A chapter from "Artificial Intelligence Safety and Security", ed. R. Yampolskiy, CRC Press, 2018

13 AI Safety
A First-Person Perspective

Edward Frenkel

CONTENTS

Though I am not an expert in the field of artificial intelligence (AI) safety, I'm not that far from it. As a mathematician and University of California, Berkeley professor, I have been exposed over the years to various topics in computer science pertinent to AI. Although my current research is on the Langlands program, on the interface of math and quantum physics (see References 1 and 2), when I was a college student I was enrolled in an applied mathematics program that involved quite a bit of computer science and applied math. My first research projects were in fact on the development of decision-making algorithms in medical diagnostics (see Reference 2, Chapter 12), and I have retained interest in topics such as data analysis and machine learning. I also have an interest in the ways technology is transforming our society, and what this means for humanity. AI safety surely presents some of the most important and difficult challenges in this domain. When I was invited to contribute to this volume, I accepted the invitation because of these interests, and also because I want to talk about one essential aspect of this challenge that is rarely, if ever, discussed.

I call it the "first-person perspective," as opposed to (and complementary to) the "third-person perspective" that is much more prevalent in scientific inquiries today. The third-person perspective means that one analyzes objective phenomena that are perceived as being "in the world," outside of oneself, who is viewed as an independent observer (i.e., a "third person"). The first-person perspective suggests that one also observes oneself, and in particular, one's relationships to the world and to objective phenomena. This perspective therefore necessarily includes subjective elements.

For centuries, scientists viewed the world as a collection of objects interacting with each other but independent of the observer, and for this reason any hints of the subjective were criticized and rejected. However, since the beginning of the twentieth century, with such landmark achievements as quantum mechanics, general relativity, and Gödel's incompleteness theorems, science has been teaching us that we can't really disentangle the observer and the observed. Thus, the taboo of the first-person perspective was lifted and gradually, more scientists started to pay attention to it. However, there is still a long way to go for us scientists to welcome and integrate the subjective and the first-person perspective into our science-informed worldview. Furthermore, I think we urgently need to do that today precisely because of the challenges posed by AI safety issues.

This brings to mind the words [3] of the great American philosopher, psychologist, and physician William James, whose works influenced Bertrand Russell and Niels Bohr [4], among many others:

> There is nothing in the spirit and principles of Science that need hinder Science from dealing successfully with a world in which personal forces are the starting-point of new effects. The only form of thing that we directly encounter, the only experience that we concretely have, is our own personal life. The only complete category of our thinking is the category of personality, every other category being one of the abstract elements of that. And this systematic denial on Science's part of personality as a condition of events, this rigorous belief that in its own essential and innermost nature our world is a strictly impersonal world, may, conceivably, as the whirligig of time goes round, prove to be the very defect that our descendants will be most surprised at in our own boasted Science, the omission that, to their eyes, will most tend to make it look perspectiveless and short.

How is this relevant to AI safety? The answer is that, ultimately, it is the humans who program the machines. Our past experiences, our fears, insecurities, and other "issues" and "hangups" inform and influence our beliefs and our behavior and are therefore relevant to what AI researchers and practitioners do when they create and maintain AI systems and protocols. Therefore, if an AI researcher or programmer is not fully aware of what's driving them, he or she may not be able to perform one's duties to the fullest capacity to ensure the safety of the system under one's control and the safety of the people who are affected by it.

Let me illustrate this with an analogy: the tragic story of the Germanwings flight 9525 that crashed in the Alps on March 25, 2015, killing all passengers and crew on board. Evidence showed that shortly after takeoff, the copilot locked the pilot out of the cockpit and deliberately crashed the aircraft into a mountain. Furthermore, it was later revealed that the copilot had been treated for suicidal tendencies but he hid this information from the airline. Thus, he was able to circumvent the airline safety regulations and commit suicide crashing a plane and taking 150 innocent people with him. Needless to say, all of those people had assumed prior to boarding the plane that their safety would be ensured according to the usual safety protocols. Yet, those protocols were powerless in the face of a single person with suicidal intent.

It is tempting to dismiss a similar scenario in the case of AI safety as too far-fetched. But consider the following variation: Suppose that a person tasked with ensuring safety of an AI system secretly believes that humans are terribly and irreparably flawed as a species, and their only positive role in evolution is to create and empower a superior "robot race" that would presumably then subjugate or exterminate "useless humans." Ideas of this nature are being seriously proposed and discussed in the open these days by a number of well-known and well-respected AI researchers (see, e.g., Reference 5), and a lot of people seem to be quite sympathetic to this line of thought.

Now, let's think about that. Putting aside the veracity of these ideas, it is clear that if a person harboring such ideas is involved with AI safety, he or she might be inclined to use his or her position in order to sabotage even the most stringent AI safety protocols, thereby putting other humans at risk. Moreover, this person would not have any moral reservations about doing that, for he or she would believe to be fulfilling what they perceive as their "evolutionary duty."

Clearly, no safety protocol can be made 100% proof if the people who design and implement these protocols do not act in good faith. Especially, if they are driven by their personal "metaphysical ideas" (for lack of better word) of the kind discussed above, which are in direct contradiction with the Asilomar AI Principles [6] (such as "Value Alignment" and "Human Values" on the alignment between the tasks performed by robots and the interests and values of humans). As a result, humanity may well find itself in the position of the passengers of flight 9525.

An obvious suggestion to counter this possibility would be to subject all AI safety personnel to strict psychological testing. But the fact that a pilot was able to evade such testing, as discussed above, shows that this would not be 100% proof either. Ultimately, the only guarantee of safety lies in each individual involved in the process taking his or her personal responsibility. The stakes are especially high in the area of AI, and that's why, in my opinion, in order to ensure safety, it is necessary to educate AI researchers and practitioners about the first-person perspective. At the moment, as far as I know, this perspective is not represented in the curricula of most of our computer science programs. I believe that for humanity to survive and prosper while using increasingly sophisticated technology, this has to change.

To summarize: our past experiences, our fears, insecurities, and other issues inform and influence our beliefs and our behavior. So, am I proposing that all AI researchers learn more about themselves? Yes, I do, but not in a way in which we are used to learning things. And that's because those psychological issues may well be "under the radar" of our thinking and rational reason: we may not be consciously aware of those issues due to our psyche's defense mechanism called "dissociation."

It is this defense mechanism that makes the task of finding and resolving our "issues" significantly more difficult and subtle, and creates a kind of "catch 22" situation (especially, for the "left brainiacs" like myself). Put simply, studies show that one can follow certain destructive impulses *without being* *aware of it* and hence not being able to control those destructive impulses. That’s why it is necessary for every AI researcher to learn about this defense mechanism and apply this knowledge in their personal life.

This phenomenon is actually well known in modern psychology (as anyone can see from the books in References 17–21 or simply by googling the word “dissociation”). Carl Jung, the father of modern analytic psychology and one of the deepest thinkers of the twentieth century (in this regard, I note his collaboration with the Nobel prize-winning physicist Wolfgang Pauli on the possible links between psychology and quantum mechanics [7]), has given perhaps the clearest model for understanding dissociation, as well as other phenomena in psychology.

According to it, the individual psyche has a “conscious part” and an “unconscious part.” The former is like the tip of the iceberg. That’s what the individual is aware of in ways that can be described and explained—that is to say, can be verbalized. The latter is like the vast underwater part that functions according to its own rules. While its contents are nonverbal, they influence our behavior just as much as the conscious part (if not more). Those contents, according to Jung, “wish” to be liberated, to burst out of the “underwater” unconscious into the “abovewater” conscious part, so that the individual would become fully aware of them. Jung wrote [8], “Everything in the unconscious seeks outward manifestation, and the personality too desires to evolve out of its unconscious conditions and to experience itself as a whole.”

The difficulty is that, by the very definition of the unconscious, its contents cannot be communicated directly, on the verbal plane. Therefore they can’t be “teased out” of the unconscious by logical thinking and reasoning alone. Instead of words, these contents present themselves through symbols. The symbolic information may be conveyed by certain persistent situations in the person’s life, by recurring thoughts or images, or by one’s dreams (see, e.g., Jung’s popular account in Reference 9).*

Now I am ready to state my main points:

Among our past experiences, some of the most consequential, in terms of defining our character, beliefs, and behavior, are the most difficult and painful experiences of our childhood and adolescence. At the moment that such an experience occurs, a young psyche is likely to be overwhelmed by it to the point that it simply cannot cope with the experience. As the result, the part of the person’s psyche that is fully aware of the experience, including the immediate pain and suffering (from the first-person perspective), gets “split off” and pushed to the unconscious. This is a defense mechanism well known in psychology and often referred to as “dissociation” [17–21]. The result of dissociation is that in the conscious part of the psyche the memory of this event is either entirely erased, or it is stripped of its emotional charge (pain, suffering, etc.) and is reduced to a third-person view. This can have tremendous consequences for the person’s life, relationships, and well-being.

Rather than “unpacking” these points in an abstract fashion, I switch now to the first-person perspective and tell you about one of my childhood traumas, the dissociation triggered by it, the impact it had on my life, and how I finally came to face it.

When I was 16 years old, I applied to *Mekhmat*, the Department of Mechanics and Mathematics of Moscow State University. This was the only place in Moscow that offered a program in pure mathematics. In the preceding year, under the mentorship of a family friend who was a professional mathematician, I fell in love with math. Becoming a mathematician was my dream, and I applied to *Mekhmat*.

What I didn’t know (being from a small town) was that in those days (in 1984, to be precise) the door to *Mekhmat* was closed to me and other students with Jewish-sounding last names. This had

* Roman Yampolskiy, the editor of this volume, has pointed out to me a similarity with the “problem of explainability” in deep neural networks, in which the intermediate layers of the network function largely as “black boxes.” At the moment. it’s not clear to me whether this is more than an analogy. In fact, as Jung and others have argued (and as I try to show in this article using a concrete example), the unconscious contents, even though they are initially non-verbal, refer to something definitive that can be communicated by symbols and ultimately even verbally (once these contents are brought into conscious awareness). No such mechanisms are currently known to exist in neural networks. However, a possible connection between the two phenomena definitely deserves further study.

nothing to do with religion (my family was not religious), but had everything to do with ethnicity, and blood (indeed, my father is of Jewish descent). Unbeknownst to me at the time, there was an elaborate system of identifying and ruthlessly failing "undesirable" students like myself (with a quarter or more of "Jewish blood," as I learned later) at the entrance exams to *Mekhmat*. Since officially all nationalities in the Soviet Union had to be treated equally, the examiners didn't tell us the truth: that the real reason we were failed was anti-Semitism. Instead, specially trained examiners did everything they could to prove to us that we weren't good enough. Our exams lasted much longer than those of other students, we were given the so-called "coffin problems" which looked innocent at first glance but were in fact much harder than the problems given to other students (for some examples, see Reference 10), and so on. All of this is by now well known and well documented (see, e.g., Reference 11).

As it happens, I was failed at my oral math exam, after a Kafkaesque interrogation by two examiners that lasted more than 4 hours. I described this experience in detail in a chapter of my book [3] (it was also published separately in Reference 12), so I will not get into the details here. Rather, I want to focus on the long-term effects this ordeal had on my psyche—of which I was totally oblivious for a long time.

In fact, for many years after the exam I told myself a story that "yes, this was a tough experience, but I was so strong and so sure of myself that it did nothing to sway my confidence and determination to become a mathematician." It's true that I was able to overcome the adversity. I was accepted to an applied math program at a technical school in Moscow, and in my second year there, as luck would have it, I got introduced to two great mathematicians who became my mentors. With their guidance, I wrote my first mathematical papers that were smuggled abroad and became known, bringing me an invitation to come to Harvard as a visiting professor when I was 21 (merely 5 years after that fateful Moscow University exam). That was the story of "overcoming adversity" that I told myself. But actually that was only a small part of the real story, I now realize.

I remembered my exam, of course. I remembered every detail of it, in fact, and even recounted them in my book. But I remembered all of it from the third-person perspective. It took me 30 years to learn what really happened to this 16-year-old boy that day. And when I did, it was very different … I don't know another way to say it: He died. A part of me died that day.

Try to put yourself in the shoes of a 16-year-old kid subjected to this kind of psychological torture. A kid who had done nothing wrong, singled out simply because of his last name. And just like that, the door was shut in front of him. There was no way around it, or so he thought. His dream was crushed, and he was crushed. What was the point to go on living?

When we are young, we are not equipped to handle this kind of pain. It's just too much, and we break. The only way to survive was to cut off that part of my psyche that knew what really happened. That's the genesis of dissociation. I left that boy, as if frozen in time, on that battlefield and I continued on, pretending that everything was fine.

But it wasn't fine. I survived, but at a high price: I had to sever my emotional connection to that boy, and therefore to the world. As the result, I could no longer be open, vulnerable, spontaneous, the way kids are. I felt insecure, I doubted myself, I constantly grasped for safety and control. As the result, I became increasingly closed and mistrustful of others. I dreaded uncertainty and would feel extremely uncomfortable at the slightest indication that I wasn't in control. I wanted to rely only on logic and rational reason, refusing to trust my intuition. I approached my life's situations as though they were mathematical problems, and would feel deeply disappointed and blame myself, or others, if my purported "solutions" didn't work out as planned. Naturally, this led to suffering and misery.

There were other, very tangible, effects this dissociation had on me as well. For instance, I would become very anxious every time I had to speak in public. No matter how much I prepared, anxiety would come, and there was nothing I could do to change that. The trauma also affected my relationships: an off-hand insensitive comment would trigger a harsh, disproportionate response on my part that would cause heartaches and breakups. What's even worse is that I didn't even see that these situations were directly linked to what happened at my exam.

And how could I possibly see that? I had never heard of “dissociation” or “childhood trauma” (even though I now realize this is a subject well studied in modern psychology and discussed widely in books and online; see, e.g., References 17–21). Or perhaps I did hear about it, but I dismissed it outright as “some new agey stuff” that’s “not scientific” and hence “not for me.” How convenient! Together with the faux masculinity of “real men don’t cry and never look back,” the prison I had built for myself was complete. In retrospect, I see clearly how I could actually *prefer* to stay in the dark for the same reason that the dissociation happened in the first place: I could not, or did not want to, deal with the pain that would inevitably arise if I dared to learn the truth about what really happened to me.

That’s what I meant earlier by a “catch 22” situation: our logical mind, working extra hard to shield us from pain, actually *prevents* us from knowing the truth. So, the more we rely on the mind in these matters, the farther we are from the truth. Paradoxical, isn’t it? This is especially troubling for us scientific types, because we rely on our logical thinking even more than other people (and of course, it’s not by chance). It’s as if our mind is constantly telling us: “Don’t look there, don’t look there,” as we crawl on the road of life, not being wholly ourselves and not even being aware of it.

But the pain is still here, suppressed, pushed to the unconscious. And that’s why an inner conflict arises—because the part of the unconscious that is aware of the traumatic experience wants to be liberated. As the result, we find ourselves again and again in similar situations, and when we do, this unconscious part lets itself known. In this respect, I recall the words of philosopher George Santayana: “Those who cannot remember the past are condemned to repeat it.” One could say that in those moments our rational side gets overpowered by the remembrance of the trauma, and the person starts acting seemingly irrationally. In my case, for example, it was the irrational fear of public speaking or my overreacting to every remark that I perceived as an affront.

These situations actually create an opportunity for the person to bring those memories back into his or her conscious awareness—and to finally “meet” that part of himself/herself that had been split off. But as we have discussed, such a “meeting” necessarily involves processing and accepting the pain that was refused in the first place. Therefore, there is typically a great resistance to that. Until the next episode when we get overpowered by the past trauma, that is, and so on. Thus, one gets into a vicious circle in which the unconscious keeps “acting out” without the person being consciously aware what’s really going on or being able to control it. In the process, more “defensive barriers” are erected and the psyche becomes ever more fractured, which leads to suffering and even diseases.

I was lucky: 30 years after my exam, I was finally able to break through to my 16-year old self and learn what really happened. It wasn’t easy. I was again a defenseless boy who was tortured at that exam. All the pain that I had refused to accept then came crushing at me as a giant wave. It felt like a tsunami [13]. When this happens, it feels as though the pain will never stop. It feels as if it can destroy you, and it’s quite scary. So, that’s why I was afraid to “go there”! But it’s important to understand that this pain is finite, and eventually it *will* stop. You have to let it go through you, you have to let it out, and then it will dissolve, so you can move on. It takes time and effort, but we can do it, and we will be better off for it.

Suddenly, I felt like my 16-year old self was back with me, alive. I had lost him for 30 years, but not anymore. He is here now, and I will never let anybody hurt him again. This is what it’s all about. It’s about overcoming our fear, and having the strength and the courage to bring back our little ones whom we leave, like fallen soldiers, on the battlefield of life. We have to hold their hand, take away their pain, and give them our love. To liberate them, to liberate ourselves.

And then you start seeing what’s going on. How our logical mind is constantly trying to steer us away. How we build those prisons trying to protect ourselves from the pain and discomfort of the truth. But being human doesn’t just mean happiness, joy, and laughter. It also means occasional sadness, grief, pain. There’s a whole spectrum of emotions. You can’t truly have one without the other [14]. You have to accept all of them in order to be whole, to be fully yourself.

I remember, years ago, reading about that famous maxim “Know Thyself” inscribed on the Temple of Apollo in Delphi. I wasn’t quite sure what it meant. “How could I possibly not know myself?” I thought. Well, it turned out that I didn’t know. There were parts of me of which I wasn’t even aware.

But if I don't know myself, how can I possibly make judgments about the fate of humanity? If I don't know what ails me, how can I possibly know what ails others, what ails the world? If there are such big blind spots in my field of vision, how can I possibly see reality as it is, without projecting my own insecurities and fears?

I can attest that when I did connect to my childhood experiences (and there have been other traumatic experiences, besides the exam), I realized that pretty much all my prior theories and beliefs about humanity, AI, evolution, consciousness, etc., were hopelessly naïve, narrow minded, and vapid [15,16]. And it's clear why: they were being spun by my frightened ego as a way to escape from learning who I am.

Pascal said, "The heart has its reasons of which the reason knows nothing." We have to listen to our heart every once in a while, even if our mind tells us otherwise. And the good news is that if we are willing to do it, then we *can* overcome our fears and connect to our little ones inside. We can learn who we are, we can be whole again. Those severed emotional connections can be restored, so we can see the world again as it is, without filters. We can then connect to each other anew and create a better world together.

I believe that we all have the responsibility to get to know the little ones within ourselves. But this especially applies to AI researchers and practitioners because there is so much responsibility vested in this field today. This is not about any religion or mysticism. It's about finding out who you are.

If you would like to learn more, I recommend the books in References 17–21. This is a good starting point. However, because we are talking about subtle and deeply personal phenomena, one shouldn't expect a "one size fits all" solution (no wonder Jung compared the process of self-realization to alchemy). What is important to realize, in my opinion, is that all of us have experienced difficult situations in our childhood. No one grows up in isolation, in the proverbial "greenhouse." Life intrudes on us, and we get hurt, especially when we are most innocent and vulnerable. There is nothing shameful in it. This is a normal, and perhaps even necessary, part of human development. The question is: What are we going to do about it when we grow up and become stronger? Will we have the courage to revisit those moments and heal the trauma?

Each of us has his or her own path, and this has to be respected. But we definitely need improvements in our education system, so that everyone has an opportunity to learn about these issues early on. In particular, I think that psychology courses in which trauma and dissociation are discussed should be in the core curricula of our computer science programs, taught by well-trained psychologists. Consultations with experienced therapists should be made available to students in those programs (on a voluntary basis). But equally important, in my opinion, is that we, scientists, and especially AI researchers and practitioners, take matters into our own hands and bring forward the first-person perspective in our conferences, discussions, books and articles we write, and so on. We should create a safe space in which we are not afraid to share our personal stories. When we do that, we will realize that actually our stories have a lot in common. When one of us speaks, so will others. And when we share our stories, we will help each other to heal and to become more aware of who we are.

REFERENCES

1. More details can be found on the math page of my website: http://edwardfrenkel.com/mathematics
2. Edward Frenkel, *Love and Math*, Basic Books, New York, 2013.
3. William James, Address of the President before the Society for Psychical Research, *Science*, vol. 3, No. 77, (1896), pp. 881–888.
4. Jan Faye, *Niels Bohr: His Heritage and Legacy: An Anti-Realist View of Quantum Mechanics*, Springer, Heidelberg, 1991 (Section II.2).
5. Ben Goertzel, *Kafka in the morning*, "Multiverse according to Ben" blog entry, September 12, 2016. Retrieved from http://multiverseaccordingtoben.blogspot.com/2016/09/kafka-in-morning.html
6. Asilomar AI Principles. Retrived from https://futureoflife.org/ai-principles/

7. Carl G. Jung and Wolfgang Pauli, *Atom and Archetype: The Pauli/Jung Letters, 1932–1958*, Princeton University Press, Princeton, 2014.
8. Carl G. Jung, *Memories, Dreams, Reflections*, Vintage Books, New York, 1989.
9. Carl G. Jung, *Man and His Symbols*, Dell Publishing, New York, 1968.
10. Tanya Khovanova and Alexey Radul, Killer Problems, *American Mathematical Monthly*, Vol. 119, No. 10, 2012, pp. 815–882. Electronic version retrieved from http://arxiv.org/abs/1110.1556
11. M. Shifman (ed.), *You Failed Your Math Test, Comrade Einstein*, World Scientific Publishing, Singapore, 2005. Electronic version retrieved from http://www.ftpi.umn.edu/shifman/ComradeEinstein.pdf
12. Edward Frenkel, *The Fifth Problem*, New Criterion, October 2012. Electronic version retrieved from http://www.newcriterion.com/issues/2012/10/the-fifth-problem-math-anti-semitism-in-the-soviet-union
13. See the video of my talk at the 2014 Science and Nonduality conference when this was still fresh in my mind: https://youtu.be/YnqQ-BWMHrE
14. An elegant and moving artistic rendition based on recent studies is presented in the film “Inside Out,” 2015.
15. I spoke about this in more detail in the following lectures available online (see also Reference 13): “I am or AI am,” keynote address at the Aspen Ideas Festival, July 2015. Retrieved from https://youtu.be/lbLI9aX5eVg
16. “Mathematics and love in the age of AI,” 60th anniversary lecture at METU, Ankara, November 2015. Retrieved from https://youtu.be/TXQP0mPFskw
17. John Bradshaw, *Homecoming: Reclaiming and Healing Your Inner Child*, Bantam Publishing, New York, 1992 (Reprint Edition: 2013). Electronic version available from https://www.amazon.com/Homecoming-Reclaiming-Healing-Inner-Child-ebook/dp/B00C4BA4I6/
18. Janina Fisher, *Healing the Fragmented Selves of Trauma Survivors: Overcoming Internal Self-Alienation*, Routledge, Abingdon, 2017. Electronic version available from https://www.amazon.com/Healing-Fragmented-Selves-Trauma-Survivors-ebook/dp/B06X9YWZMM/
19. Bessel van der Kolk, *The Body Keeps the Score: Brain, Mind, and Body in the Healing of Trauma*, Penguin Books, Westminster, 2017. Electronic version available from https://www.amazon.com/Body-Keeps-Score-Healing-Trauma-ebook/dp/B00G3L1C2K/
20. Peter A. Levine, *Waking the Tiger: Healing Trauma*, North Atlantic Books, Berkeley, 1997. Electronic version available from https://www.amazon.com/Waking-Tiger-Healing-Peter-Levine-ebook/dp/B002IYE5XO/
21. Jeremy Abrams (ed.), *Reclaiming the Inner Child*, Tarcher, Los Angeles, 1990.