A science of suffering needs an understanding of the Self through embodied artificial general intelligence

"'This is myself and this is another.'
Be free of this bond which encompasses you about,
And your own self is thereby released."

(The Buddhist Scriptures, p. 179. 1959.)

Without question, depression and other psychological suffering are some of the largest challenges of our time. According to a 2017 report from the World Health Organization [1], depression is the leading cause of disability around the world, affecting 4.4% of the world's population. The report also highlights that 3.6% of the global population is affected by anxiety disorders. Not only are these conditions highly prevalent, they have also increased substantially during the last ten years.

Depression and anxiety disorders are highly multi-dimensional in nature, being a result of complex interactions of biology and life experiences. As a clinical psychologist and researcher in clinical psychology, I have struggled to understand these conditions from multiple perspectives. My own conclusion is that it all boils down to concept of the Self. A cardinal symptom of depression is self-hatred — talking to oneself in a way that would be impossible to do to someone else, labeling ourselves using concepts that are associated with the most severe forms of contempt, such as "I am a horrible person". We constantly relate ourselves to others in all imaginable ways, such as "I am much less valuable than others" or "I am not worthy of love". A very common theme among people struggling with anxiety is fear related to the perception of being judged by others, such as "Other people think I'm pathetic" or "I need to perform for other people to like me". At the root of this is the Self. We identify with the labels we put up to describe ourselves, and we derive an endless amount of self-labels in the endless relating

towards others. This process of forming a conceptual self, which I will call *selfing*, is what I see as the very root of psychological suffering.

This idea is not new. As illustrated by the quote above, a fundamental thesis in Buddhism is that there is no such thing as an unchanging, permanent Self in living beings, and, the clinging to a belief in a Self is the source of all suffering. Hence, it has been a claim for more than 2500 years, since the time of the Buddha, that selfing is what causes suffering. Importantly, *selfing is something we do*, a constant ongoing relating of the conceptual self to other concepts. If selfing is a behavior, then it can also be changed.

During the last 30 years, behavioral scientists have made a substantial amount of work towards an experimental understanding of selfing as relating. The contemporary behavioral science theory *Relational Frame Theory* [2] argues that language and cognition, including selfing, is possible to study as an act of complex relating. While this science has made an impressive amount of progress there is simply a limit to what can be done when doing experiments with human beings. Even if behavioral scientists had an unlimited amount of time and resources, which they clearly don't, there will always be research that cannot be conducted. Some complex forms of selfing and relating will be impossible to arrange in an experiment. Other forms of experiments will be deemed unethical to conduct. But above all, as the methods of the science of relating are working right now, the science cannot scale to levels of progress that the world needs right now, to solve some of its most urgent issues.

This is where artificial intelligence comes into the equation. During the last decade, we have seen an impressive increase in applications of AI, very often in the medical domain. Can AI help behavioral science to scale? Yes and no. While today's technology has a lot to offer regarding simulating the mind, there are several challenges that needs to be addressed. One problem with traditional AI is the challenge of transferring what has been learned in one domain into another [3]. This is of course highly relevant for selfing and relating, as the whole point with Relational Frame Theory is to understand the very fact of how we relate ourselves to things never experienced, like "I want to be like that person, because then I would be much more happy". One way to put it is that a lot of clinical problems and negative thinking about the self can be understood as involuntary and constantly ongoing transfer of

knowledge between domains. The problem of transfer is only one of the problems with AI as we normally see it [4].

The field of Artificial General Intelligence (AGI) is both an old and a new field of research [5]. It was the original goal of AI, to create "thinking machines" in the sense of general-purpose intelligent systems that could carry out a range of tasks across different domains. However, most AI as we see it today is solutions to highly specific problems in narrow domains, such as image recognition and game-playing, where the algorithms used typically vary from task to task. While this form of problem-solving is very effective and undoubtedly something that will help human society to evolve, it is far from AGI. Importantly though, during the last 10 years, there has been substantial progress in the field of AGI itself. One of the most important attempts of progress in AGI as a science is to establish a theory of general intelligence in itself. With a level of description that is abstract enough, intelligence should be possible to describe as a phenomena that is possible to study in both humans and machines. A theory of general intelligence should be able to account for what we have learned from human psychology, but be formal enough to be possible to implement in a computer system [6]. Importantly though, AI in the form of problem-solving in specific domains does not provide a theory of intelligence that can be used to guide research and practice in the field of AGI.

While many interesting approaches exist in the AGI field, one particular model that I see as promising is Pei Wang's Non-axiomatic Reasoning System (NARS) [7]. It implements a "non-axiomatic" logic that is a formalization of Pei Wang's theory of general intelligence. In essence, NARS takes a logical approach to intelligence, using formal inference rules to derive new knowledge from existing knowledge and experience. NARS uses a formal language of relations to describe all of this. The meaning of a concept for a NARS system is how that particular system has experienced a concept in relation to other concepts. Hence, all "thinking" is NARS can be seen as acts of relating. While NARS is very much in development, the design of it opens up for a theoretically coherent understanding of thinking, feeling, and consciousness in an AGI system. In addition, it also provides an account of the Self as a concept, being part of an endless ongoing act of relating. In summary, to be able to study selfing as an ongoing process in machines, we will highly likely need AGI. Among the currently available AGI models, Pei

Wang's NARS seems like a very interesting candidate, with its strong theoretical foundation in how to view intelligence as an experience-grounded inference process.

Another key aspect of contemporary artificial intelligence is the role of the body [8]. Is it possible to create human-level AI without the embodiment in the world? Most scientists today, including myself, would say No to that question. A longer answer is that it depends on what kind of AI you're trying to build. The approach NARS takes towards AGI, that intelligence (including selfing) is an ongoing act in context, where meaning is grounded in experience, assumes the existence of a body. Importantly though, the body doesn't need to be a human body. Sensing enabled by the body become part of a system's experience, independently of which "sense organs" the body provides.

In my own research, I have spent the last two years working with a highly sophisticated robot body, the Furhat robot, created by Furhat Robotics in Stockholm. My goal is to create a simulation of psychological suffering such as depression and anxiety disorders, taking place in a robot body. Using the Furhat software, we have created prototypes of patients expressing their own suffering. In parallel, we have been experimenting with NARS on its capacity for showing the processes Relational Frame Theory suggests needs to take place for selfing to happen. Very much work still needs to be done, but the results this far are encouraging. I hope to be able to gradually incorporate NARS into the Furhat robot, creating a simulation of suffering that is not only expressed clearly by the robot, but also implemented similar in process to that in human beings.

To summarize, depression, anxiety and other psychological conditions are some of the world's largest challenges. Contemporary behavioral science, as well as the 2500-year-old doctrines of Buddhism, suggest that the Self is at the root of human suffering. To accelerate progress in developing the most effective means for dealing with suffering, we need a science of selfing as a process working together with AGI models, grounded in bodies and world. Another way to put it, to understand ourselves and our own suffering to its full extent, we need to develop robots with a Self.

References

[1] World Health Organization. Depression and other common mental disorders: global health estimates. World Health Organization; 2017.

[2] contextualscience.org/what_is_rft

[3] en.wikipedia.org/wiki/Transfer_learning

[4] medium.com/@GaryMarcus/in-defense-of-skepticism-about-deep-learning-6e8bfd5ae0f1

[5] en.wikipedia.org/wiki/Artificial_general_intelligence

[6] cis.temple.edu/~pwang/Publication/TheoriesOfAI.pdf

[7] www.frontiersin.org/articles/10.3389/frobt.2018.00020/full

[8] medium.com/@ashishkumar191295/embodied-cognition-the-road-towards-true-artificial-intelligence-d0306c514670