This is a pre-publication draft version of:

Mandik, Pete (in press). Sliders. For a special issue of *Journal of Consciousness Studies* edited by François Kammerer and Keith Frankish on their target article "What forms could introspective systems take? A research programme."

Sliders

Author

Pete Mandik (William Paterson University)

Bio

Pete Mandik is Professor in the department of History, Philosophy, and Liberal Studies at William Paterson University of New Jersey, where he teaches courses in philosophy, psychology, and cognitive science. He is author of *Key Terms in Philosophy of Mind* (2010), *This Is Philosophy of Mind: An Introduction, 2nd Edition* (2022), and *Physicalist Theories of Consciousness* (under contract). He writes and draws the *Mind Chunks* philosophical comic strip at dailynous.com. He hosts the philosophy and science podcast, *SpaceTimeMind* (spacetimemind.com).

Abstract

“Sliders” are a speculative introspection-enhancing future technology allowing humans with cybernetic brain implants to precisely and voluntarily modulate moods and other mental states that vary along a one-dimensional scale. Such future humans may, for example, use the Sliders interface to temporarily present a COWARDLY-COURAGEOUS “slider” in their visual field, and with a mere act of will change their level of courage from a 60 to a 65 on the 100-point scale. The present article discusses the implications of such a technology in the form of an epistolary fiction in which the author’s future persona warns the reader of the dire consequences and hard-earned insights arising from wide-spread Slider use and abuse.

Hello from the future. I write with a dire warning. Never mind how, or from when, for none of that matters. I hardly understand it myself. Just trust me when I say that we've solved all of the time-travel paradoxes you can think of, plus a few more. But do understand this. Despite the vast powers we future humans have at our disposal, a horrible doom awaits future humanity unless past humanity, you and your contemporaries, act immediately to avert it. And that terrible doom bears this simple name: "Sliders"

"Sliders" is what we call the number-one all-time most popular app for the brain-computer interfaces that each of us future humans have implanted in our brains. The

very idea for Sliders derives from several science fictions of your era (Egan, 1998; Doctorow & Stross 2012; Bakker, 2015). The name "Sliders" comes from the primary user interface in the app, an interface inspired by interfaces that you past humans are quite familiar with. You encounter them frequently in your personal computing and smart-phone technologies. If you've ever adjusted the audio volume on one of your devices, you've seen a slider. You've seen a visual representation of a scale, from loud to quiet, as a line segment along which you slide a tab to any desired position from the very quiet to the very loud. Such virtual sliders are of course modeled on the earlier technology of physical slider switches, as you may find controlling a dimmable light source. Our sliders, here in the future, appear at will in our visual fields and allow us to directly modulate our own mental states.

Suppose it occurs to you that you're feeling sad and you wish to be less so. If you have the latest Slider software running on your brain implant, then with just the slightest mental action, the faintest act of will, you command the interface to present you with the relevant slider in your visual field. In this case, the relevant slider is one representing a range of moods linearly arranged—your SAD-HAPPY slider. To assist in precisely setting the slider, the private visual display includes a numerical display, so that you can see, privately, that your current mood is set at 40 (with 1 being maximally sad and 100 being maximally happy). As tempting as it might be to jam the slider all the way up to 100, you may nonetheless choose to just bump it up to a modest 60—just enough to put a bounce in your step, but not so much as to call up a torrent of overwhelming ecstasy that would make for an embarrassing public show. Self-cheered, you continue about your day.

This is just a simplified example, designed to ease you into the cold water of understanding this terrible technology. Before we go on to consider more complicated examples, examples that hopefully put the requisite terror into you, let's pause to consider some key similarities and differences between Sliders and old-fashioned, standard-issue, un-augmented introspection.

To bring your own introspective faculty into view the old fashioned way—which, for you, living in my relative distant past, is the only way you can—suppose someone were to ask you what mood you were in. Suppose they asked specifically how happy versus sad you are. What would you do to answer that question? One thing that you might do is *introspect*. You would attempt to shift your attention inward, so to speak, and attend not to the external world, to your immediate physical environment, but to attend instead to your own mind. One immediately apparent difference between old-fashioned introspection and Sliders is the difference in degrees of precision between the two. With Sliders, you can see in just a quick, albeit massively assisted, glance that your mood is a 60 on the 100-point SAD-HAPPY scale. Contrast this with someone asking you, an old-fashioned, to express your mood numerically on a 100-point scale. Likely you'll feel like you would just be shooting in the dark—you couldn't put that precise of a numerical score on it. You might feel like you can only pin down your mood with a coarse grain partition of 1. extremely sad, 2. kind of sad, 3. neither happy nor sad, 4. kind of happy and 5. extremely happy.

The really marked contrast between Sliders and old fashioned introspection is not the precision in the *self-knowledge* enabled, but instead the immense scope of the *self-control* it facilitates, where increase in self-control here means increase in scope and depth of self-determination. Hopefully this gets clearer for you with further examples. Suppose you wanted to muster the courage to march into your boss's office and ask them for a raise. Without sliders, what would you do to gather up your reserves of bravery? Try to imagine yourself as a favorite courageous fictional character? Give yourself a pep talk? Look long and hard in a mirror and tell yourself "You can do it!"? Whatever the attempted efforts, how confident are you now that you could raise your courageousness beyond a faint nudge? Probably not very. With sliders come the utmost confidence—literally. You call up the COWARDLY-COURAGEOUS slider, and slide the tab right up to 70. Feeling the immediate surge in confidence, you think that perhaps you should tone it down a few notches. Maybe a 70 will be too much swagger, and will negatively influence any ensuing negotiations. You roll it back a bit, deciding the 67 feels about right. But anyway, notice the precision and ease of the literal self-control you can access with Sliders. Further, you can continue to make adjustments on the fly (although it does take some practice to make sure you don't look like someone attending to their private sliders, which can be a turn-off for your interlocutors). While talking to your boss, you may decide, based on certain turns of the conversation, to nudge up to a bold 72, and with other turns, slide it down to a cool 63.

If you can see the appeal of such a system, then you wouldn't be surprised to learn that as soon as this technology became commercially available to the general public, it was a massive hit. Once a person grasps the basic idea of Sliders, the advertisement copy practically writes itself.

- A student flagging during a late-night study session tweaks his INTEREST slider and is instantly able to overcome what would otherwise be debilitating levels of boredom. Despite the late night, further slider nudges keep their concentration and energy levels up long enough to enable them to ace the big exam before they return home for some much needed sleep.

- A husband returning home from an especially stressful day at the office bumps down his ANGER slider. Coincidentally, his spouse, who was expecting him home for dinner an hour ago, bumps down her own as well. Mutually respectful slider adjustments become the cornerstones of successful relationships here in the future, and couples review their adjustments in weekly audits (Bakker, *ibid*.).

- A trio of young women, returning to their shared apartment after a late night party, adjust both their COURAGE and their AGGRESSION sliders, giving them the right level of nonverbal “do not mess with us” signals to ward off any would-be assailants during their walk home.

Like many of my fellow engineers of the Slider interface, I myself was an early adopter of Sliders. Like many in that first wave of adopters, I take pride in the skills with sliders that I've curated and cultivated over the years. I'm using them right now as I compose this message to you. I have a personalized recipe of INTEREST and

EQUANIMITY that I consider essential for intellectual work such as composing the current communication. Layered on top of that I have some additional slider settings customized for this particular communication, requiring me to keep a judicial yet supple control of my HORROR slider. I do this to strike the optimal balance between feeling sufficient urgency to compose this warning, but not so much as to be plunged irretrievably into despair. Please forgive my delay in conveying the warning part of this message. There are further basics about Sliders and Slider culture that you must understand.

Besides the skill that we early adopters take pride in, we additionally take pride in our taste and restraint. Aside from the literal danger of slamming any slider up to 100 or down to 1, we're guided by a refined aesthetic of *less-is-more*. This carries over into a benign Luddism—not only do we value avoiding extreme slider settings, we also take pride in often refraining from slider use altogether—like a classic samurai who only unsheathes their blade when absolutely necessary. As is typical with generational gaps, the more restrained find themselves on the older side of the chasm. Newer users were less subtle than us early adopters in what they pushed their Sliders around for.

I'm skipping over a lot of detail in this potted history of Sliders. We didn't jump right into commercial availability. There were many iterations of safety review and testing. Some of the problems we ran into were anticipated by science-fiction authors of your era.

Right up front we were worried about "wireheading," a concept inspired by scientific work in the 1950s. Scientists put stimulators in the pleasure centers of rats' brains. The rats died of starvation and dehydration when stimulating their own pleasure centers became their dominant priority (Olds & Milner, 1954). Science-fiction authors picked up on the idea (Niven, 1991; Robinson, 1982) and explored wireheading as the most dangerously addictive possible human activity. Later, AI researchers applied the idea as a problem for AI training systems that have reward functions, where "wireheading" refers to AIs hacking their own reward channel (Yampolskiy, 2014). Wireheading is perhaps the most pernicious and extreme exemplar of "Goodhart's law," which can be stated in as the adage that "when a measure becomes a target, it ceases to be a good measure" (Strathern, 1997).

For researchers of my own era, working on the early development of Slider technology, our immediate fear was that a PLEASURE slider would be cranked all the way up to 100, rendering the user catatonic, locked into a recurring loop of maximal ecstasy, immobilized and rigid. We found out in early human trials, conducted on prisoners, that they will indeed die like the rats wired in the 1950s.

Another tragic outcome we identified during the trials on prisoners is one that researchers labeled "the bloody stumps problem". In groups of trial subjects that had access to anything at all resembling a WILLPOWER slider, they would drive their bodies to damaging extremes of performance. One especially gruesome example was an incident that gives the problem its name. A human subject used their WILLPOWER slider to fulfill a whim to walk across the continent of North America. The decision to

make this journey was truly made on a whim. There was no preparation. They didn't even have appropriate footwear. But (as revealed in a postmortem forensic Slider audit) they did happen to have their slider for willpower pumped up to almost 100. As soon as the idea to walk across the continent entered their mind, they immediately got up and set out on their voyage. They were not deterred by blisters or by a sprained ankle. Their handlers were authorized to intervene only if they presented a threat to others. Self-harm was allowed so that I and my fellow researchers could see how far they would go.They were allowed to continue without intervention and literally walked themselves to death, bleeding out after crossing the finish line on "bloody stumps," the extremities of their limbs having been ground down in the absence of the life-saving weakness of will that would have stopped and spared an un-augmented human.

It became clearer in the later stages of the human trials that we are going to continually need to monitor for new uses and potential abuses of the sliders. During gamified economic simulations, we found that the slider users were messing with their own shopping preferences. This had disastrous effects on their simulated economy, effects that would obviously spook our corporate investors (cf. Bakker, *ibid*.). We realized that we were not going to be able to simply hardwire a solution to the anticipated problems, and needed a way of remotely monitoring and remotely installing fixes. Thus every brain implant was continuously, and largely outside of the awareness of the user, streaming info to governmental and corporate watch dogs, as well as to their AI proxies. Thus was the famous, or infamous, *Slider-Monitor "Paradox"* born. Paradoxically, the more that humans sought to expand their self-knowledge and their self-control, the more they expanded the domains of things to lack of knowledge of, and to lack control over. Everyone wanted to expand the scope of their self-insight and self-determination, and they wound up instead with a cop in their head, literally monitoring their thoughts and reporting back to a central command that wielded remote access.

There were worse problems to come once we made the Sliders app and the brain implants to run it available to the general public. The rollout was widespread and inexpensive for consumers, thanks to generous governmental subsidies and corporate incentives. With an increased base of Slider users came an increased exploration of the forbidden and unknown potentials of this technology. A growing and dangerous proportion of users sought intensities of experience that no human should have been exposed to. Thus emerged the underground Slider culture of intensity-hacking.

One kind of intensity-hacking that the DIY Sliders crowd loved was "double dipping," an activity anticipated by a writer of your era (Bakker, *ibid.*). Consider some stimulus that is likely to get a strong reaction from you, like maybe a comedian who you think is pretty much always funny. And suppose that before going and seeing this comedian's live act, you slide your HILARITY slider to its maximum setting. When the jokes hit you, you laugh so hard that the risk is extremely high of breaking a rib, asphyxiating, seizing, and slipping into coma. You have "double dipped"—your natural mirth response was amplified by the artificially boost from the Sliders app. Each time some doomed explorer stumbled upon a fatal (or near-fatal) slider setting, we uploaded another block-and-lock routine to the implant-monitoring AI watchdogs. Upon detecting

an escalating "double-dip" toward the red zone, the sliders app would reset the user's Slider levels to some pre-implantation baseline, and lock the user out of Sliders access for a 60- minute stretch. Hackers made most of their income by jailbreaking the implant monitor-and-lock systems for the machine-headed experience junkies. National governments dedicated vast numbers of their police and military personnel to dealing with this lucrative and fast growing new crime. Societal chaos escalated as the casualties of the War on Hackers far superseded its precursor, the War on Drugs.

Despite the emerging societal costs of allowing the continued existence of the DIY hacker culture, the heads of our industry were loath to eliminate this comparatively inexpensive source of research-and-development. One of the most profound innovations to emerge from the hacker scene was an AI-assisted implementation of Doctorow and Stross's (2012) "fractal" scheme of nested sliders, wherein each slider can be decomposed into a branching structure of four additional, and more fine-grained sliders. For an example directly out that inspiring fiction, "ANGRY-DELIGHTED" can be decomposed into the separate sliders "FED UP-RESIGNED, SICKLY FASCINATED-CONTEMPTUOUSLY ALOOF, RIGID-INCANDESCENT, ASHAMED-RIGHTEOUS", each of which can be further decomposed, until the slider labels eventually abandon natural language descriptors in favor of "a specialized set of intricate ideograms that appear to categorize all human experience as belonging to one of several million recombinant subjective states." A problem some of us suspected, but had no idea how to evaluate the potential risk of, was that by plumbing depths that exceed our natural-language and folk-psychological understanding, the DIY hackers dredged up some literally unspeakable horror with unforeseen and unforeseeable potential for destruction. Naked greed kept us from lingering to contemplate these possible ramifications.

For the first few years of the commercial release of Sliders, our problem-detection and mitigation strategies were able to keep up with the trickle of emergent ways to kill yourself by overdosing on experience. Yearly fatalities were low; it was one of the safest lifestyle purchases a person could make (as long as one stayed on the right side of the law).

When things truly fell apart globally, the fall was as rapid as it was cataclysmic. Widespread use of Sliders gave rise to death, insanity, and destruction in an ever-widening and recurring circle of death, insanity, and destruction. The leading edge of the death wave was the DIY hacker culture. The Slider hackers were seeking the extremes of human experience. Working in groups, each group member goaded the others to even further extremes. They sought not just extremes of intensity, but also novel combinations of extreme experiences. With their amazing new controls they now wielded amazing powers to edit their base personalities. If a peer suggests yoking the intensities of erotic pleasure to the depths of fear and torment, then whatever natural inclinations might otherwise have been inhibiting—perhaps a feeling of revulsion or horror in the face of madness—could now just be suppressed with a few bumps to the relevant sliders. Before long, rates of violent crime made a sharp uptick. And even if it wasn't for the testimonies of the few surviving victims, it still would have been manifestly evident that the perpetrators were intensity-hacked Slider users. Investigators arriving

on the crime scenes could see this right away from the state of the victims' grotesquely ruined bodies. It was obvious to investigators that, while still alive, the victims were subjected to ordeals of cruel ritualistic complexity—levels of complexity and cruelty that could only be perpetrated by someone with their PERSISTENCE and SADISM sliders turned way up and their EUSOCIALITY and REMORSE sliders turned way down.

Despite efforts to suppress their own revulsion at their abominable acts, some perpetrators of these murders succumbed to old-fashioned remorse and despair. The old natural inclinations were increasingly finding their ways past the vigilance of the self-sculpting DIYers. Murders were increasingly supplanted by murder-suicides. And it became clear in investigations of multiple such incidents, that a lot of the suicides in the murder-suicides were due to the murderers not acting fast enough to decrement their REMORSE slider. This is precisely how I lost my spouse after she murdered our three children while I was away at a company retreat.

Those of us who have so far survived have had to tamp down more and more of our humanity to endure a world sliding deeper into what cannot be called sane. If all that I have reported so far isn't enough to scare humanity away from Sliders, I despair that I don't know what else there is to say. It is probably futile to urge you to appreciate what you have already in your pre-slider, un-augmented introspective capacities, but given the horrors that overrun your future an attempt must nonetheless be made.

I leave you with a summary of the main insights we few who remain think we've learned about old-fashioned introspection, insights that hopefully enhance the vital project set forth by your contemporaries, Kammerer and Frankish (2023). The insights are two. First, introspection is perhaps best thought of in concert with what actions or behaviors it facilitates—self-knowledge makes sense only alongside self-determination. The second insight, which may seem mundane to you but is crucial from my future vantage, is that introspection is already improvable without cyber-surgical augmenting.

Regarding the first point, that introspection is perhaps best thought of in concert with what actions or behaviors it facilitates—self-knowledge in concert with self-determination—it perhaps helps to appreciate this point by thinking of introspection in the context of Darwinian natural selection. Even if the core of the concept of introspection is about self-knowledge, Darwinian thinking encourages us to wonder what downstream effect on adaptive behavior such self-knowledge could possibly have. There are a few points worth making here. One point is that nature is unlikely to care much about how well we understand ourselves, unless that understanding has a downstream effect on behavior. It seems quite clear that, from an evolutionary perspective, knowledge of our environments was far more important than knowledge of our own mental states. Self-knowledge was a "good" but only in a context of relative scarcity. The innate curiosity we have about our mental states is importantly analogous to our innate desire for sugar. We evolved in situations with meager access to it, and when exposed to situations offering abundant access, the effects can be (and will be) disastrous.

Regarding the second point, that introspection is already improvable without augmenting, note that humanity has amassed thousands of years of knowledge about precisely this. That knowledge is codified in various contemplative traditions that provide instruction for the cultivation of greater degrees of control and acuity in self-directed attention. These various traditions for cultivating one's ability to meditate offer many of the benefits that Sliders tempt, but come also with certain safeguards that Sliders culture just steamrolled right over. The main safeguard is how difficult it is to augment introspection in this deeply old-fashioned manner. Changes are incremental and require dedication and effort. The average un-augmented human cannot turn themselves into a murderous psychopath with a flip of a switch. Other safeguards come in the form of communities of ethically committed peers and teachers. None of these safeguards are perfect and there are still risks (Britton, 2019), but none of the risks rise to the level of the existential risks presented by Sliders.

Your generation is poised on the edge of great technological change, and when presented with the opportunity to break down the barriers to greater self-understanding and control, I beg you to keep in mind Kant's bird who foolishly wished for the absence of air: "The light dove, in free flight cutting through the air the resistance of which it feels, could get the idea that it could do even better in airless space." (Kant, 1965). On behalf of the dwindling future of the human race, I bid you goodbye and good luck.

Acknowledgements

For helpful discussions of early versions of some of the ideas here, I am grateful to Brian Kobylarz and David Roden. For detailed and helpful feedback on the penultimate draft, I thank François Kammerer, Keith Frankish and an anonymous reviewer. For the main gloomy inspiration of this piece, I thank R. Scott Bakker.

References

Bakker, R. S. (2015). Crash space. *Midwest Studies in Philosophy*, *39*, 186-204.

Britton, W. B. (2019). Can mindfulness be too much of a good thing? The value of a middle way. *Current opinion in psychology*, *28*, 159-165.

Doctorow, C., & Stross, C. (2012). *The Rapture of the Nerds: A tale of the singularity, posthumanity, and awkward social situations*. Tor Books.

Egan, G. (1998). "Reasons to Be Cheerful." In *Luminous*, 191–227. London: Millennium.

Guyer, P. & Wood, A. W. (eds.) (1998). *Critique of Pure Reason*. Cambridge University Press.

Kammerer, F. & Frankish, K. (2023). What forms could introspective systems take? A research programme. *Journal of Consciousness Studies*.

Niven, L. (1991). *Death by ecstasy*. Orion Press Inc.

Olds, J., & Milner, P. (1954). Positive reinforcement produced by electrical stimulation of septal area and other regions of rat brain. *Journal of comparative and physiological psychology*, *47*(6), 419.

Robinson, S. (1982). *Mindkiller: A Novel of the Near Future*. New York: Holt, Rinehart and Winston.

Strathern, M. (1997). 'Improving ratings': audit in the British University system. *European review*, *5*(3), 305-321.

Yampolskiy, R. V. (2014). Utility function security in artificially intelligent agents. *Journal of Experimental & Theoretical Artificial Intelligence*, *26*(3), 373-389.