, but the bigger problem is they don't learn how to negotiate, and bargain, and deal with any kind of pushback. Because AI never pushes back. It just says–oh, you're so smart, what a wonderful question, oh, you're so thoughtful. It's really great you asked that, right? I mean, I can't stand it for like five seconds, but a lot of people get very addicted to that.

And then when you reach a real person who says–I don't think that's right–the response is to withdraw, right, and to say, I don't want to talk to you. I don't want to hear from you, because you're not validating me and telling me what I want to hear.

And so that kind of danger that happens with AI isn't just restricted to your interactions with AI. It bleeds into your interactions with diplomacy, with negotiation among leaders. And that, I think, is an additional danger that we haven't really thought about in a serious way.

Sheena Chestnut Greitens: It reminds me a little bit of one of my undergraduate advisors, Bill Perry, who is famous for saying "hope is not a strategy," which is now a widely used saying, but I always hear his voice when I think about this.

Rose McDermott: I didn't know that that was his line. That's great!

Sheena Chestnut Greitens: I don't know if it was originally, but he was the first person I heard it from. And so, whenever I hear someone say that, or convey that principle, I hear it in his voice.

Rose McDermott: That's great. Yeah, and that basically is what overconfidence is, and you know, it can bleed into all kinds of—not just planning—but it can bleed into all kinds of situations where, like, let's say a leader goes into a negotiation. They may be overconfident about their skills and abilities to persuade the other side—their own, you know, charm. Of course, a lot of these things are exacerbated by narcissism, but they're also distinct from narcissism.

Ryan Vest: There's a lot to unpack there. I love it. And, you know, we see this all the time throughout governments and even in everyday life. Very intuitive.

[00:10:58] Bias Two Planning Fallacy

Ryan Vest: But the second judgmental bias you identified, to me, was a little bit less intuitive, and you called that the planning fallacy—a little bit more difficult to maybe get my arms around than overconfidence. I was wondering if you could talk a little bit about planning fallacy, and how does it affect decision-makers?

Rose McDermott: Yeah. I think it does relate to overconfidence, but it's distinct from it. And what the planning fallacy does is it basically says, you aren't always accurate in how you plan. So the original intuition was actually, Amos Tversky's intuition was–whenever somebody would ask, how long is it gonna take to finish that paper? He'd say, three weeks. And of course it doesn't take three weeks. It takes a lot longer. And so his joke was, however long you think something's going to take, multiply it by four, double it, and that's probably about right. Right?

And my undergraduates do this all the time: I could write a paper in a night. But that night there's a game, or, you know, there's some other competing interest, and so it doesn't get done, because you think you can get it done in this window of time, where maybe you get sick, or something else gets in the way. And so you're not actually able to do that.

The reason it matters for things like strategic stability is, you may not always leave enough time to actually do the planning you need to do, prior to a crisis. So, you know, of course, you know the great line, "the fog of war" can change everything, but it's still good to go into it with a plan. And if you don't have a plan, if you left it till the last minute, if you didn't understand all the component parts you needed.

I think there's another aspect to it, and I know Steve Sloman has done some work on this, which is that we often think that we understand how things work, but we don't. So the classic example he had is flushing a toilet. People think they know how flushing a toilet works, but if you actually ask them to explain the mechanics of it, very few people can.

But nobody sits around and thinks–gee, I don't know how my toilet flushes. And you can think of many situations where people think that they understand a

technology, because they use the technology, but it doesn't mean that they actually understand the underlying foundation of that technology.

And, you know, that's as true for AI, or hypersonic weapons, or nuclear weapons, or chemical, or biological weapons, as it is for flushing a toilet. Like the underlying dynamic of–do you really understand what goes into this–is the same. It's just that people, because they don't actually have to figure out how it works, and they can use it without any problem, means that they think that they don't need to engage in the more systematic planning.

And they underestimate the effort, they underestimate the complexity, and they certainly underestimate how long things will take.

Sheena Chestnut Greitens: That makes a lot of sense.

[00:13:35] Bias Three Illusion of Validity

Sheena Chestnut Greitens: The third bias that you talk about is the illusion of validity. And this strikes me as being a little bit similar to the planning fallacy, but how should we understand the illusion of validity and why it's important, again, in this context around emerging technologies and strategic stability?

Rose McDermott: I think it goes back to the Type 1, Type 2 thing in the sense, you're right, that they're connected, but I think it goes back to Type 1 or Type 2, in the sense that, what happens with the illusion of validity is that, we're just more likely to believe something will happen if it has a good narrative, or a good story around it.

So this relates to how people judge the likelihood or probability of an event happening in the future, right? So how likely is it that you think a nuclear weapon will be used. And people may have some abstract notion: 10%. But then you say, gee, you know, Putin's flank is collapsing in Eastern Ukraine, he's run out of men, there's an economic crisis. He thinks the way he can get out of it is to use a teeny tiny nuclear weapon in, you know, Ukraine. What's the probability of that? And then people will say, oh gee, 40%, right? So if you tell a story that makes sense, people judge the probability to be really different.

So it's the illusion, right? Because the overall probability of using a nuclear weapon is way more statistically, than using one in Ukraine. But because you can tell a story, a coherent story, about how it would happen in Ukraine, people judge that to be more likely than the overarching thing, which statistically is

more likely, right? So you're using your gut feelings as opposed to the abstract, you know, boring, Type 2 sort of way of thinking about things.

And the reason that I think that that becomes problematic is that, it means that people who make decisions, are much more likely and much more easily manipulated by those who can tell a good story, right?—who can say, oh, if you don't do this, then this will happen–and that this looks like a coherent story. Well, you know, if his flank is collapsing, if he's run out of men. Xi Jinping will invade Taiwan because he wants to control the computer chip industry. You know, if you can sort of tell a story, it makes people much more likely to believe that that's true.

But statistically, it doesn't make it more likely that it's true. So you make a decision based on something that's an illusion of what the actual probability is.

[00:16:05] Bias Four Prominence Effect

Ryan Vest: As emerging technologies increase time pressures on leaders, the dangers of the prominence effect, which is your fourth bias, become a lot more apparent. Can you explain how this effect changes the pressures of shorter timeframes, and how it affects decision-making?

Rose McDermott: Yeah. So, the prominence effect, you know, back in the old days would've been called guns versus butter, right, so that you have this sense of the trade-off that you invest in, let's say, social welfare programs versus what you invest in the military. And, the prominence effect says people really privilege security, right?

They really privilege a sense of safety. And so they're willing to spend enormous amounts of money to achieve that, or to think that they achieve that, even if it comes at a tremendous cost to social welfare programs to, you know, domestic programs, healthcare, education, you know, the other things that you care about, because it's considered to be much more important.

In a fundamental psychological sense, that makes sense, right? Like even Maslow's hierarchy of needs, the most fundamental thing is to feel safe and secure. The problem is that at some point when you privilege that over the other aspects of society, you actually undermine your own safety, but in a different way.

You're undermining your safety for things like erosion of democracy, you know, civil war, you know, things that happen inside societies, as opposed to things

that happen, you know, between societies. And so part of the problem is that that's very instinctual, you know, it's hard to tell people–oh, don't worry about being attacked. What you really need to do is invest in, you know, long-term education policy.

People are like–what? It just kind of doesn't make intuitive sense, but that doesn't mean there isn't a tipping point where people can really privilege security in a way that actually undermines their own actual safety, as opposed to their, you know, psychological sense of safety.

[00:18:01] Rethinking Rationality

Sheena Chestnut Greitens: So I listen to you walk through these biases, and I'm actually just struck. You know, I remember when I first read people like Schelling in graduate school. Deterrence seems so fundamentally rooted when we first learn it in rationality and this human, very rational cost-benefit calculation, about the risks of escalation in nuclear war.

So when I'm listening to you, I guess I'm thinking if there's one thing, or one assumption that policymakers, military leaders, presidents, other decision-makers in the international environment—one assumption they should unlearn after listening to, or reading your work, or listening to this conversation— what would it be, and what would you replace it with?

Rose McDermott: Well, I think that it's that people aren't as rational as we like to believe, or as policy-makers and others like to believe. For me, it's a real difference in kind of disciplinary foundational assumptions. And by that I mean Schelling—and he was a genius, I don't deny that—was an economist, right?

And so the notion of rationality there is this kind of cost-benefit sense of rationality. You know, when I tell my psychologist friends–hey, did you know that political scientists believe in rationality? They sort of laugh hysterically and say–why would they do that? What a crazy thing to believe.

You know, it's like telling somebody you believe in, you know, some really obscure, you know, conspiracy theory or, you know, other kind of wacky idea, because that's just alien to the notion of the descriptive accuracy of how people actually think, and how they actually behave. It doesn't mean that they aren't rational.

They're just not rational in the economic cost-benefit sense that economists mean, and part of it is a different sense of timing. So for an economist, they're

really thinking very short term. Even if they think that it's long term, like 50 years, that's nothing when you think about the history of humanity, which is millions of years old, right?

And so their notion of rationality is like, can I make money on this in the next, you know, 10 years? And it's also always about money, right? There's never this kind of like, what really matters is–Are you thirsty? Are you hungry? Are you tired? I've done various work with a lot of special forces guys, and I don't know that many of them would take, you know, $10 over water, if they're really thirsty, right?

Like, it's just not how human nature works. And so, from the psychological perspective, there is a rationality, but I think of it as ecological rationality. And by that, I mean, it's about the likelihood that you're going to get your genes into the next generation.

I am absolutely mortal, but technically I'm immortal. My genes are immortal if they keep going into the next generation. In my case, that's not true because I don't have kids, although there's, you know, cousins and things like that. But if you behave in ways that increase your probability of actually pushing your genes into the next generation, that's rational.

It's not about money. It's not about what happens next week. It's about what happens across millions of years. And even if the differences from behavior are super tiny, you're still talking about a neuro-computational program that works across billions of people, across millions of years. So super tiny differences in the way people have learned to behave that promote the survivability of their ancestors, increases the likelihood that you'll see it in the current generation.

It doesn't mean it's adapted to the current environment, right, because it's backward looking, but you have a sense where there is a real rationality. It's just not economic rationality. And so when I think about it, that's the way I think about how real people make real decisions. It's about survivability, which may be why you get things like the prominence effect. But it's not about making more money, and it's not about short time spans.

And I think there's a really simple example. You know, if you look at really devastated war regions in Sudan, in Gaza, in, you know, Ukraine. What keeps those people alive is not money. It's not an economic exchange. It's a series of social networks that allow them to help each other in ways that maximize the probability of surviving. You basically eliminate economics in those situations,

and that's how I think about what real rationality looks like. It's social networks. It's actually not money.

Now that doesn't mean social networks don't have problems. I mean, I hate social media, but that's a different thing than I'm talking about. I'm talking about real social networks. So it's not the Facebook thing where you have, you know, a thousand friends. It's that in real life, who can you depend on to drive you to your colonoscopy, right? That is something that not as many people have now, than had 20 years ago.

And that decline in social networks is what I think makes people increasingly anxious, depressed, detached, and that, you know, is a problem that I think social media contributes to, but is not the only thing that contributes to it.

[00:23:11] Evolutionary Mismatch Today

Ryan Vest: It's really fascinating because in later sections, you bridged cognitive psychology with evolutionary psychology, and you argue that many of these biases were once adaptive, but now they're dangerous. Why has human psychology failed to keep pace with the destructive potential of modern weapons and technologies?

Rose McDermott: Because reproduction is a lot slower than technological innovation, right? So evolution is like trying to fix the car while it's running. And it's going to happen across generations. It's not going to happen from OpenAI, ChatGPT Version One, and OpenAI ChatGPT Version Four, you know, which occurs across a couple of years.

It's just not even close to happening in the same kind of time. And just because it worked in the past, doesn't mean it'll work in the future. And, you know, that's the way I think about nuclear deterrence, and why I really worry about it is that, just because we haven't had a nuclear explosion in war in 80 years, doesn't mean that we'll never have it happen.

And, you know, the way that I think about this maladaptation in the modern world has to do with sugar preference, right? So, most people, including babies who are hours old, will show a strong preference for sugar water over plain water. We just have a very strong inbuilt bias, because our ancestors who lived in environments of incredible food scarcity, were much more likely to live and have kids if they liked honey as opposed to, you know, leaves.

And so we've kept that preference. But in the modern world where I can go to Safeway and buy bags of Reese's Peanut Butter Cups, that is not a good strategy for putting off things like diabetes or cardiovascular disease.

Now, you know, millions of years ago that wasn't a big deal because nobody lived to be 60, or 70, or 80 years old, right? You died much sooner. And so the evolutionary system didn't have to adapt to people living to be a hundred years old. So they didn't have to worry about things like cardiovascular disease if you liked honey.

Not that honey causes cardiovascular disease, but you know what I mean. Like, if you have a sugar preference, you're much more likely to eat more of it, gain weight, have higher risk for health disorders, particularly in older age that weren't a problem in the past, because nobody would've been able to get that access to that much sugar, and they wouldn't have lived long enough to have complications from it.

[00:25:35] Authoritarian Leaders and Risk

Sheena Chestnut Greitens: Let me switch gears for a minute here and ask about one of the things that was one of my favorite threads, that you pulled a little bit in the article, which is about the nature of regime type in the modern nuclear era. And one of the things that you noted is that a lot of the countries that have nuclear weapons right now, in the world today, are run by personalist or authoritarian leaders.

And the role of those leaders in those systems may be particularly susceptible to bias, because any bias held by an individual isn't as subject to institutional constraints, even if it's through a party institution, or a military institution, or the democratic process.

So does that—I'm really trying to think through what that means for strategic stability—does that mean that it's become much more dependent on regime type, and on the particular individuals, who are the leaders of nuclear armed states today?

Rose McDermott: Yeah, I think there's two parts to that. The first is we just have more nuclear armed states than we had in, say, 1945. And so when you just think about the probability of interaction among nine actors, as opposed to two actors, that just raises the probability of accidents, you know, inadvertent use, but also, you know, intentional use.

I do think that when you have leaders who are less constrained institutionally, it has particular consequences. And it's not like that never happened, right? Stalin didn't have a lot of constraints on him, but at that time there was also one other nuclear power, right? So it wasn't like trying to deal with this three, or six, or nine body problem.

But with individual leaders where they have either personalist dictatorships, or authoritarian tendencies. I think you get a couple of things. One is exactly like you suggested, where you have less institutional constraints, and less norms binding people.

But secondly, I think, what really happens is that you surround yourself with sycophants, right? You surround yourself with people, where you privilege loyalty over competence. You want them to be really loyal to you no matter what–to believe me, or your lying eyes, believe me, right? Like, it's one of these kinds of things where: I want you to be so loyal, that I actually don't care if you're very good at your job. I just want to make sure that I can trust you.

What that means is that there's a much higher risk of getting incorrect advice, and I think that was part of what happened with Putin going into the war in Ukraine, where he thought it was going to be over in three days. I think that had he had more sophisticated advice that said–hey, do you remember that after the Second World War, Ukraine kept fighting both the Germans and the Russians?

Like, this is not a population that gives up easily, right? Like, you don't imagine that he got the most sophisticated advice, because he's going to surround himself with people who agree with him on principle.

It also means that there may be problems with the competence of maintaining certain weapons. I think about this, for example, you know, in the Russian system, they have to, I think, replace tritium in some of the weapons, you know, every 10 years. It's really expensive. There's a handful of guys who do it. So there's an incredible incentive for those people to kind of sell the stuff on the open market, for their own wealth. And so that reduces the probability that those weapons are dependable when you use them.

So what happens, for example, if Putin launches a nuclear weapon and–oopsie it doesn't work. You know, that is a very interesting problem. And you can imagine other ways that surrounding yourself with loyalty, and not competence, can get you into real trouble in other circumstances, not just nuclear weapons.

I think about that with the pandemic, right? There were a lot of countries that did a lot better than other countries, and it had a lot to do with the institutions that were in place in public health, and the trust that happened between the population and the public health advisors. In the United States itself, it really differed, you know, city by city. But the United States did not do it very well relative to countries like New Zealand, for example. And so, you know, when you have those kinds of situations, that's particularly problematic.

The last thing that I'll say is that when you have these authoritarian tendencies, leaders don't learn from their mistakes, or from other people's mistakes, the way that those in a democratic regime might. And I think there's two reasons for that.

First of all, they don't learn from their own mistakes because they don't believe they make mistakes. They believe that if something goes wrong, it's his fault–it's some other guy's fault, and I'm going to take him out and then it's all going to be good. But they don't learn from others' mistakes because they think they won't make those mistakes.

And they don't have other people pushing back on them. And so learning is really restricted. And that matters both for crisis bargaining, but it also matters for technological innovation over time. And so I think that those things become, you know, real limitations over time in, you know, the stability of those regimes, but also raises the risk that you'd have accidental, or inadvertent use, not just of nuclear weapons, but biological innovations that are gain of function, artificial intelligence that tells people to kill themselves, you know, like you can imagine it going off in all [kinds] of different directions.

[00:31:14] Policy Fixes and Safeguards

Ryan Vest: I'd like to turn a little bit to policy implications, to build on what we've been talking about. You argue that decision-making itself should be treated as a skill that can be trained and improved. What kinds of institutional safeguards or decision processes are most realistic for mitigating the prominence effect in real world crises?

Rose McDermott: Yeah. You know, I'll say something about places where this has worked first, which is that when you put certain protocols in place that seem really simple, it can make a huge difference. So the best example I can think of is: there was a series of protocols that were put in place in hospitals, I don't know, 10 years ago, where before, for example, you do a surgery, you have to go through this checklist of like 10 things.

And they're really simple things. They're things like, did you wash your hands? Do you know which leg you're actually operating on? So you put a button on the leg that says yes, and a button on the leg that says no, so that you don't do a knee replacement on the wrong leg. Within a year, it saved a hundred thousand lives. I mean, it just had this enormous impact.

And so, you know, I think that there's other policy things we can put in place when we think about strategic stability. I think with regard to nuclear weapons, one of the really simple things that we can do is that, as the memory of the use of nuclear weapons dies off among policymakers, right—you get policymakers who were not alive when Hiroshima, Nagasaki happened, or you know, weren't old enough to remember the Cuban Missile Crisis, or whatever—trying to make it more salient before they have to make a decision with vivid imagery, like watch a one minute movie about how bad it was, could help.

Having a checklist and a protocol of–go through these procedures. And one of those things should be what are your central values, right? Like your central value isn't just getting back at the guy who just attacked you, but do you want humanity to survive? Do you want civilization to survive? Do you want to be the one who precipitates the 6th extinction? You know, like really simple ways to remind people of the broader, larger values.

I think that making sure that the choice protocols include non-nuclear response, or non- chemical, or biological response, so that it's not just like your only choices are to do this incredibly destructive thing, but that you have other choices. The use of conventional weapons, you know, cyberattacks, whatever it is that allows the possibility to think more broadly in the moment.

And I think the hardest part is about time pressure, which is that as technology gets more and more advanced, things happen faster and faster, we're more likely to relegate it to AI. But AI is simply, you know, the evidence I've seen is that it just replicates human psychological biases at scale: racism, sexism, you know, and the other biases we talk about.

So, turning it over to AI just makes those mistakes faster. I don't think it makes it less likely, because the people who program them are people, and they share those biases. The algorithm may be more predictable, but it doesn't necessarily eliminate the prospect of the biases that we've been talking about.

And so anything that we can do to sort of slow things down, to allow a little bit of time for the person to be like, okay, is this really what I wanna do? Is it how I do it? Making sure it's not entirely AI, right?—that Doomsday Machine thing. I

think any of those kinds of things at the margins may prove helpful, but the last thing is like with regard to nuclear weapons.

I really think moving away from sole use authority is really important, right? And Bill Perry's been, this has been Bill Perry's thing, that he's really cared about now for decades, where not just in the United States, but elsewhere, you know, that it can't just be one person who decides–I'm going to use nuclear weapons–that you have to have other people.

Now in authoritarian regimes, that's going to be difficult. You can imagine that it would be very difficult for somebody to say no to Putin. But, at least there'd be, you know, there's a structure in place for others to be involved in that decision in Russia, in a way that's not true in the United States. So, you know, thinking through what collective action would look like in those moments, I think is important.

Sheena Chestnut Greitens: It's fascinating, and you mentioned this, that these biases are known to create potentially huge problems, or risks in human decision-making, when you're talking about stakes like the use of nuclear weapons, and yet there are things, and there's evidence showing, that some of this training and skill can actually be improved with practice.

We recently ran a piece at *TNSR* that looked at overconfidence, and one of the findings was by Dr. Friedman. And it talked about the fact that if you give people a couple minutes of training that shows them where and how they're overconfident, they get better.

And I found that really fascinating.

Rose McDermott: That's true, particularly with overconfidence. It really can make a difference.

[00:36:11] Can AI Be Used Responsibly

Sheena Chestnut Greitens: So, I guess I wanted to ask, you know, you talked about the risks of AI in decision-making, and that this is not, this can replicate human biases at scale. Are there ways to use AI as an emerging technology responsibly in decision-making, or is this really just a–this is risky, we should constrain it, and slow it down?

Rose McDermott: You know, it's such a good question, and I don't feel like I have a good answer for it. And part of it is that I think it's new enough that we don't actually know, right? What I've seen is no one showing an interest in constraining it, right? Not any of the serious tech bros, who actually do the coding and stuff like that. I don't see them wanting to constrain it.

You get even, you know, this recent thing with Elon Musk not wanting to stop being able to make people naked on Grok, right? Like, if they can make money off of it, that's what they care about, and not the larger societal consequences. And so whether it can be used responsibly is, I think, different than the question of whether it would be used responsibly.

There may be ways to do it. I know that in the face of some lawsuits, where parents have accused some of these chatbots of teaching their kid to kill themselves more effectively, like, you know, the person types in–how do I do a noose, you know?–That they've tried to make the bots more sensitive to those kinds of things, to be like–maybe you shouldn't kill yourself, you know, maybe you should see a real therapist.

But the problem, again, is you get upwards of 30% of people now saying that they'd rather talk to ChatGPT than their doctor. Now, on the one hand it makes sense, your doctor doesn't have enough time, and your doctor probably isn't as patient, or as nice as, you know, a sycophantic machine.

The problem is the sycophantic machine makes stuff up, right? It's going to hallucinate a lot of stuff. And so it's hard for me to know whether, I mean, obviously AI has particular uses, and I think it could be super effective for things like generating new medications for cancer or, you know, new antibiotics and things like that.

But I think in human interaction around decision-making, a lot more effort would have to go into actually constraining some of these biases, and building the algorithms, than I have seen evidence so far. It doesn't mean it can't be done, I just don't see any interest in any of the guys building it, to have it be done.

And I think that's unfortunate, but I think it's also–they replicate themselves. You know, there's a particular kind of person who does that kind of coding, and that kind of person is really interested in technological innovation, and making money. And, you know, that's a different set of goals, necessarily, than protecting the, you know, ethical integrity of human dignity.

[00:38:51] Vetting Leaders and Wrap Up

Ryan Vest: So we've covered a lot of ground today, which is great. This is one of the best conversations we've had in a long time. As I think about all of this, if you could change one thing about how leaders are trained to prepare, or to make these decisions involving highly destructive emerging technologies, what would it be?

Rose McDermott: You're not going to like my answer, but my answer is I think that they should be psychologically vetted before being allowed to be in office, and medically evaluated, right? I think that there are particular personality disorders that are different than biases, right, that are real disorders, that are particularly destructive: narcissism, psychopathy, paranoia. And if you have someone who really has that and, you know, look around the world at the main leaders now, they've all got it, right? Like, this is not a peekaboo–It's like where's Waldo, right? It has a particular pattern.

So for example, if someone is a narcissist, the minute you prick their ego, they will respond with aggression. It is like turning on a light, and having the light go on. It's just automatic. That's dangerous. So if we had a way to vet, and, you know, there have been lots of discussions of this in the presidential health literature, for years about health, like cancer and stuff like that. But I think there should be psychological vetting, that if they can't actually pass, and I don't mean cognitive tests, I really mean personality disorder vetting, and have certain thresholds that if somebody is you know, above or below a certain threshold, they're not allowed to run.

Is that realistic? No. But I think it would be the single most effective thing that you could do. I just don't think it'll happen.

Sheena Chestnut Greitens: That would be a tough bar to clear in any political system, but especially, frankly, some of these authoritarian ones, where any constraint is difficult to conceive, and if anything, sometimes the trend over time, in certain political systems, is toward personalization, not toward more constraint institutionally, or otherwise.

Rose McDermott: I think of this in very evolutionary terms, right? So, like on my wall, you can see I have pictures of chimps. I don't know if you can see it. It's, you know, in the olden days, right? And I don't just mean humans. I mean, you can still see it in chimpanzee and gorilla populations now. The alpha is usually, you know, the most hierarchical male, is usually someone who actually takes care of his community. And if he starts exploiting that community, what

happens is that the beta males get together, and they toss him from power. You know, in a lot of animal worlds that looks like assassination, you kill it.

But in the modern world in democracy, you throw them out of office. The problem is, it's very hard to track exploitation in large communities, right? So if you're a group of 150 people, you can totally tell if your leader is exploiting you, in a group of 330 million people, that's much harder.

And so I think that it's very difficult to think about how you'd get, you know, less powerful people to try and overthrow a more powerful person. And, like when you see the case of the generals who tried to overthrow Hitler–what happened was he found out and he killed them, right? And so you have to be someone, where you actually believe that everybody else in your group gets more value out of being in your group, than telling the leader–oh, these people are out to get you.

He kills them. And then you rise in power. And so that trust mechanism, around people who are in a position to overthrow, doesn't work logically. And I think it makes it very difficult to overthrow leaders, you know, short of, like what happened to Saddam Hussein, or something like that.

It doesn't mean it can't happen, it's just more difficult.

Sheena Chestnut Greitens: Well, and there are often significant costs along the way. If you think about what it took to topple some of the recent leaders who've lost to power–the Civil War in Syria, and the human cost of that's pretty staggering.

Rose McDermott: It's enormous. I mean, just look at the riots in Iran now, right? I mean, it's just, and who knows if that'll be successful? You may have hundreds and thousands of people who die, and they're not capable of pulling the regime down. So that's the challenge. You know, who has control of the instruments of coercion?

Ryan Vest: Well on that wonderful and uplifting note, I think we've about hit the end of our time. Rose, we want to thank you very much for joining us today.

This has been a really interesting discussion.

Rose McDermott: I'm just really grateful for the chance to have this conversation with both of you, and really grateful for the chance to air the arguments in the *Texas National Security Review*. And I think that the other

articles in that collection are really valuable too, and so I'm hoping we can just continue the conversation, but I really appreciate the thoughtfulness of the questions that you've asked.

Ryan Vest: Thanks for joining us on *Horns of a Dilemma* from the *Texas National Security Review*. Our guest today has been Rose McDermott, author of the article, "The Influence of Psychological Factors in the Search for Strategic Stability," which as always can be accessed for free on our website, *TNSR*.org.

If you enjoyed this episode, be sure to subscribe and leave a review wherever you listen, and you can always find more of our work *TNSR*.org. Today's episode is produced by *TNSR* Digital and Technical Manager Jordan Morning, and made possible by the University of Texas system. This is Ryan Vest and Sheena Chestnut Greitens. Thanks for listening.