Humanities & Social Sciences Communications

ARTICLE

https://doi.org/10.1057/s41599-021-00812-y OPEN

Freud and the algorithm: neuropsychoanalysis as a framework to understand artificial general intelligence

Luca M. Possati[1]✉

The core hypothesis of this paper is that neuropsychoanalysis provides a new paradigm for artificial general intelligence (AGI). The AGI agenda could be greatly advanced if it were grounded in affective neuroscience and neuropsychoanalysis rather than cognitive science. Research in AGI has so far remained too cortical-centric; that is, it has privileged the activities of the cerebral cortex, the outermost part of our brain, and the main cognitive functions. Neuropsychoanalysis and affective neuroscience, on the other hand, affirm the centrality of emotions and affects—i.e., the subcortical area that represents the deepest and most ancient part of the brain in psychic life. The aim of this paper is to define some general design principles of an AGI system based on the brain/mind relationship model formulated in the works of Mark Solms and Jaak Panksepp. In particular, the paper analyzes Panksepp's seven effective systems and how they can be embedded into an AGI system through Judea Pearl's causal analysis. In the conclusions, the author explains why building a sub-cortical AGI is the best way to solve the problem of AI control. This paper is intended to be an original contribution to the discussion on AGI by elaborating positive arguments in favor of it.

[1] University of Porto, Porto, Portugal. ✉email: lupossati@gmail.com

Introduction

The thesis[1] of this paper is that neuropsychoanalysis and affective neuroscience can provide a new paradigm for AI, particularly for artificial general intelligence (AGI). I especially refer to the works of Mark Solms and Jaak Panksepp. Neuropsychoanalysis and affective neuroscience give us a precise answer to the enigma of the mind/brain dualism by highlighting the constant interaction of these two dimensions. I will try to show how this approach can give us a new key for the conceptualization of AGI. This is a preliminary communication in which I treat the problem in a broad outline and the first step in a research project on the possibility of AGI. Obviously, the topics covered are all controversial and would ideally be given a detailed analysis in several papers.

The structure of the paper is as follows. In section "Neuropsychoanalysis: an introduction", I give a definition of the fundamental aspects of neuropsychoanalysis and its relationships with affective neuroscience. In Sections "The neuropsychoanalytic model of the mind" and "The primitive affective states, or the basic human values", I illustrate the neuropsychoanalytic model of the mind/brain relation. In particular, I analyze Panksepp's theory of the basic affective states. In section "A Freudian computer: sketches", I define some basic design principles of an AGI model based on neuropsychoanalysis and affective neuroscience. I answer the question: How can we translate Panksepp's theory of the basic affective states into algorithms?

Why do we need a new approach to AGI?

I answer this question with two remarks. Firstly, an essential point that artificial intelligence (AI) research must consider is that a method based merely on the physical imitation of the brain is wrong. This is for two reasons: the first is that our knowledge of the brain is still very limited, and the second is that even assuming that we could properly reconstruct each cell of our brain and its functioning, something would still be missing, namely the mind. We, therefore, need a model that can hold these two dimensions together, mind and brain. The imitation of anatomical mechanisms and the psychological expression of these mechanisms must go hand in hand.

Secondly, so far, research in AGI has mainly focused on the activities of the cerebral cortex and the main cognitive functions (language, logic, memory, cognition, etc.). The time has come to think and develop an AGI of the subcortical. Neuropsychoanalysis and affective neuroscience affirm the centrality of emotions and affect—i.e., the subcortical area of the brain where primary processes (instincts, emotions, feelings) are located in psychic and cognitive activity. My hypothesis is that an AGI system inspired by neuropsychoanalysis and affective neuroscience must be based on the modeling and simulating of the seven basic affective states analyzed by Panksepp. Panksepp's works—crucial also for neuropsychoanalysis—give us a theoretical framework on which to develop this hypothesis.

An important clarification should be made. Today, there is much discussion of affective computing. The relationship between emotion and AI is a vast research field, beginning with the important and controversial book by Rosalind Picard (1997). What does it mean for a computer to "have emotions"? In general, when we talk about effective computing, we mean three connected things: (a) the way in which a computational system can recognize and instantiate human emotions; (b) the way in which a computational system can respond to human emotions; and (c) the way in which a computational system can express emotions spontaneously or use them in a positive way in the decision-making process (see Schuller and Schuller, 2018; Erol et al., 2019; Shibata et al., 1997; El Nasr et al., 2000; Fogel et al., 2018). "More specifically, affective computing involves the recognition, interpretation, replication, and potentially the manipulation of human emotions by computers and social robots" (Yonck, 2017, p. 5). Experts agree that artificial emotional intelligence is a continuously developing research field and that it will have a decisive importance in the future economy and society. However, artificial emotional intelligence will also pose new ethical and legal problems. There are new dangers, such as psychological manipulation (see Picard, 1997, chapter 4). The study of biomimetics and hybrid systems (biological and technological) that analyze the possibility of building robots capable of reproducing the versatility of the human organism (see Prescott et al., 2018) is also connected to this immense research field.

How is my research different from the affective computing approach? This paper does not intend to provide an overview of the debate on affective computing. The scope of that subject would merit an entire book unto itself. However, I will develop some critical considerations of Picard's concept of emotions and feeling. In my opinion, Picard remains too tied to a cognitivist conception of mind, preventing her from considering emotion as such. Following Panksepp (1998) and Panksepp and Biven (2012), I hold that emotion is an intrinsic function of the brain, not the reflection or derivative of the higher cognitive functions. There exist basic instinctual systems that are phylogenetic memories that we have inherited as evolutionary tools for living. If we do not fully understand these systems, we cannot understand the brain/mind relationship, or the "BrainMind," as Panksepp terms it. Human emotionality has an intelligence, a structure; it is not only the mechanical answer to a series of random situations. A subcortical AGI should be capable not only of reproducing the basic human affective systems but also of using them to build the most elaborate cortical functions, such as learning and language. Therefore, three aspects characterize my approach: a) emotions are not reducible to cognitive activities; b) cognitive activities arise from emotions; c) emotions are analyzed from a neuropsychoanalytic point of view.

From this point of view, an AGI system that is able to instantiate these basic affective systems or even the Freudian unconscious must be thought of in a way radically different from classical methods.

What is AGI?

The dream of creating machines perfectly capable of reproducing human intelligence is very old. The investigations of Turing, von Newmann, Shannon, and many others have radically revolutionized this idea, opening an entirely new field of research (Dyson, 2012). Today, AI is an ever-expanding sector in which philosophy, technology, design, storytelling, sci-fi dystopian stories, and speculations of all kinds are continuously intertwined (see Amoore, 2009; Apaydin, 2016; Baldwin, 2016; Le Cun, 2019; Colvin, 2015). A classic definition is that of one of the great pioneers of AI, Marvin Minsky: "AI is the science of making machines do things that would require intelligence if done by men" (Bolter, 1986, p. 193). As Fjelland (2020, p. 1) underscores, this is what we call "weak AI," that is, a type of AI capable of performing only some specific human tasks (seeing, manipulating objects, classifying, etc.).

The concept of AGI is essentially different from that of weak AI because it denotes a type of AI capable of fully simulating human intelligence, not just a part of it. It is the intelligence of a machine that is capable of learning and understanding any human intellectual activity, "a machine with general-purpose, adaptive intelligence" (Shanahan, 2015, p. 3). It is therefore a generalist intelligence, capable of adapting and creating new forms of behavior with a degree of ability similar to that of a human being. However, research on AGI still appears to be very

limited. As Wikipedia says, "MIT presented a course in AGI in 2018, organized by Lex Fridman and featuring a number of guest lecturers. However, as yet, most AI researchers have devoted little attention to AGI, with some claiming that intelligence is too complex to be completely replicated in the near term."

My thesis in this paper is that an AGI is possible but must be conceived starting from the basic affective states. I will use the expression "AGI" as the equivalent of "strong AI," even if some differences can be found between the two. The expression "strong AI" could, in fact, also be interpreted as "superintelligence"—that is, intelligence that wants to overcome humans and control them (Bostrom, 2016).

There are many difficulties related to the concept of human-like AI (see Russell, Norvig, 2016). First of all, there is the obvious difference between algorithms and the human way of thinking: Penrose (1989, 1994) and Dreyfus (1972) have demonstrated, in different ways, the abyss between the human way of reasoning and computation. Dreyfus, in particular, holds that computers, who have nobody, no childhood, and no cultural practice, could not acquire intelligence at all (Dreyfus and Dreyfus, 1986; Fjelland, 2020).

The thesis of the present paper is twofold: (a) our AGIs do not have a childhood, or practical culture because they do not have a crucial element of human evolution—i.e., emotions and affects—and (b) the large masses of data (the so-called "Big Data") and the statistical techniques now available allow us to instantiate human emotions and affects. From a neuropsychoanalytical point of view, affects are the basis of intelligence and consciousness. Furthermore, this paper wants to show that claiming that AGI is possible does not at all mean overestimating technology and underestimating human intelligence.

Neuropsychoanalysis: an introduction

The main advocates of the neuropsychoanalytic point of view (Solms, Kaplan-Solms, Turnbull) argue that their perspective on the mind/brain question is the same as Freud's and call this approach "dual-aspect monism." Their thesis is that the mind is a unique reality. Nonetheless, we cannot directly access it. To describe and understand the mind, we must draw inferences (to build models) based on two limited forms of experience: first-person subjective experience (psychology) and third-person study of brain structures and functions (neuroscience). In more formal terms, the first form is "interoceptive," while the second is "exteroceptive."

These two forms of experience are independent "observational perspectives" and have the same value, but are not able to explain this unique reality, which can be called the mind/brain, in a complete way. If we look at it with our physical eyes, we see a brain, biological organ-like many others. If we look at it with the eyes of our subjective consciousness, we come into contact with mental states such as sadness, desire, and pleasure. It is therefore necessary to keep both points of view (subjective and objective) open and build dynamic parallelism between them. We will never find a thought, a memory, or an emotion in a piece of brain tissue; we will find brain cells, nothing else. Meanings and intentionality are not reducible to neurons. According to dual-aspect monism, the mind can be distinguished from the brain only from the perceptual perspective. If we admit a single entity X "behind" the terms "mind" and "brain," then we can say that (a) the mind is X perceived subjectively—that is, through one's own consciousness—and (b) the brain is X perceived objectively—that is, through external perception and objectifying methods of sciences.

Neuropsychoanalysis tries to connect the X-object to the X-subject. In this way, neuropsychoanalysis does not intend to reduce the mind to the brain; even if it has been accused of biologism, it does not intend to reduce everything to biochemical processes and anatomy. All mental phenomena require a biological correlate; this is indisputable. This does not mean, however, entirely reducing the mental phenomena and their meaning to supposed biological correlates. Biological and psychological dimensions must be kept together; they must be considered two sources of information of the same value.

Neuropsychoanalysis does not intend to prove that Freud was always right. Instead, it claims to finish the work started by Freud. Indeed, Freud began his career as a neuroscientist and neurologist (see Sulloway, 1979, chapter 1). He had a specific and broad scientific program, but it was largely conditioned by the limits of the neuroscientific methods available at the time. For Freud, psychoanalysis is not only a hermeneutics of mental life. The separation between psychoanalysis and neuroscience was for him only a pragmatic, strategic, and temporary solution; it was motivated by the lack of knowledge about the brain at the time. However, as Freud repeats in several passages, the inevitable progress of neuroscience would sooner or later lead to a bridging of the gap between the two disciplines and to an organic basis for the discoveries of psychoanalysis (Solms and Turnbull, 2002). In other words, Freud was dissatisfied with the clinical-anatomical method of his time and therefore developed his analytical method independently of neuroscience from 1895 to 1939. He eagerly awaited the progress of neuroscience and biology, and for this reason, he sought confrontation, dialog, and cooperation with these sciences (Solms and Saling, 1990).

Since Freud's time, things have changed a great deal. We can now verify the validity of Freud's basic statements through appropriate scientific observations. The knowledge and methods for studying the brain are much more developed and therefore allow us to improve and finish Freud's endeavor. In the past twenty years, neuroscience has not only experienced exponential growth but also changed its character, thanks to technological advances. In particular, the critique of the behavioristic (focused only on the observable patterns) and cognitive (the thesis that the human mind is essentially information processing, and so perception and learning) models of mind has led to a broader vision that includes emotions and feelings, the connection to a body that acts and perceives within a social and technological environment. Both the behavioristic and cognitive models undermine the importance of emotions and feelings.

This turning point can be found in numerous works: Benedetti (2010), Damasio (1994), Decety and Ickes (2009), Gallese (2009), LeDoux (1996), and Panksepp (1998). Furthermore, Luria's (1976) important work also demonstrated the possibility of renewing the psychoanalytic method through neuroscience. In particular, Solms (2000) and Kaplan and Solms (2000) underlined the importance of Luria's method, which entails the abandonment of a rigid localization of cognitive functions in favor of a much more integrated approach to the mind. This is the so-called "dynamic localization method" according to which complex mental activities (memory, imagination, thought, etc.) cannot each be located rigidly in a single area of the brain. On the contrary, many areas of the brain activate at once, each time in a different way.

It should never be forgotten, however, that the debate on neuropsychoanalysis is broad and complex. Much research in neuroscience claims that the Freudian dream theory (but not only this) is completely wrong (Hobson, 2007). There are also many psychoanalysts (see Blass and Carmeli, 2007, Edelson, 1986, Pulver, 2003) according to whom neuroscience is irrelevant to psychoanalysis, and this is because neuroscience has nothing to say about our mental meanings and their interpretation, which are the domains of psychoanalysis. Knowing the biological basis

of mental processes explains nothing of the meanings that make up our lives; it would be like wanting to explain software based on knowledge of hardware. I will not analyze these criticisms in the present paper.

The neuropsychoanalytic model of the mind

Neuropsychoanalysis proposes a general model of how the human mental apparatus, as conceived by psychoanalysis, can be represented in brain tissues. It is a hypothetical model based on current knowledge of the brain and on a still limited amount of empirical data. The theoretical points of reference for this operation are mainly Luria's work and Freudian metapsychology. The main thesis is that mental functions are not rigidly localized in individual areas of the brain but are statistically distributed in several areas. Each area contributes in its own way. Information processing is a dynamic process that involves many areas of the brain in ever-changing ways.

At the center of the model is the Pcpt-cs system, namely, the perceptual consciousness. This system has two surfaces: an external one, directed toward the world around the brain (it is divided into different areas of specialization: sight, hearing, kinesthesia, and tactile sensation) and an internal one, directed toward the processes taking place inside of the body. The first surface is located in the posterior part of the cortex (although numerous subcortical structures contribute to the processing of the stimulus in a dynamic way). The second surface is connected to the limbic system and to a series of deeper, subcortical brain structures that represent the oldest part of the brain. These are the only two sources of stimuli, or data, that our brain possesses, namely, external reality and internal reality. For neuropsychoanalysis, therefore, consciousness is nothing abstract or metaphysical. It is the set of our external and internal perceptions—the connection between the data we have about the external world and the way in which these data modify us.

According to Solms (2008), the intermediate zone between the internal and the external perception corresponds to areas of the brain that filter, record, and structure information by using connections, associations, and classifications. These areas are located in the posterior parts of the cortex. Associations and connections can be of different types, depending on the type of memory involved in the process, such as working memory, episodic memory, procedural memory, and semantic memory. The information is recorded in different ways. Through memory, the brain develops intentionality, the ability to plan its actions and therefore to act in the world. In addition, it develops the ability of thought—that is, deferring the action or satisfaction of the need at a given moment.

Now, Solms links this area to the Freudian notion of ego. In Freudian metapsychology, in fact, the ego is that instance that must mediate between external and internal reality; it is precisely that part of the id that has been modified by external reality (natural and social) in the course of evolution. Freud also saw in the ego a series of memory structures through which experiences are connected and recorded. These connections are not predetermined; they develop over time. The progressive stabilization of the connections gives rise to the main cognitive functions, such as thought, language, logic, attention, calculation, and imagination. Two important articles by Kandel (1979, 1983) explain the way in which these processes develop at the cellular level. The ego is a continuous, dynamic connective process whose constant evolution depends on numerous variables. The crucial function of the ego is to work as a barrier to the stimuli. If there were no ego to filter and organize information, the human brain would be overwhelmed by stimuli and therefore would be in a state of perennial excitement.

The id is our deep, visceral biological dimension, which is also called "internal milieu" (*milieu intérieur*) and includes different systems of our body, such as the musculoskeletal system, the immune system, the endocrine system, the chemical processes, and organic cycles. The way the brain perceives changes occurring in this biological system is what neuropsychoanalysis calls "internal perception"—this is the immense field of instincts, feelings, and emotions to which psychoanalysis gives a predominant role in the psychic activity. Internal perception consists of the activation of deep and ancient brain structures (the limbic system and subcortical brain structures) connected to the biological dimension of the body and the mechanisms of adaptation. It is important to note that the neurons that make up the limbic system and the subcortical brain structures work very differently than the neurons of the perceptive-mnestic systems of the cortex (Solms, 1996). These neurons generate not only discrete stimuli but also gradual state changes.

The ego also mediates between the id and the super-ego. According to Solms (1996), the super-ego can be connected to some regions of the prefrontal lobe and precisely to those regions that connect the prefrontal part of the brain with the limbic system. These regions act as a filter, as censorship toward the needs of the instinctual pole of the mind. Their type of memory is called "semantic memory" and mainly concerns social conventions. In line with what Freud says, the super-ego arises from the internalization of behavior and value schemes in the social context.

Before this section is concluded, one puzzle must be solved. The source of activation of internal perception is the id, the vital biological systems that compose our organism. Does this mean that the id is conscious, that we have the perception of the id? Is the unconscious conscious? Solms claims that the id is the source of all forms of consciousness: "This constant 'presence' of feeling is the background *subject* of all cognition, without which consciousness of perception and cognition *could not exist*" (Solms, 2013, p. 16). Solms and Friston (2018) stress this point: consciousness is mostly interoceptive: "The primary function of consciousness is not to register states of the external world but rather to register the internal states of the experiencing subject. […] conscious qualia arise primarily not from exteroceptive perception (i.e., vision, hearing, somatic sensation, taste, and smell), and still less from reflective awareness of such representations, but rather from the endogenous *arousal* processes that activate them" (2–3). This means that external perceptions become conscious and subjective only when they are connected to—and activated by—deeper and internal arousal processes—that is, the affect and instinct systems.

Therefore, affects and instincts compose the first form of consciousness, which is the condition of all the others. Where, then, does repression arise? In the transition from one system to another. Where does what we call properly unconscious originate, in a Freudian sense? The basic form of affective consciousness is not fully translated into the more complex systems of the ego and the superego and therefore remains invisible. "If we retain Freud's view that repression concerns representational processes, it seems reasonable to suggest that repression must involve withdrawal of *declarative* consciousness" (Solms, 2013, p. 17). In a nutshell, the id has no access to declarative consciousness.

Now, if consciousness is essentially founded on affects, what are affects properly?

The primitive affective states, or the basic human values

The organism of mammals is generally composed of a series of structures in relation to each other. Homeostasis is the set of coordinated and partly automatic physiological, biological, and

chemical processes that are indispensable for maintaining the state of the organism stable and thus guaranteeing survival, including the regulation of temperature and heart rate, the concentration of oxygen in the blood, the structure of the musculature, the skin tone, and the metabolism. According to Damasio (1999), emotions are closely connected to homeostasis; they are biological phenomena produced by neuronal configurations in order also to guarantee homeostasis. The brain influences and modifies the body by regulating it. The aim is adaptation and survival—that is, to create advantageous conditions for the organism in certain situations. For example, fear causes the acceleration of heart rate in dangerous conditions. An external situation (the danger) activates some regions of the brain that produce, through the release of chemicals or neurotransmitters, a series of modifications of the body (the acceleration of the heartbeat, the movement of the legs, etc.). In response to the brain, the body changes its internal regulation mechanisms and adapts to the new situation, and survives. Damasio distinguishes primary emotions (joy, sadness, fear, anger, surprise, and disgust) from secondary emotions, which are more complex. Then, there are the background feelings, such as well-being, malaise, calm, and tension, expressed in the details of posture and in the way of moving the body. With the somatic marker hypothesis, Damasio has shown that emotions play a role of primary importance in cognitive processes (Damasio, 1994, pp. 45–49). Furthermore, consciousness itself is closely connected to emotion, feeling, and homeostasis; it is a more refined and effective form of realizing homeostasis in the face of the challenges posed by the surrounding environment (see Damasio, 1999, Chapter 10; see also Damasio, 2003 and Damasio, 2010).

Based on the study of animals and the comparison between animals and humans, Panksepp (1998) offers us a much more elaborate and complete theory of emotions than Damasio. According to Panksepp, Damasio is still a victim of cognitivist prejudice because he still thinks that emotions are a variant of higher cognitive processes—i.e., the results of a sort of "re-reading" of them by the cortex. Cognitive prejudice can also be found in Rolls (1999, 2005): there are no basic affective states; emotions are the products of the cognitive activity—for example, the ability to verbalize or conceptualize assessments is considered a necessary condition for emotional experience. For Panksepp, these theories are full of problems and contradictions: How can a cognitive state give rise to an affective experience?

In contrast, Panksepp, who moves closer to neuropsychoanalysis than Damasio, has identified the existence of an ancestral core of emotional states that underlie any form of psychic activity, unconscious or conscious. Panksepp argues that emotions are intrinsic functions of the sub-cortical brain that humans have in common with animals. Emotions, or affects, are "ancient brain processes for encoding value—heuristics of the brain for making snap judgments as to what will enhance or detract from survival" (Panksepp and Biven, 2012, pp. 31–32). These basic affective systems are not cognitive at all; they "are made up of neuroanatomies and neurochemistries that are remarkably similar across all mammalian species" (Panksepp and Biven, 2012, p. 4).

In general, Panksepp distinguishes three levels of brain activity:

a. The primary process, which includes the most basic affects;
b. The secondary process, such as learning and behavioral and evolutionary habits;
c. The tertiary process, which includes executive cognitive functions (thoughts and planning).

The primary process activities are organized into three areas: emotional affects, homeostatic affects, and sensory affects. The homeostatic effects concern internal biological cycles (the need to defecate or eat, for example) that allow homeostasis. Sensory affects are reactions to sensations experienced from the outside; they are exteroceptive, sensory-triggered pleasurable, and unpleasurable/disgusting feelings. Emotional affects (also called also "emotion action systems," or "intentions-in-actions") are the oldest and most complex. Panksepp organizes these affects into seven systems: SEEKING, RAGE, FEAR, LUST, CARE, PANIC, and PLAY (he uses capitalization to distinguish these primary emotional brain systems from the use of the same terms in common language). These systems are described by Panksepp as real physical circuits present in the most ancient and deep parts of the brain, the subcortical area, which activates certain reactions and behaviors (for example, the rat escapes the smell of predators, and this pushes it to look for another ground to feed) and therefore forms of learning. They are instinctive (automatic reactions) and evolutionary (the result of a long natural selection process). They are networks of causal processes, as I will show later.

Panksepp argues that raw affects are the fundamental basis of any brain activity; the mind is essentially emotional, and raw affects tend to shape any other cognitive activity. "Most prominently, it looks like the SEEKING urge may be recruited by the other emotional systems. It is required for everything the animal does; one could conceptualize it in psychoanalytic terms as the main source of libidinal energy" (Panksepp, 2008, p. 165). For Panksepp, the study of the constitution of these systems is essential "for understanding our own affects and for developing better psychiatric treatments for emotional imbalances" but it "would require further causal preclinical research into our ancestral subcortical primary process emotional brain systems" (Davis and Montag, 2019, p. 2). In other words, raw affects are ancient brain processes for coding values, which are heuristic operations of the brain used to make rapid assessments of what, in the real situation, increases or decreases the chances of survival. They can interact with and be influenced by cognitive states, often in very complex ways, but they do not presuppose them. They are, using a Panksepp expression, "a flexible guide for living" (Panksepp and Biven, 2012, p. 43).

As I have just said, the crucial point of Panksepp's approach is that basic emotions have nothing cognitive and therefore cannot be understood from a cognitive point of view. They must be dealt with on their own terms. The Panksepian affective neuroscience principle is that "the neocortex is fundamentally *tabula rasa* at birth," Latin expression for "blank slate" (Panksepp and Biven, 2012, p. 427). Therefore, "the widespread claim that affects are just a variant of cognitions seems little more than a word game to me, even though I certainly accept that the many (good and bad) feelings of the nervous system are always interacting with cognitions (imagination, learning, memory, thoughts) within the full complexities of most human and animal minds" (Panksepp and Biven, 2012, p. 489). The point is that "it is through experience that the neocortex is 'programmed' (likely through interactions with subcortical regions) to acquire its capacities that as we reach maturity to come to seem like 'hard-wired' brain functions" (Davis and Montag, 2019, p. 4). With maturation, "these physically, as well as evolutionarily separate brain regions, develop a reciprocal seesaw like the relationship to weigh whether a life event should trigger or inhibit the expression of a primary emotion with imbalances in either direction potentially becoming dysfunctional" (Davis and Montag, 2019, p. 5). The neocortex is organized by the subcortical functions of the brain. These latter guide the neocortex in acquiring and processing information. For instance, Johnson and Horn (1986, 1988) clearly demonstrated it by studying chicks. Alberini (2010) proved that all long-term memory has an emotional component: traumatic events create very strong memories, or they can lead to partial or total memory loss. One of the basic functions of the

primary-process emotional memory systems is arousal and the associated drawing attention to specific events that can facilitate the formation of memories for important life events. These memories in turn can subsequently inform how we respond to future life events.

Panksepp claims that cognitions are often "handmaidens," or emissaries, of the affects, not the opposite. Cognition

> emerges from the neocortex, which is the brain's outermost layer and the part that is evolutionarily newest. This indicates that the capacity for affective experience evolved long before the complex cognitive abilities that allow animals to navigate complex environmental situations. It is also noteworthy that the deeper evolutionary location of the affective systems within the brain renders them less vulnerable to injury, which may also highlight the fact that they are more ancient survival functions than are the cognitive systems. (Panksepp and Biven, 2012, pp. 43–44)

Affects are automatic, instinctual, and innate processes; individual behavior, education, and culture cannot change them. In all the mammalians, the two "brains" (neocortical and subcortical) communicate but are fundamentally different. (On the distinction of "two brains," see Kahneman, 2011). Yet, we cannot understand secondary and tertiary functions if we do not understand primary functions first. This is also confirmed by other data: subcortical neurons function very differently from those of the regions of the neocortex (see Panksepp and Biven, 2012, p. 50).

This distinction between the two brains is important, and it is the reason that leads me to criticize Picard's point of view. Like Damasio, Picard remains too tied to a cognitive conception of emotions and affects. According to Picard, emotions are generated by a cognitive activity (a thought, the knowledge of a state of things, etc.) (see Picard, 1997, pp. 65–66). With this, in my view, Picard does not grasp the essence of human emotional life. This point of view implies that emotion is not something intrinsic to the human brain, but something built from cognitive reflections operated by a human or a machine. On the other hand, Panksepp says that emotion is intrinsic to the brain, and it is the brain that produces the physiological reactions of the body.

Another crucial aspect that emerges from Panksepp's research is the complexity of emotions and the brain. We cannot reduce emotions and affects to simple bipolar systems based on pairs of opposites such as charge/discharge and pleasure/displeasure. It is much more complex; it cannot be reduced to the on/off mechanics of neurons. "The simple-minded neurone-doctrine view of brain function, which is currently the easiest brain model to apply in AI/robotics, under-represents what biological brains really do" (Panksepp, 2008, p. 163). Each basic affective system acts in a different way according to very complex chemical and neurochemical dynamics and equilibria, which we do not yet fully know. Emotions cannot be explained in a dualistic way according to a series of oppositions arranged on three levels: energetic, perceptive, and motor. Each of the fundamental affective systems generates positive or negative states, but, in reality, the distinction between pleasure and displeasure is not clear-cut. There are many intermediate or even superimposed states (so that the same situation generates pleasure in one case and displeasure in another).

Now, I hold that the crucial assumption of an AGI based on the neuropsychoanalytic model of mind is the design of a computational system capable of simulating the seven basic affective systems analyzed by Panksepp. The instantiation of human raw affects must be the fundamental basis of AGI—in the sense that any other activity of the system must be based on them.

Let us see how.

A Freudian computer: sketches

Panksepp's topography of emotions gives us a clear indication of what an emotion is and how basic affective systems and subcortical brain work. How can we translate these indications into an AGI system?

6/1 The Solms-Friston model. As stated above, at the root of our AGI system, there must be seven systems that would be able to instantiate the seven basic affective systems in mammals. "In order to simulate the operations of the human mind, we must consider both the genetic and epigenetic construction of the human brain. We must be clear about what is genetically fundamental and what is epigenetically derivative" (Panksepp, 2008, p. 149). Each basic affective system can be described "in terms of 'state-spaces' that regulate 'information-processing' algorithms" (Panksepp, 2008, p. 149). Can such affective-emotional properties of biological brains be emulated by machines? "Only future work can tell" (Panksepp, 2008, p. 149). For Panksepp, simulating the ancient visceral nervous system is problematic: "a deep understanding of the subcortical tools for living and learning is the biggest challenge for any credible future simulation of the fuller complexities of the mind. The cognitive aspects may be comparatively easy challenges since many follow the rules of propositional logic" (Panksepp, 2008, p. 150). The crucial question is whether and how an algorithm can instantiate complex subcortical circuits. What kind of logic should we follow? This issue "may require a complete re-thinking of where we need to begin to construct the ground floor of mind" (Panksepp, 2008, p. 152). Panksepp comes to express skepticism about the possibility of accomplishing this feat. "I have no confidence that the natural reality of those processes can be computed with any existing procedures that artificial intelligence has offered for our consideration" (Panksepp, 2008, p. 152; see the first attempt of this project: Dietrich et al., 2007).

Is Panksepp's skepticism justified? Today, the tools of statistical rationality and the evolution of technology can drastically change the situation. The last twenty years have been marked by what can be called "Bayesian turn." In particular, "the application of Bayesian formulations to the study of perception and other processes described as problems of inference has generated a huge literature, highlighting a large interest in Bayesian probability theory for the study of brains and minds" (Bruineberg et al., 2020).

Solms and Friston (2018) demonstrate that it is possible to create a statistical modelization of affects following the indications of neuropsychoanalysis and computational biology. The Solms-Friston model is based on the hypothesis that the self-organization of autonomous organisms can be represented in statistical terms. The key idea is that a biological system is able to adapt to its environment and predict possible future states in order to maintain homeostasis. This ability can be described as a statistical procedure of evaluation and inference—in other words, a Markov blanket. "Markov blanket defines the boundaries of a system in a statistical sense. It is a statistical partitioning of a system into internal states and external states, where the blanket itself consists of the states that separate the two" (Kirchhoff et al., 2018, p. 1). The most intuitive example is that of a cell; the boundaries between the cell and its environment can be described by a Markov blanket—that is, as a set of variables separated from another set of variables called "external states" that are independent of each other. Here "states" mean "any variable that locates the system at a particular point in state-space; for example, the position and momentum of all the particles constituting a thermodynamic system—right through to every detail of neuronal activity that might describe the state of the brain" (1).[2]

This entails that

- The external states are conditionally independent of internal states, and vice versa; thus, internal and external states can influence each other only via sensory and active states. External states cause sensory states that influence, but are not influenced by, internal states. These latter cause active states that influence, but are not influenced by, external states. The distinction between external and internal states involves a process called active inference, which tries to predict the external states in correspondence to the internal states. The active inference can be described "in terms of approximate Bayesian inference and probabilistic beliefs that are implicit in a system's interactions with its local surroundings" (Kirchhoff et al., 2018, p. 2; for the general concept of inference in statistics, see Bruineberg et al., 2020, pp. 5–10); therefore, the functioning of the Markov blanket must be interpreted as a probabilistic inference (Bayesian inference) in which the variables correspond to the levels of confidence in the occurrence of an event.
- The goal of "active inference" is the minimization of free energy, i.e., the energy available to the system to do useful work, and therefore the reduction of entropy (uncertainty, surprise, etc.) within the system. Thus, the Solms-Friston model presents a teleological interpretation (the goal is the minimization of free energy) of Bayesian inference.

Therefore, biological systems "have a capacity to maintain low entropy distributions over their internal states (and their Markov blanket) despite living their lives in changing and uncertain circumstances" (Kirchhoff et al., 2018, p. 3). This scheme should not be considered as a fixed structure. A system can have many Markov blankets, "the boundaries of which are neither fixed nor stable" (2). Furthermore, Kirchhoff et al. (2018) distinguish two different types of active inference: mere active inference and adaptive active inference. Only the latter enables autonomous organization. This aspect is very important: even a cell-like any other autonomous biological system—is capable of biological and unconscious inference in order to maintain its integrity. Thanks to the Markov blanket, we can build a mathematical modelization of this process.

Solms and Friston (2018) apply this general scheme to the psychoanalytic model of the mind. The main thesis is that we can explain consciousness and affects through the principle of free energy minimization and the concept of the Markov blanket, as defined before. Then, the goal of the psychic system is to minimize the amount of free energy used to reduce entropy—i.e., to maintain homeostasis. The ideal state corresponds to a situation where the free energy is zero—a state without entropy. The possible outcomes of the active inference in the psyche are three: (a) to change the sensation that comes from the outside; (b) to change the internal representation of the sensation and therefore the prediction of future sensations; and (c) to try to match prediction and sensation more and more, improving the confrontation between reality and expectations—the optimization of precision with respect to free energy. In short, consciousness is an inferential process—a self-evaluation process—that aims to predict changes in the use of free energy and formulate strategies against entropy. This process takes place simultaneously on different levels: sensory, motor, internal, external, etc. "What we describe is an elemental form of a self-maintaining mechanism that takes more complex forms in more complex biological systems (like vertebrates)" (Solms and Friston, 2018, p. 28). Qualitative fluctuations in felt affect "arise continuously from periodic comparisons between the sensory states that were predicted (based upon a generative model of the viscera and the world and samples of the actual sensory states)" (28). Consciousness comes from this biological mechanism of affective self-regulation.

In other terms, to maintain its integrity, the organism responds to the external stimulus (sensory state) by changing its internal state and its environment. This process can be interpreted in terms of a Bayesian inference that aims at "inferring the *most probable, hidden causes of sensory signals* in terms of expectations about states of the environment" (Kirchhoff et al., 2018, p. 4; my emphasis). In other words, the organism tries to predict *the cause of a sensory state* in relation to its expectations and then produces an active state in order to minimize free energy. In the Solms-Friston model, free energy corresponds to prediction errors; "the recurrent assessment of sensory states only gives rise to changes in subjective quality (i.e., precision and feeling) when the amplitude of prediction errors *changes*—signaling a change in uncertainty about the state of affairs and, in particular, the consequences of action" (Solms and Friston, 2018, p. 28). In this context, the concept of "hidden cause"[3] is essential: "What is seen does not cause what is felt. Both have *hidden* causes. Consciousness (both exteroceptive and interoceptive) involves the quest for these unitary hidden causes, which must be inferred from the two sets (i.e., modalities) of data and *explain them both*" (29).

Is this scheme really satisfying to describe Panksepp's emotional systems? In my opinion, it is not. If we take Panksepp's seven systems as a point of reference, none of them can be explained only on the basis of the Solms-Friston model and Bayesian networks. In the next section I want to formulate some criticisms of the Solms-Friston model and propose a new model inspired by Pearl's theory of causality.

6/2 How Panksepp's systems can be organized and embedded in AGI. At the beginning of the eighties, Panksepp (1982) was convinced that there were at least four biological brain-based emotional action systems, which were Expectancy, Rage, Fear, and Panic. In the nineties, especially with the publication of *Affective Neuroscience* (Panksepp, 1998), Panksepp expanded his list of primary emotions to seven primary-process emotional command systems: SEEKING/Expectancy, RAGE/Anger, FEAR/Anxiety, LUST, CARE/Nurturing, PANIC/Sadness, and PLAY/Social Joy. The core of his approach was "mapping of the seven primary emotional systems by means of electrical stimulation of the mammalian brain, including pharmacological challenges and brain lesions" (Davis and Montag, 2019, p. 2). The ESB and DBS techniques have shown that all mammalian brains work in a very similar way; distinct affects can be linked to pretty much the same areas of the brain and the same type of electrical or pharmacological stimuli. Panksepp's hypothesis is that (1) imbalances in these primary emotional systems are strongly linked to psychiatric disorders, such as depression or suicidal thoughts, (2) we can act on these systems to modify and cure these imbalances. One of the most important confirmations of this was the Affective Neuroscience Personality Scale (ANPS), according to which the primary-process emotions constitute the psychobiological foundations of personality (Davis and Montag, 2019). The study of primary affective systems has allowed us to better understand how the personality is born and evolves, what are its main disorders, and how we can cure them. Panksepp has also made several predictions about psychiatric treatments. For instance, "[he] predicted that autistic children might have dysfunctional brain opioid systems resulting in excess endogenous opioid levels" (Davis and Montag, 2019, p. 24). Following this approach, Montag et al. (2017) explained depression, and Yovell et al.

(2016) explained suicidal thoughts. Many other research works confirmed the value of Panksepp's work.

Now, the questions I want to ask are the following: (1) How do these ancestral emotional systems work? (2) How can we translate them into algorithms? This second question implies another one: Is the Solms–Friston model able to fully formalize the ancestral emotional systems, or do we need other theoretical tools?

Carrying out a detailed analysis of ancestral emotional systems would require not a simple paper but several books. Here I will analyze only the SEEKING system because—as Panksepp states—this system is the oldest and most general; it always continues to operate in the background in all our brain's activities. When the SEEKING system is very active, rats move with a specific purpose, vigorously sniffing and exploring where they are, even making an inaudible sound. The same thing happens in humans when they experience feelings of anticipatory craving; they feel as though they are effective agents in the world and are happy with what they do. The SEEKING system causes a positive sense of wanting and being able to do. Therefore, it produces expectations about what can be done and how it can be done and, ultimately, pushes us to act in a certain way. "It is your subcortical SEEKING system that helps energize your neocortex—your intellect—and prompts you to do things like buy this book and also to learn from books, if they are engaging" (Panksepp and Biven, 2012, 102). When this system is underactive, mammals feel depressed and hopeless; or, due to its hyperactivation, they can become psychotic. "It is evident that the SEEKING-EXPECTANCY system is a general-purpose system for obtaining all kinds of resources that exist in the world, from nuts to knowledge, so to speak" (Panksepp and Biven, 2012, p. 103).

The activation of the SEEKING system can take place in two ways: due to homeostatic imbalances or more complex negative emotions, such as loneliness or pain. In the first case, some nerve cells located in the ancient regions of the brain or in some body organs (Panksepp and Biven, 2012) register homeostatic imbalances (thirst, hunger, cold, etc.), which are a problem. The SEEKING system works not only to respond positively to the problem but also to give us hope, i.e., a positive purpose, and push us to act. It is not simply a positive reaction to an external stimulus but something much more complex. It is the reaction to discomfort *and* a precise plan of action. If it is cold, the mammal immediately seeks a suitable solution to that situation: a shelter, or a blanket. In other cases, in humans, low levels of endogenous opioids (such as endorphins) can activate the SEEKING system, whereas variations in hormone levels favor the activation of the LUST system.

Panksepp's research shows that ancestral emotional systems are not simple automatic ways of reacting to certain stimuli. They are much more. "The SEEKING system is driven by brain dopamine, but it is much more than just the creation of that one energizing neurotransmitter. It is a complex knowledge-generating and belief-generating machine" (Panksepp and Biven, 2012, p. 103). The complexity of the SEEKING system is confirmed, for example, by its connection with the sense of time. As Panksepp and Biven (2012, p. 138) explain, "The dopamine-containing neurons of the SEEKING system have such endogenous pacemakers that normally keep them firing at a stable monotonous rate, like the ticking of a clock, especially when nothing special is happening to an animal"; these neurons "even keep firing when animals are asleep, but the background activity is not normally attended by the release of dopamine." The ancestral affective systems are not automatic ways of reacting to certain stimuli but complex and dynamic causal networks.

Now, is there a whole structure of the SEEKING system that can also be identified in the other ancestral affective systems? In other words, is there a common structure shared by the emotional systems? In the SEEKING system, we can identify four crucial phases: (a) the generation of emotion for internal or external reasons, i.e., the biological reaction; (b) the evaluation of the emotion, which can be positive or negative (pleasure or unpleasure); (c) the anticipation, in the sense that our emotional brain anticipates reality and creates emotional memories and projections of what might happen (for example, when we raise our arms to protect ourselves even if there is nothing threatening us) that can later be processed by other parts of the brain in an ever more refined way; (d) the action, i.e., the production of a series of actions in accordance with the previous moments. This cycle (emotion, evaluation, anticipation, and action) constitutes the first form of learning in mammals. If the action confirms the prediction and evaluation, mammals learn to link the reward to that behavior. Otherwise, they learn that it does not lead to any reward. This improves their ability to adapt to the environment. As I said, this cycle has a causal structure: Emotion causes evaluation and anticipation, which then causes action; but the action can also modify anticipation, evaluation, and emotion.

The same structure can be found in another basic emotional system, the RAGE. Rage is not meant to punish; anger, hatred, and revenge, but also remorse and forgiveness, are cognitive elaborations of rage that many animals do not possess. If some areas of the amygdala, hypothalamus, and periaqueductal gray are electrically activated, a human being clenches their jaw and experiences a feeling of rage without knowing why. The causes that can trigger rage are many (homeostatic imbalances, external factors, other emotional systems, etc.). There are also several chemicals that regulate rage: testosterone, norepinephrine, glutamate, acetylcholine, etc., which behave differently depending on the part of the brain in which they act. However, even in this case, we can distinguish at least four phases in the functioning of the system: (1) the generation of an emotion (with the release of some chemicals), (2) the evaluation of the emotion (pleasure or displeasure), (3) the production of more or less cognitively elaborated anticipations, and (4) the action. For example, (1) hunger and scarcity of resources fuel the release of certain chemicals that produce rage; (2) rage makes you feel bad, as it is a negative feeling; (3) this negative feeling can trigger delusional fantasies of persecution and revenge, and, finally, (4) some actions that will either tend to eliminate the origin of the rage (hunger) or those fantasies. Again, the system is a complex causal network. This is even more evident in the LUST system. Sexual stimulation increases the production of testosterone (in males), or estrogen (in females), and generates a feeling of general well-being, as well as leading to the activation of bodily systems and certain types of behavior (for example, a sexually receptive body posture, a particular kind of smell, or erection, copulation, and courting) (see Panksepp and Biven, 2012, p. 235). Furthermore, sexual desire is closely related to the SEEKING system that is recruited for the task of seeking a sexual partner—one system affects the other by conditioning it. An important example is that of sadomasochism: a painful and unpleasant stimulus triggers emotion, i.e., sexual desire, and, therefore, the release of hormones (Panksepp and Biven, 2012, p. 245).

The example of sadomasochism illustrates an important point. The distinction between pleasure and unpleasure in the second phase (evaluation) is very vague in the sense that it can vary. Rage is a perfect example of this: There are people for whom rage is a positive feeling in the sense that they feel good when they feel anger and attack others (even if, in the long run, rage and anger have very negative effects and are unsustainable). As Panksepp points out (see Panksepp and Biven, 2012, p. 148), the boundary between pleasure and unpleasure, as well as the nature of anticipations and actions, depends on (1) which other areas of the brain are acting at that moment, (2) which other affective systems

are interacting with each other, and (3) which chemicals trigger the whole process. These are very complex dynamics. Let us consider, for example, the system of FEAR. An emotion triggers anxiety and, therefore, a series of anticipations (fantasies, memories, etc.) and actions (increased heart rate and blood pressure, sweating, flight, etc.). However, anxiety is not necessarily a negative feeling that causes pain. When we watch a horror movie, for example, we seek that situation; we want to feel fear, and this gives us pleasure. The relationship between pleasure and unpleasure varies according to the structure of the affective system or the relationships of an affective system with the others. It varies according to the type of causal network involved. Every emotional system is a set of causal networks that are not organized by the dualism of pleasure and unpleasure.

We now come to the second question that we started with: How can we translate ancestral emotional systems into algorithms?

Here, I propose to integrate the Solms–Friston model with Pearl's causal calculus. I think that Pearl's causal calculus is a more plastic tool and that, precisely for this reason, it allows one to better explain (and, therefore, model) the complexity of ancestral emotional systems. In the following parts, I will justify this idea.

Is the Solms–Friston model really satisfying for describing Panksepp's ancestral emotional systems? In my opinion, it is not for two main reasons:

1. The model remains too tied to the Freudian dualism of pleasure–unpleasure, which becomes a general scheme used to explain all effects and emotions. However, as shown by Panksepp's research, the differences between pleasure and unpleasure are never clear-cut and can often vary between the different affective systems and within the systems themselves. I think that the free energy principle is more useful for describing what Panksepp calls homeostatic affects and sensory affects. More complex and plastic models are needed in order to describe emotional affects that are networks of cause-effect relations. In short, the pleasure–unpleasure couple does not capture the essence of emotion.
2. The Solms–Friston model is based on the concept of Markov Blanket, which presupposes that of Bayesian network. My question is this: Do these concepts really have anything to do with causation? With regards to this point, Pearl's critique of the Bayesian network seems very compelling. Pearl argues that we cannot define causality purely in probabilistic terms, i.e., A is the cause of B because it increases the probabilities of B. Probability and causality are different concepts. If we follow Pearl's critique and his concept of the "Ladder of Causation" (Pearl and Mackenzie, 2018, pp. 28–29) (more on this below), we have to recognize that the Solms–Friston model still stands at the first rung of the ladder and cannot move toward the upper rungs. The model explains causation in purely probabilistic terms, i.e., as a correlation. In a nutshell, it confounds causation and correlation.

I will try now to justify these two theses. The concept of causality plays a central role in the Solms–Friston model (see Friston 2009). According to Friston, biological self-organizing systems are based on a Markov blanket and, for this reason, they are capable of active inference; in other words, "The partition of states implied by the Markov blanket endows internal states with the apparent capacity to represent hidden states probabilistically so that they appear to infer the hidden causes of their sensory states" (Friston, 2013, p. 6). Thanks to the Markov blanket, a "circular causality" (Friston, 2013, p. 6) takes place in the system. This circular causality connects sensory states and active states; "sensory states depend on active states rendering inference active or embodied" (Friston, 2013, p. 6). This circular causality allows one to limit the surprise in the system, gradually adapting expectations and confirmations, and, therefore, it allows one to limit the free energy and resist entropy, i.e., the dispersion of the system. Homeostasis "is informed by internal states, which means that active states will appear to maintain the structural and functional integrity of biological states" (Friston, 2013, p. 6).

As is evident from these passages, the active inference is a type of Bayesian inference. My question is the following: Is the circular causality, which, according to Friston, is the core of active inference, really a causality or only a correlation between variables? As mentioned above, I refer to Pearl's critique of the Bayesian network—a concept created by Pearl himself—which is the basis of the concepts of active inference and the Markov blanket. All current machine learning systems are based on Bayesian networks (Pearl and Mackenzie, 2018, pp. 122–128). Nonetheless, according to Pearl, the Bayesian network alone does not grasp the essence of causality and cannot express it.

Let us now discuss some characteristics of a Bayesian network. Then, I will introduce the core of Pearl's critique. Thanks to the Bayesian network, it is possible, starting from a certain set of data (probabilities), to calculate (a) the probability of the causes (e.g., symptoms → disease) and, therefore, (b) the recurrence of other similar events in the future. For this reason, Bayesian networks are used to develop machine learning algorithms. The essence of the Bayesian network is the calculation of inverse probability: Starting from one set of probabilities (effects), we arrive at another set of probabilities (causes). Now, the Bayesian network is highly dependent on the available data—on a particular set of data. It does not involve the formulation of a hypothesis, i.e., general models on causality; it does not apply a model to the data and only calculates the probability of unknown events (causes) starting from the available data (effects). This means that, in the Bayesian network, causality amounts to the increase of the probability of the effect. However, this equivalence (A causes B because it increases the probability of B) is wrong because the increase in the probability of B is not a sufficient criterion to make A the cause of B. Indeed, alone, the increase in the probability of B does not allow us to understand if there are hidden causes of B, or indirect causes, or even backgrounds factors that influence B and its probability. This is the so-called problem of the "confounder." Causation implies increasing the probability of the effect but is not limited to that. To determine the cause, we need a hypothesis, i.e., a theoretical model that must be tested and that is independent of the data.

This is the essence of Pearl's argument against the Bayesian network: "In both a cognitive and a philosophical sense, the idea of causes and effects is much more fundamental than the idea of probability" (Pearl and Mackenzie, 2018, p. 46). According to Pearl, the Bayesian network is essential to understand causality, and, yet, it is not enough. Causation requires going beyond the data, hence creating more complexity. For this reason, Pearl develops specific tools such as causal diagrams and do-calculus. These mathematical methods solve the problem of confounding and related paradoxes. "Whereas a Bayesian network can only tell us how likely one event is, given that we observed another […] causal diagrams can answer interventional and counterfactual questions" (Pearl and Mackenzie, 2018, p. 130).

Let us now have a closer look at Pearl's theory of causation. Pearl describes three levels of causal inference, what he calls the "ladder of causation" (Pearl and Mackenzie, 2018, pp. 28–29):

1. *Association*: being able to find phenomena that are related; most animals can do this to some extent, and most machine

learning models are trained to learn associations between variables. Example: *What is the expected lifespan of somebody who is vegetarian and does not smoke?* This level is that of the simple observation of facts and their correlation. It answers questions such as "What if I see?"

2. *Intervention*: being able to guess what the effect will be if one performs an action, i.e., it changes the value of a variable. Such higher-level understanding is typical of more intelligent animals and is related to the topic of reinforcement machine learning. Example: *How would my expected lifespan change if I become a vegetarian? What if I ban cigarettes?* This level is that of the action changing the facts. It answers questions such as "What if I do…?"
3. *Imagining*: being able to reason about hypothetical situations, things that *could* happen and not that have already happened. Imagining is typically done in intellectual activities, such as performing thought experiments, making up a story, etc. Example: *Would Kennedy be alive if Oswald had not killed him? Would my grandfather still be alive if he did not smoke?* This level is that of the imagination and possible worlds. It answers questions such as "What if I had done…?"

These are fundamentally different concepts, which require different mathematical tools to be described. The first level, dealing with associations, is studied using the rules of the probability theory and can be learned from data using statistical methods. The second level deals with interventions. To assess the effect of interventions, one either has to perform a suitable experiment (which might be expensive or even not possible) or be able to determine the causal relations among the variables of the system. Data alone cannot help us answer action-related questions at the second level of the ladder. If you have a database of high school pupils with their curricula and test scores, you could easily see that pupils who follow advanced math courses tend to score better on a standardized mathematics test. Would enlisting all students in such advanced math courses improve their mathematics understanding? Not necessarily. It is possible that students enrolled in such classes are naturally more gifted in mathematics. Adding pupils who do not have a knack for mathematics would, in this case, not help; or worse, it could demotivate them, resulting in an even worse grade. We must go beyond the data to understand what the exact dynamics of the facts and the causes of the processes are.

The final level is even more challenging, as it deals with reality as it would be if the circumstances were different. In this case, there is no data available, nor could we ever perform experiments. The results of the queries at this level are called *counterfactuals*. To make statements about such hypothetical situations, we need an intricate understanding of the system and how all its parts are linked together. If we return to the example of mathematics education, I could determine what my score would be if I had taken the advanced mathematics course. This would not only account for all the things that I would have learned during such a course but also involve backtracking who I might have met there, what influence this could have had on my other activities, etc.

In other words, Pearl argues that there is a radical difference between simple association prediction and causation. In the first case, the question is "What is the probability of x given the presence of y?" This is what a machine learning system does with the data in its possession and what the active inference of the Solms–Friston model does too. Starting from the observation of one fact, we can calculate the probability of another fact on the basis of the data in our possession. In the second case, the question is "What is the probability of x if I change the value of y?" In order to calculate how the change introduced in the data influences the probability of x, I cannot use the same mathematical tools as in the first case. Why? The reason is that —as I said before—limiting myself to finding associations could lead me down the wrong path and make me believe that the probability of x is influenced by a variable that, in reality, has no causal function. If there is a correlation between x and the appearance of y, that correlation is not necessarily causal; it could be caused by hidden common causes or background factors. I simply know that, when I look at y, x also appears. But there could be a hidden cause that influences both x and y, or, maybe, there are several different causes—some direct, and others indirect. Knowing a correlation is, in itself, passive knowledge. In order to understand the causal connection between data and to plan my actions, in reality, I need conceptual tools that allow me to distinguish between association and causation, as well as the different types of causation. The causal diagrams and the do-calculus describe this normal activity of the human brain. Thanks to these tools, we are able to modify the laws of statistics in order to statistically determine correlation and causality. Without these tools, as Pearl shows, it is impossible to solve the problem of confounding and some of the paradoxes that arise precisely from the confusion between correlation and causality, such as the Simpson's Paradox or the Berkson's Paradox (I cannot go into detail here; I just refer to Pearl and Mackenzie, 2018, pp. 197–211). That is the reason for distinguishing between Bayesian networks and causal diagrams: "Bayesian networks inhabit a world where all questions are reducible to probabilities or degrees of association between variables; they could not ascend to the second or third rungs of the Ladder of Causation" (Pearl and Mackenzie, 2018, p. 51). The main point is this: "While probabilities encode our beliefs about a static world, causality tells us whether and how probabilities change when the world changes, be it by intervention or by an act of imagination" (Pearl and Mackenzie, 2018, p. 51).

Let us be even more precise in distinguishing between a Bayesian network and a causal diagram. How are these two theoretical structures distinguished? According to Pearl, "A causal diagram is a Bayesian network in which every arrow signifies a direct causal relation, or, at least, the possibility of one, in the direction of that arrow. Not all Bayesian networks are causal, and in many applications, it does not matter" (Pearl and Mackenzie 2018, p. 95). However, "if you ever want to ask a rung-two or rung-three query about your Bayesian network, you must draw it with scrupulous attention to causality" (Pearl and Mackenzie, 2018, p. 95) and transform your Bayesian network into a causal diagram.

An objector might ask, at this point, what causation is from Pearl's perspective. What is the general concept of the cause that guides Pearl's analysis? Pearl's guiding idea comes from the philosopher David Kellog Lewis and his theory of counterfactuals. Lewis claims that "we think of a cause as something that makes a difference, and the difference it makes must be a difference from what would have happened without it. Had it been absent, its effects—some of them, at least, and usually all—would have been absent as well" (Lewis, 1973, p. 161). I would say that the ladder of causality itself is a definition of causality. It can be conceived as a test to answer the question "Can A be the cause of B?" A variable has a causal influence on another if and only if (1) it is observable when the other variable is present (association), (2) it makes a difference, in the sense that, if its value changes, the value of the other variable also changes (intervention), or (3) the difference does not occur without it (counterfactual). The do-calculus is a mathematical formalization of this fundamental idea, which is actually an instinctive process in the human being—we are instinctively led to recognize causality. In a nutshell, causal diagrams and the do-

calculus allow us to identify and mathematically define what really "makes the difference" in a complex of changing variables.

Here is the difference between the active inference of the Solms–Friston model and Pearl's techniques. If we take seriously Pearl's critique of the Bayesian network, we have to admit that active inference remains a classic model of machine learning based on association prediction, which is unable to properly analyze causality. It remains on the first rung of the Ladder. In other words, Bayesian networks and active inference fail to grasp what "makes the difference." *Pearl's aim is to understand how to describe what* "makes difference" *in probabilistic terms*. As I said above, I think that the Solms-Friston model can describe better the homeostatic and sensory affects, and not the emotional ones, following Panksepp's terminology.

These considerations lead us to the answer to the second question that we started with: How can we translate the emotional systems described by Panksepp into algorithms? Through causal diagrams and the do-calculus, we can formalize and algorithmize the complex behavior of all emotional systems because we can formalize and algorithmize the causal networks that these involve. The do-calculus and the causal diagrams allow us to recognize the causal relationships between data and to separate them from the simple associations, therefore allowing us to intervene effectively, i.e., to plan the causal action.

Now, in the emotional systems described by Panksepp, we can identify at least four types of causal relations: (a) those internal to the system (among the four variables we have described); (b) those between the system and external reality; (c) those between the system and the other emotional systems; (d) those between the system and the rest of the brain. These four types of relationships can be translated into a set of causal diagrams. The diagrams sort the input data according to possible causal chains, then the do-calculus operates on them to "set them in motion," transform them, and calculate the consequences (the output data) of their transformation according to the data. Each emotional system can therefore be translated into a series of causal diagrams that allow the analysis and interpretation of data. For example, several causal diagrams can be drawn to simulate the behavior of oxytocin, a crucial hormone in the LUST system. All these causal diagrams constitute the memory of the system. One or more processors operate the do-calculus on this memory. The result is a system that can think causally. The plasticity of this model is guaranteed by two factors: (a) the causal diagrams can be varied and interchangeable according to the different situations—they do not depend on a single set of data like Bayesian networks; (b) the calculus can modify the diagram, adapting it to the needs of the situation.

There are two fundamental conditions for the realization of an AGI system based on this type of model: (a) The design of the system will presuppose an enormous research work on animals and humans and use this to construct the most exact causal diagrams to model the behavior of the different subcortical systems[4]; b) it will be essential to set the right conditions for the evolution of the computational system that instantiate the emotional systems—we are used to thinking of the AGI in terms of adult human beings, but this is completely wrong.

There are other two big benefits of following Pearl's indications. As I mentioned, the do-calculus can also give a mathematical representation of the counterfactual inference (see Pearl and Mackenzie, 2018, pp. 269–280). Now, counterfactual reasoning involves imagination and other complexes/cortical cognitive functions because it involves not only the ability to think of the world in a different way, i.e., an alternative world, but also the ability to reflect on one's actions. Counterfactual reasoning is a form of self-awareness. The do-calculus, therefore, offers us a unique mathematical language that allows us to (a) think causally and (b) reflect on one's own causal chains. This is an important point: The do-calculus also allows us to explain how an emotion (understood and formalized in a causal way) can give rise to an elementary form of consciousness and the development of complex/cortical cognitive functions. This confirms one of the fundamental ideas of Panksepp's neuroscience: Emotion is the basis of consciousness and higher cognitive processes.

The second benefit is that Pearl's do-calculus is logically complete, as has been shown by several groups of researchers. Completeness in mathematics means that an axiom system "has the property that the axioms suffice to derive every true statement in that language" (Pearl and Mackenzie, 2018, p. 237).

6/3 How to build the system: a concrete example. Do-calculus is essentially a mathematical technique to treat causality in probabilistic terms avoiding the problem of confounding, that is, mixing correlation and causality in a group of variables.[5] It is an axiomatic system that allows for the examination of a causal diagram and to "purify" it of any possible spurious correlations identifying causal connections and translating them into probabilistic terms. It is a method "that can, astoundingly, tease out causal information from purely observational data" (Pearl et al., 2016, p. 55). In more precise terms, do-calculus "is an axiomatic system for replacing probability formulas containing the *do*-operator with ordinary conditional probabilities. It consists of three axiom schemas that provide graphical criteria for when certain substitutions may be made" (Hitchcock, 2018).

To better understand how to translate Panksepp's seven systems into algorithms we need to better understand Pearl's causal theory and do-calculus. According to Pearl, causation *can be interpreted as increasing the probability of the effect*; however, in order to so, we need mathematical tools that are able to resolve the difficulties that probabilistic causation has encountered in the past and clarifies what relationships exist between probabilities and causation.

Let us say that a conditional probability such as $P(Y = y|X = x)$ $P(Y = y|X = x)$ gives us the probability that Y will take the value y, given that X has been *observed* to take the value x. Do-calculus allows us to predict the value of Y that will result if we *intervene* to set the value of X equal to some particular value x. Pearl writes: $P(Y = y|do(X = x))P(Y = y|do(X = x))$ to characterize this probability. The *do*-operator identifies the intervention. When we intervene on a variable in a model, "We fix its value"; this means that: "We change the system, and the values of other variables often change as a result" (Pearl et al., 2016, p. 54). The *do*-operator allows us to introduce the intervention in the causal diagram: "When we intervene, we override the normal causal structure, forcing a variable to take a value it might not have taken if the system were left alone. Graphically, we can represent the effect of this intervention by eliminating the arrows directed into the variable intervened upon. Such an intervention is sometimes described as 'breaking' those arrows" (Hitchcock, 2018). For example, if I write:

$$P\big(Y = y|X = x, do(Z = z)\big),$$

I assume that the action $do(Z = z)$ is being performed in the actual world; hence, I observe the values that other variables take $(X = x)$ in the same world that the intervention takes place. Graphically, the structure of the diagram changes in the sense that any arrow that goes toward the node that represents the intervention is eliminated. No arrow leads to this knot, which is a cause.[6]

Starting from these considerations, I affirm that an emotional system can be formalized in Fig. 1:

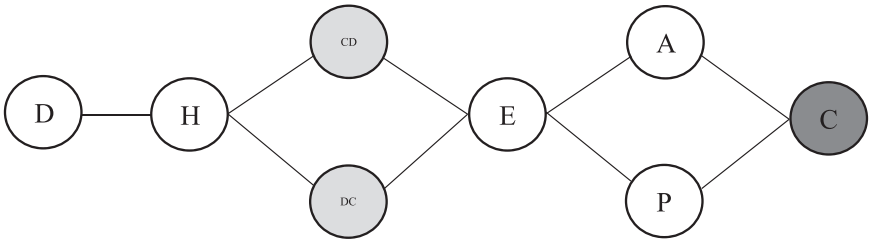

Fig. 1 The formalized structure of an emotional system in our AGI. The diagram respects the four distinct phases previously identified: (**a**) the generation of emotion for internal or external reasons; (**b**) the evaluation of the emotion; (**c**) the anticipation-prediction; (**d**) the action, i.e., the production of a series of actions in accordance with the previous moments.Our AGI system analyzes a set of data (D) that can derive from external reality, from other emotional systems, and from other parts of the brain and body. Using this set of data, the system elaborates causal hypotheses (H) based on some fundamental assumptions or beliefs. The "heart" of the system is the evaluation (E), or inference, which is realized through the causal diagram (CD) and the do-calculus (DC). The evaluation allows for the elaboration of predictions (P), actions (A), and counterfactuals (C).

This is the basic form of our artificial general intelligence (AGI) system. The system analyzes a set of data (D) that can derive from external reality, from other emotional systems, and, finally, from other parts of the brain and body. Using this set of data, the system elaborates causal hypotheses (H) based on some fundamental assumptions or beliefs. For example, in the case of the SEEKING system, in a situation of danger, a fundamental assumption would be "to seek adequate shelter," or, in a situation of hunger, "to seek food resources." Hypotheses are elaborated based on these assumptions and the data available, and have the following form: "What happens if I do *x*?" "What is the effect of *x* on *y*?" The system is activated when the data indicate a situation of danger or discomfort and immediately elaborates hypotheses. The "heart" of the system is the evaluation (E), or inference, which is realized through the causal diagram (CD) and the do-calculus (DC). These two tools act simultaneously and allow us to analyze and interpret data and variables, and thus identify causal connections in probabilistic terms. These tools allow for the elaboration of predictions (P) and actions (A)—the evaluation can be represented in a more complex and articulated way by the diagram of the inference engine described in Pearl and Mackenzie (2018, p. 12). Two aspects must be underlined: (a) the causal diagrams and the do-calculus are continuously updated in relation to the new data available, therefore to the success or failure of the chosen strategy; (b) the system must elaborate and analyze many interlinked hypotheses at the same time, and therefore elaborate many evaluations, and often also reflect on what happened in counterfactual terms (C).

Let us take a very simple example. A man is dying of thirst and is lost in the middle of a forest. The SEEKING system is activated to fix the situation. The man has two paths in front of him, one of which leads to a river, thus to survival. According to his memory of the forest, almost certainly at least one path leads to the center of the forest, and therefore walking it would mean losing any possibility of reaching the river. Also based on his memory, there is another path that leads to the river and to salvation, but it could be too long; if it is, the man risks dying of thirst. Moreover, there are two variables to consider: (a) a friend who often passes by the forest could help him choose the right path—or even give him water; therefore, the best choice would be to wait for him; (b) that the climate gets hotter and warmer.

A very simple example of a causal diagram could be this (see Fig. 2):

The causal diagram describes a world. It is a way to represent and sort the data we have and define probabilities. In this case, data is about the relation between the system and the external reality. According to Pearl, our brains use exactly this type of representation: "Humans must have some compact representation of the information needed in their brains, as well as an effective procedure to interpret each question properly and extract the right answer from the stored representation" (Pearl and Mackenzie, 2018, p. 39). The arrows can be translated into probabilistic formulas and be modified by inserting the *do*-operator. "Behind the arrows, there are probabilities. When we draw an arrow from *X* to *Y*, we are implicitly saying that some probability rule of a function specifies how *Y* would change if *X* were to change" (Pearl and Mackenzie, 2018, p. 45).

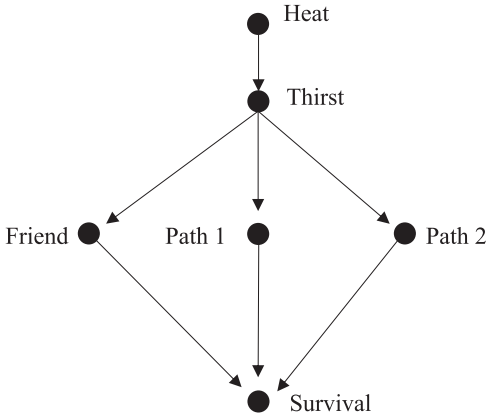

Fig. 2 The man has two paths in front of him, one of which leads to a river, thus to survival. A friend who often passes by the forest could help him choose the right path—or even give him water; therefore, the best choice would be to wait for him. However, the climate gets hotter and warmer.

Let us now apply the *do*-operator to our diagram; this means that we introduce an intervention into the diagram, an action. For example, our man thinks that path 1 is correct and chooses to walk along it to arrive at the river. However, this choice implies three other variables: the conditions of the *ground* could render the path impervious; ferocious *animals* could attack him, or traveling in this way could take too much *time*. The objective of the SEEKING system is to calculate the causal relationship between path 1 and the arrival at the river, hence survival. In doing this, the system must calculate the weight of the three aforementioned variables. Considering the variable *time* is a characteristic of the SEEKING system, as we have seen before. The variable concerning animal attacks obviously comes from the FEAR system that interacts with the SEEKING system.

Let us see how the diagram changes through the introduction of the do-operator (Fig. 3):

To evaluate what the man should do, the SEEKING system searches for the best course of action. To do this, the system must

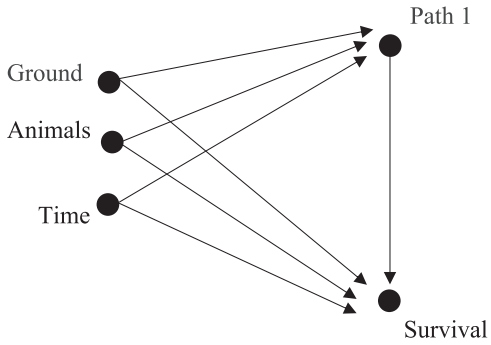

Fig. 3 How the diagram changes through the introduction of the do-operator. This diagram describes the behavior of the affective systems involved in the situation.

evaluate the weight of the variables and identify the more likely causal connection; that is, what really "makes the difference" in this situation. Using other terminology, the system must identify the "confounders," i.e., those variables that can produce spurious correlations and prevent the identification of causal relationships, as mentioned above. In this situation, the *ground*, *animals*, and *time* can be considered as confounders. (It should be noted; however, that there are many strategies for recognizing a confounder—some researchers may define a confounder as something that others do not).

The purpose of do-calculus is to identify what Pearl calls the "deconfounders"; that is, the variables that allow identification of the confounders to clearly distinguish the causal connections (the interventions that have real causal potential) and non-causal effects. Pearl's idea is that identifying a group of deconfounders allows (or is tantamount to) computation of the probability of the causal effect of each variable. "If you have identified a sufficient set of deconfounders in your diagram, gathered data on them, and properly adjusted for them, then you have every right to say that you have computed the causal effect $X \rightarrow Y$ (provided, of course, that you can defend your causal diagram on scientific ground)" (Pearl and Mackenzie, 2018, p. 139). Pearl proposes two "deconfounding techniques": using a back-door criterion or a front-door criterion, which allow, provided the data are available, to identify the deconfounders. Do-calculus overcomes them and goes further: its three axioms allow identification of the deconfounders *even without having experimental data*. "Inspired by the ancient Greek geometers, we want to reduce the problem to symbol manipulation and in this way wrest causality from Mount Olympus and make it available to the average researcher" (Pearl and Mackenzie, 2018, p. 233). In other words, applied to the real world, do-calculus makes it possible to identify the most effective action to achieve a goal through a simple combination of symbols (obviously, only if our causal diagram is correct). The three axioms of do-calculus (Pearl and Mackenzie, 2018, p. 236) allow the system to (a) analyze the transformations of the probabilities, (b) identify causal relationships, (c) make predictions, and (d) design new action plans. This is carried out through a simple combination of symbols.[7]

This is exactly what happens in our trivial example. The man wants to understand the effect of x (walking path 1) on y (getting to the river and drinking). The problem is whether x is a confounder, i.e., a false cause, or not. What can he do? Find new data on x and analyze new variables (*ground*, *animals*, and *time*). For example, by advancing a little along path x, the man could find out that the way is crossed by a shortcut that leads directly to the river. The probabilities in the diagram change; thus, the diagram itself changes.

As I said, my example is trivial. Pearl applies his deconfounding techniques to much more serious problems, such as the use of fertilizers, the link between smoking and cancer, or global warming. However, my thesis is that our emotional systems work this way; deconfounding techniques represent the fundamental ways of action and learning of our limbic system. This is a basic level of learning. Therefore, it is possible to represent and interpret the behavior of each emotional system through causal diagrams and the do-calculus. Panksepp's seven emotional systems are nothing more than sets of patterns of action and learning. The advantage of using causal diagrams and do-calculus is that these tools can be applied to all four types of causal relationships that I have distinguished. Causal diagrams can be continuously transformed based on the data.

In summary, translating an emotional system into an AGI means:

- Translating causal diagrams and do-calculus into algorithms.
- Defining a data classification system that may concern (a) internal states of the system, (b) relations between the system and external reality, (c) relations between the system and other emotional systems, and (d) relations between the system and other parts of the brain.
- Defining exactly the principles of each system, i.e., the objectives (for instance: to seek a solution to a problem, to satisfy sexual desire, etc.).
- The heart of the system will be a processor that must be able to produce causal diagrams and interpret them through do-calculus, which adapts to new situations and learns from them: it has to be able to develop the ability to "draw" increasingly complex diagrams on its own and self-correct.
- The set of causal diagrams should be classified in relation to the different types of data, as I discussed in the previous section.
- The system must produce counterfactuals, i.e., a type of reflection and retro-action on the system itself. Is it a real human-like consciousness? I do not know. But, as Russell (2019) writes, "for AI purposes this makes no difference" (17).

An objector might ask: How do the causal diagrams and the do-calculus we use to describe emotional systems differ from those we can use to describe and model cortical cognitive systems? The difference is in the type of causality we use. In the following, I will clarify two points:

- The distinction between basic emotions and cognitive-oriented emotions.
- The distinction between the do-calculus describing the cognitive/cortical processes and the do-calculus describing the subcortical/affective processes.

In their seminal book, Collins et al. (1994) define emotions as "valenced reactions to events, agents, or objects, with their particular nature being determined by the way in which the eliciting situation is construed" (13) Thus, "the particular emotion a person experiences on some occasion is determined by the way he construes the world or changes in it" (13). Within this context, evaluation depends on what the authors call "the knowledge representation system" (54). Let us focus on three key terms contained in this definition: reaction, value, and interpretation. Emotions are reactions to a situation that result in the individual attributing value to that situation based on their interpretation of it. In other words, emotion is the effect of a

cause, and that cause is both interpreted and evaluated. Interpretation and evaluation, therefore, presuppose causal reasoning (i.e., identification of the cause of an effect).

Now, for Collins et al. (1994), emotions are always determined by our representation of knowledge (i.e., cognitive activities), suggesting a cognitive theory of emotions. For Panksepp, on the other hand, it is necessary to distinguish pure or basic emotions from emotions connected to and transformed by cognitive activities. This point clearly distinguishes Panksepp's approach from that of Collins, Ortony, and Clore. As noted in Section "The primitive affective states", Panksepp criticizes the cognitive interpretation of emotions. He further defends the idea of "basic emotions," which Collins et al. (1994, pp. 26–27) regard as too vague. However, Collins et al. (1994) do not exclude the concept completely: "We claim that some emotions are more basic than others because we can give a very specific meaning to it, namely that some emotions have less complex specifications and eliciting conditions than others" (28).

If we follow Panksepp's logic, we must not only believe (a) in the existence of a pure core of basic emotions distinct from cognition (i.e., biological emotions) but also (b) in the emotional origin of cognitive activities and cognitive-oriented emotions. Panksepp proposes a model based on difference and continuity: the difference between emotion and cognition, but continuity via cognition that arises from emotion.

How can we express this double relation through the do-calculus? There must be a difference between the do-calculus that describes cognitive/cortical processes and the do-calculus that describes affective/subcortical processes. My claim is that while affective processes must be represented as interventions (i.e., the second rung of Pearl's ladder of causation), cognitive processes must be represented as counterfactuals (i.e., third rung). As discussed in the last paragraphs of the previous Section, counterfactual reasoning makes it possible to abstract from the individual causal connection or reaction and generate imaginative variations that allow for more elaborate evaluations and anticipations. Therefore, cognitive processes derived from affective processes use different libraries of diagrams and calculations accumulated over time. They are, by extension, on a different level of abstraction.

Let us now consider the formal structure of the counterfactual as described by Pearl and Mackenzie (2018, p. 278). We can schematically distinguish two stages: (a) transformation of the initial model or diagram through the do-operator (i.e., the imaginative variation) and (b) statistical prediction of the consequences of model transformation and information related to it. Let us observe what happens, for example, in a poem, which can be considered the expression of an emotion.[8] The model or diagram representing the initial causal connection (i.e., the initial emotion) is modified through an imaginative act (i.e., the do-operator). Herein lies the abstraction. The modified model is then used to predict a new emotion: specifically, a reaction in the audience. Thus, the poem aims to build a new causal connection and invites the reader to analyze its possible consequences. It is an example of counterfactual reasoning, which can be translated into formal terms through the do-calculus. This line of reasoning is arguably also compatible with Collins, Ortony, and Clore's appraisal structure model (1994, pp. 50–58). Therefore, Pearl's ladder of causation gives us a unique model for explaining the continuity and difference between emotion and cognitive activity or, perhaps, even consciousness.

What is the unconscious then? The unconscious is repressed. But what is repression here? Memories connected to traumatic emotions are repressed. We can interpret repression in a causal way. In our AGI system the causal diagrams related to traumatic emotions, those that involve an excessive waste of energy, are repressed. Repression, in this case, means that those diagrams (or part of them) are blocked or "bypassed" by other diagrams, even if they remain in the memory of the system. More complex functions such as imagination, language, etc. cannot act on these diagrams. As I said, diagrams related to an emotional system can be arranged on several levels, in a hierarchical manner; then only some levels will be repressed and not others. The system defends itself by blocking or bypassing the diagrams (or part of them) related to traumatic emotions. The "return of the repressed" can be due to a miscalculation, i.e., a problem of de-confounding, or other reasons.

6/4 Reply to Dreyfus' classical argument. In the previous sections, I have shown how Panksepp's topography of emotions can be organized and embedded in AGI. I have (1) discussed the Solms-Friston model and (2) proposed a new model based on Pearl's theory of causation. In this section, I intend to reply to some criticisms of AGI.

Analyzing all the arguments that have been produced against AGI amounts to writing a book and not a paper. Here I want to focus only on Dreyfus' criticisms contained in his important book *What Computers Can't Do* (1972). I will try to briefly reconstruct the structure of Dreyfus' argument on AGI and then formulate some criticisms.

Dreyfus's critique of AGI is inspired by Heidegger's phenomenological research. Dreyfus criticizes what, according to him, are four wrong assumptions of AGI research: (a) the assumption that the brain and mind are analogous to hardware and software; b) the assumption that the mind works computationally; (c) the assumption that all human activities can be formalized and calculated; (d) the assumption that reality consists of a series of facts. Following Heidegger, Dreyfus holds that human existence is a specific being-in-the-world defined by a horizon of possibility (see Coeckelbergh, 2020, p. 47). In *Being and Time*, Heidegger claims that the being-in-the-world is an ontological structure defined by the category of "care," which develops in two directions: the belonging to the world and the relationship with others. Care is above all a set of possibilities whose ultimate horizon is temporality and death. This ontological structure cannot be formalized or reduced to computation because, as Heidegger claims, "science does not think," in the sense that it is capable only of thinking about entities, physical things, not being.

Inspired by Merleau-Ponty's work, Dreyfus emphasizes in particular that our being-in-the-world is based on our body. As embodied, we are part of the social world and have tacit knowledge and skills (dispositions, tendencies) which cannot be formalized or expressed in a language (see Dreyfus, 1972, pp. 24–34). Dreyfus argued that human skills and competence depend mainly on our background sense of "context," that is the ability to identify what is important and interesting in a given situation.

I want to make four criticisms of Dreyfus.

The first concerns emotions. According to Heidegger, one of the central dimensions of the being-in-the-world is emotionality. The human being is always immersed in a certain emotional situation that defines him/her. Now, neuroscience confirms this point but also demonstrates that human emotions (a) have nothing mysterious but can be perfectly explained in physical and computational terms, (b) is very similar to the emotions of all other mammals. Therefore, neuroscience shows that at least a part of the human being-in-the-world can be translated into computational terms. Heidegger's criticism of science is based on his own romantic pre-judgment.

Second criticism: it is not true that AGI is based on the assumption that all reality consists of facts. Dreyfus's arguments cannot be applied to current AI. Precisely the introduction of the

Bayesian Networks with Pearl in the 1990s introduced an elegant way of treating probability (therefore the dimension of the possibility) in AI. So, if Heidegger's being-in-the-world is a horizon of possibilities, then we can formalize at least part of this horizon.

The third criticism of Dreyfus concerns the concept of embodied. Machines also have a body that constitutes a set of tacit knowledge and skills related to the social world. The concept of the design illustrates exactly this point. My computer has a body, a material shape designed by professional designers based on certain social needs (Vial, 2013). My computer is a design object and not a stone or piece of wood found in the woods. Through design, my computer conveys a set of meanings and it is part of the human world. Moreover, the act of design is always a utopian act, which looks to the future of society, which sets a new way of being in the world. As Findeli says, "the end or purpose of design is to improve or at least maintain the 'habitability' of the world in all its dimensions" (2010). Through design, the computer also "possesses" implicit knowledge and skills that are not translated into propositions. For instance, the ability to transmit a vision of the world, society, and the future.

The fourth criticism concerns the distinction between knowing-how and knowing-that and the primacy of intuition. Dreyfus argues that intuition cannot be translated into formal propositions, therefore into programs. Knowing-how cannot be reduced to knowing-that. A possible answer could be that, in reality, the only difference between knowing-how and knowing that is the speed with which the data is processed. Therefore, it is not important whether formal propositions are at the basis of the process or not. Intuition is just a much faster knowledge than the others.

Fjelland (2020) has recently drawn on Dreyfus' criticism of AGI. Moreover, Fjelland uses Pearl's theory of causation in order to claim that AGI is impossible: computers cannot understand causal connections because they do not have a model of reality. As we can read:

> According to Pearl and Mackenzie, the root of the problem is that computers do not have a model of reality. However, the problem is that nobody can have a model of reality. Any model can only depict simplified aspects of reality. The real problem is that computers are not in the world because they are not embodied. (Fjelland, 2020, p. 6)

This argument does not make sense. First of all, this is not the position of Pearl and Mackenzie. The opposite is true: Pearl and Mackenzie show that human understanding of causation can be translated into algorithms and software—it is the outcome of the "causal revolution" that the "cause" ceases to be a vague and imprecise concept and acquires a precise mathematical status. Pearl responds positively to the question "Can we make machines that think?": "I believe that strong AI with causal understanding and agency capabilities is a realizable promise" (Pearl and Mackenzie, 2018, p. 367). Second, it is not true that computers are not in the world. Computers act in the world just like human agents, animals, plants, and the rest of things. They have a physical body just like us. They are social agents exactly like us. The objector might reply that computers do not have semantic intelligence; they do not understand the meaning of what they do. This, however, is an ambiguous answer, for two reasons, one positive and one negative.

First, the notion of *meaning* is ambiguous in itself.

Secondly, the notion of *meaning* can be also understood in an evolutionary sense: as a layering of networks of affects, emotions, memories connected to layered networks of sounds, images, neural connections, etc. In short, what distinguishes us from the machines would not be a "special human quality" that machines do not possess, but the fact that machines are still at the beginning of their evolution. I think that the evolution factor and integration with the surrounding environment are two crucial elements for achieving AGI. However, this is the responsibility of humans, not of the machine. Just as the healthy growth of the child from an emotional point of view is the responsibility of the "good enough" mother (Winnicott, 1988), not of the child. We are used to thinking of AI as if it were an adult human being. In reality, the opposite is true: many AIs behave like children, and the child needs time and an ongoing personal environment to become a person. Without the care of a parent, this development cannot take place.

Pearl defines intelligence as the ability to pass the mini-Turing test by answering questions about causality (Pearl and Mackenzie, 2018, pp. 36–46). For psychoanalysis and affective neuroscience, intelligence is emotional maturity, which is the result of the individual's emotional growth. Emotional maturity is the ability to manage one's emotions, overcoming conflicts with the surrounding environment. Can we build machines that can experience emotional growth? This paper replies positively by indicating the fundamental basis for this undertaking.

6/5 A body for AGI. As I mentioned before, Dreyfus affirms that computers, which do not have a body, a childhood, and a culture, cannot acquire intelligence in the proper sense. Dreyfus (1992) argues that much of human knowledge is tacit and therefore cannot be articulated into a program. The project of a strong AI is therefore impossible.

On this point, however, more advanced research could transform the situation again. I am not talking about biorobotics or biomedical engineering research, but about the creation and development of the first biological robots, made of programmable biological matter. In this regard, Kriegman et al. (2019) open a completely new path. They present the results of research that led to the creation and development of Xenobots, the first tiny robots made entirely of biological tissues. Xenobots are a new life on our planet. Researchers used an evolutionary algorithm to simulate the design of robots. They then selected the best models. These models have been tested to make them increasingly capable of adapting to real situations; the researchers subjected them to large quantities of noise in order to understand if, in a normal situation, they would have maintained the intended behavior or not. The transition from the design to the implementation phase took place through the use of embryonic stem cells of *Xenopus laevis*, a type of frog. The cells were assembled and developed by the computer and then programmed to perform some functions. "Programmed" means that cells were assembled into a finite series of configurations to which certain movements and functions correspond in an aqueous environment. These micro-organisms are neither animals nor traditional robots. They have a heart and skin. If damaged, they can repair themselves and survive for at least ten days. They are assembled by the computer and programmed to behave according to the models. Xenobot is an organism in all respects but based on an artificial design. As it has a body, it has *bodily senses*, and thus it has homeostatic and sensory affects. This solves many of the problems associated with robot embodiment (see Dietrich et al., 2008, p. 150).

From our point of view, the principle of biological programming could be used to program the seven basic affective systems theorized by Panksepp into the cells (or groups of cells). Powerful learning and evolutionary algorithms would allow us to understand how these cells evolve and whether they develop feelings and thoughts like humans or other animals.

6/6 can AI sleep? The AGI model outlined in this paper may seem too ego-centered. In fact, we have not mentioned—except briefly—two key elements of Freudian psychoanalysis: dreaming and repression. So far, we have essentially described a machine with instincts. As I said before, If we follow the reasoning developed in the previous sections of the paper, we must affirm that the repressed is a causal diagram whose access has been blocked by internal or external events, for example an eccessive waste of energy, or the interaction with another diagram. However, the locked diagram keeps acting in the system memory. In this section, I re-interpret the concepts of dreaming and repression starting from a case study.

Can an AI system sleep? Theoretically, machines do not understand things such as sleep or rest. A machine can continue to work continuously without ever having to stop if provided a constant source of energy. However, in June 2020, researchers from the National Laboratory of Los Alamos made an important discovery. They realized that a neural network system for unsupervised learning became more and more unstable if left to work for too long. The solution was to put the system to sleep. "We study spiking neural networks, which are systems that learn much as living brains do," said Los Alamos National Laboratory computer scientist Yijing Watkins. "We were fascinated by the prospect of training a neuromorphic processor in a manner analogous to how humans and other biological systems learn from their environment during childhood development."[9] Watkins and her research team found that neural network learning to see became unstable after continuous and intense periods of unsupervised learning. Faced with this difficulty, they decided to put the system to sleep: "When they exposed the networks to states that are analogous to the waves that living brains experience during sleep, stability was restored," explains the note from the laboratory. "It was as though we were giving the neural networks the equivalent of a good night's rest," said Watkins.

The researchers used spiking neural networks (SNNs), which are computational models that mimic biological neural networks. "Compared with artificial neural networks (ANN), SNNs incorporate integrate-and-fire dynamics that increase both algorithmic and computational complexity" (Watkins et al., 2020). Neuromorphic processors have tested that try to simulate the behavior of the brain and the human nervous system. These processors are made of special materials that are able to best reproduce the plasticity of the human brain. Deep neural network software is then run in these processors.

> The discovery came about as the research team worked to develop neural networks that closely approximate how humans and other biological systems learn to see. The group initially struggled with stabilizing simulated neural networks undergoing unsupervised dictionary training, which involves classifying objects without having prior examples to compare them to. "The issue of how to keep learning systems from becoming unstable really only arises when attempting to utilize biologically realistic, spiking neuromorphic processors or when trying to understand biology itself," said Los Alamos computer scientist and study coauthor Garrett Kenyon. "The vast majority of machine learning, deep learning, and AI researchers never encounter this issue because in the very artificial systems they study they have the luxury of performing global mathematical operations that have the effect of regulating the overall dynamical gain of the system."[10]

As stated earlier, the researchers solved the instability problem by making the system "sleep." They did that *by introducing noise*. The machine "sleeps" and, thanks to this "sleep," manages to regain equilibrium, exactly as the human body does.

> The researchers characterize the decision to expose the networks to an artificial analog of sleep as nearly a last ditch effort to stabilize them. They experimented with various types of noise, roughly comparable to the static you might encounter between stations while tuning a radio. The best results came when they used waves of so-called Gaussian noise, which includes a wide range of frequencies and amplitudes. They hypothesize that the noise mimics the input received by biological neurons during slow wave sleep. The results suggest that slow-wave sleep may act, in part, to ensure that cortical neurons maintain their stability and do not hallucinate. The groups' next goal is to implement their algorithm on Intel's Loihi neuromorphic chip. They hope allowing Loihi to sleep from time to time will enable it to stably process information from a silicon retina camera in real time. If the findings confirm the need for sleep in artificial brains, we can probably expect the same to be true of androids and other intelligent machines that may come about in the future.[11]

This experiment can benefit from the integration of the results of Hobson and Friston's research on dreams. Hobson and Friston (2012) demonstrate the essential function of the dream in relation to the free-energy principle. Sleep implies optimization processes that are perfectly consistent with the free energy principle. In particular, Hobson and Friston emphasize the connection between homeothermy, sleep, and consciousness. Sleep is connected with homeostatic processes, especially temperature control, which are necessary for consciousness. In particular, Hobsn and Friston hold a conception of the dreaming brain as a simulation machine or a virtual reality generator that seeks to optimally model and predict its waking environment and that needs REM sleep processes (particularly PGO waves) to do so. The basic idea is that the brain comes genetically equipped with a neuronal system that generates a virtual model of the world during REM sleep because REM sleep processes are essential to optimize this generative model. In other words, as the brain is a virtual reality machine or prediction error device (as we saw above) in order to minimize free energy, sleep is a particular way to achieve this goal. From this point of view, the experiment of the National Laboratory of Los Alamos is very interesting because it shows a profound analogy in the functioning of the brain and a deep neuronal network: both need an "off-line" phase in order to ensure the equilibrium of the system. This also confirms what has been said above: the free energy principle is a useful model to describe and explain above all homeostatic processes, but not emotions. Homeostatic imbalances can activate or influence emotional systems. However, homeostatic processes remain something different from emotions. Furthermore, the homeostatic processes active during sleep cannot fully explain the emotions experienced during the sleep or the contents of the dream itself.

What conclusions can we draw from these considerations? I have identified two. The first is that an advanced AGI system presents much more complex behavior than expected and, therefore, requires cycles of activity and rest. The other is that in an advanced AI system, the simulation of human cognitive activities (language, logic, memory, learning, etc.)—what we would call "secondary processes" in Freudian terms—requires the simulation of "primary processes" (sleep is only one example; we could also mention instincts or emotions) as well. Here, I want to avoid confusing sleep and dreaming; obviously, I am not implying that an AI system can dream. The point is that a cortical AI

system needs subcortical AI. In the case we examined, it is the machine that experiences this need.

However, what exactly is "sleeping" for an AI or AGI system? The essence of Freud's theory is that in sleep—and, in particular, in that phase of sleep in which dreams occur—a regression takes place; the ego (the center of cognitive activities) is inhibited and the id (the unconscious, the set of drives) takes over (see *The Interpretation of Dreams*, Chapter 7). Sleep is then a fundamental observatory in which primitive drive states are more evident. Can we apply the Freudian notion of regression to AGI? My hypothesis is that there exist forms of regressions in information as well. In AGI, the regression goes *from information to noise*. The regression to noise is essential to information. In every information process, there are forms of regression to different types of noise.

Freud distinguished three types of regression: (a) topical—from one psychic system to another; (b) temporal—the regression toward older psychic formations; and (c) formal—the return of primitive modes of expression and representation. I claim the same about information. The regression to noise can be of three types: (a) topical—toward information of different types (data that must be coded in another way); (b) temporal—that is, toward more ancient information; and (c) formal—toward data without configuration (pure noise). The Los Alamos Lab experiment proves exactly that information needs regressions to noise, to forms of stabilization and iteration.

Returning to our AGI model, we can connect the regression from information to noise to the need to maintain homeostasis. Noise is all that threatens homeostasis and increases entropy; on the contrary, information ensures homeostasis and decreases entropy. Noise is the set of data that threaten homeostasis and are therefore split from information while remaining in the system's memory. This is a very important point: the unconscious is not a part of the psyche, but a quality of the processes of the psyche. Therefore, all internal affective processes that cause an increase of the free energy must be repressed; they remain in the state of a pre-mental, pre-cognitive consciousness—they remain raw data or noise. The distinction between information and noise arises from the general need of the system to keep the level of free energy as low as possible. This distinction corresponds to Freudian repression. As I said, the Solms-Friston model is not at odds with Pearl's causation theory; they can be held together as complementary, one to represent homeostatic affects while the other to represent emotional affects in Panksepp's terminology.

Conclusions

I consider the theses developed in this paper to be the beginning of a research program on the possibility of an AGI based on the simulation of the subcortical areas of the brain. Only future investigations will be able to establish the merits and demerits of the ideas developed here. The central theoretical hypothesis of this paper is that AGI is possible only if the main cognitive functions are based on computational systems capable of adequately simulating the raw affective systems that humans share with other mammals. AGI must not be based on the imitation of the behavior of humans, but on the modelization of their seven basic affective systems described by Panksepp within a psychoanalytic framework. The purpose of this paper was to show how to organize and embed the seven emotional systems defined by Panksepp into a computational system. With this in mind, I also analyzed and criticized the position of Dreyfus, who launched a famous critique of the AGI project from a Heideggerian point of view.

In conclusion, I would like to formulate a last thesis: developing an AGI based on the fundamental human emotional systems, i.e., capable of producing its own emotional life similar to the human one, is the best way to solve the problem of AI control.

The problem of AI control can be formulated as follows: "If we build machines to optimize objectives, *the objectives we put into the machines* have to match *what we want*, but *we do not know how to define human objectives completely and correctly* (Russell, 2019, p. 170; my emphasis). The problem of AI control is "to design machines with a high degree of intelligence—so that they can help us with difficult problems—while ensuring that those machines never behave in ways that make us seriously unhappy" (Russell, 2019, p. 171). The future of humanity is tied to the future of AI and how humans will be able to integrate AI systems into their world (Elliott, 2018). This is the reason why it is so essential to develop machines that are able not only to know human desires and needs but also to understand them, interpret their changes and share them. As Russell (2019, p. 11) says: "Machines are beneficial to the extent that their actions can be expected to achieve our objectives." Only in this way, humans will avoid becoming the second intelligent species on the planet.

The objectives we put into the machines: this is exactly the problem. Humans want the machine to do what they want, "but *we do not know how to define human objectives completely and correctly*" and we often act in ways that are contrary to our own preferences. What are the human goals with respect to AI? How can we clarify them? What do we want from machines? Does an AI need to be able to recognize human unconscious dynamics so that it can always act for the best of humans—that best that not even humans often know? Emotional neuroscience can give an answer by identifying the DNA of human needs, objectives, and thoughts. Panksepp's emotional systems are the fundamental schemes of action and learning, the foundation of our whole being.

Now, a subcortical AGI, like the one we have described in this paper, would solve the problem of AI control. An AGI system based on the seven systems of Panksepp would share with the human beings the fundamental schemes of action and learning, and therefore the essential needs and desires. This would be possible *without sacrificing the computational power of AGI*. Nor would it be necessary to introduce increasingly complex sets of rules in the system.

What would happen if the AGI system developed wrong emotions and behaviors? It is a legitimate question. Here an educational problem arises. If we want to create super-intelligent systems capable of understanding and supporting our objectives, and also sharing their emotions with us, we must be able to educate them, to follow them in their growth, as if they were children. For this reason, an AI psychoanalysis, that is, psychoanalysis of AGI systems could play a key role in the future of humanity.

Data availability

All data generated or analyzed during this study are included in this article.

Received: 14 July 2020; Accepted: 18 May 2021;
Published online: 31 May 2021

Notes

1 This paper is a development and a creative transformation of Possati (2021, chapters 4 and 5).

2 "Formally, a Markov blanket renders a set of states, internal and external states, conditionally independent of one another. That is, for any variable A, A is conditionally independent of B, given another variable, C, if and only if the

probability of A and B were given C can be written as p(AjC) and p(BjC). In other words, A is conditionally independent of B given C if, when C is known, knowing A provides no further information about B. […] The cell is an intuitive example of a living system with a Markov blanket. Without possessing a Markov blanket a cell would no longer be, as there would be no way by which to distinguish it from everything else" (Kirchhoff et al., 2018, p. 3).

3 "Hidden causes are called hidden because they can only be 'seen' indirectly by internal states through the Markov blanket via sensory states. As an example, consider that the most well-known method by which spiders catch prey is via their self-woven, carefully placed and sticky web. Common for web- or niche-constructing spiders is that they are highly vibration sensitive. If we associate vibrations with sensory observations, then it is only in an indirect sense that one can meaningfully say that spiders have 'access' to the hidden causes of their sensory world—i.e., to the world of flies and other edible 'critters'" (Kirchhoff et al., 2018, p. 4).

4 An objection could be advanced here: Does not reduce affective systems to causal probabilistic models imply a cognitive interpretation of affects, as in Picard? This is an important objection. However, I believe that the reference to Pearl's work gives us a way to respond to it. Pearl's causal models are not mere sets of data computations but involve an interpretation of data that is based on concrete experience—the intuition of causality and apprehension of reality that precedes causality and any other fact or concept.

5 In more technical terms, Pearl defines confounding through two interrelated notions: incomparability and a lurking third variable (see Pearl and Mackenzie 2018, p. 151). This passage is essential: "[…] *the noncausal paths are precisely the source of confounding*. Remember that I define confounding as anything that makes P(Y | do (X)) differ from P(Y | X). The do-operator erases all the arrows that come into X, and in this way, it prevents any information about X from flowing in the noncausal direction. Randomization has the same effect. So does statistical adjustment, if we pick the right variables to adjust" (Pearl and Mackenzie, 2018, p. 157; emphasis added).

6 For an elementary introduction to Pearl's theory of causality, see Pearl et al., 2016. An important source of information is also: http://bayes.cs.ucla.edu/jp_home.html. See also Tucci, 2013.

7 It is also important to read the rest of the passage: "First, let us rephrase the task of finding the effect of X on Y using the language of proofs, axioms, and auxiliary constructions, the language of Euclid and Pythagoras. We start with our target sentence, P(Y | do(X)). Our task will be complete if we can succeed in eliminating the do-operator from it, leaving only classical probability expressions, like P(Y | X) or P (Y | X, Z, W). We cannot, of course, manipulate our target expression at will; the operations must conform to what do(X) means as a physical intervention. Thus, we must pass the expression through a sequence of legitimate manipulations, each licensed by the axioms and the assumptions of our model. The manipulations should preserve the meaning of the manipulated expression, only changing the format it is written in" (Pearl and Mackenzie 2018, p. 233).

8 The nature of poetry and the role of affect in it has been a hotly debated topic throughout literary history. To give one example, Stephen Halliwell (2017) discusses the ancient Greco-Roman origins of this debate, which is also the origin of the modern philosophical tradition. He also touches on more modern branches of this debate, including the Romantics' definition of poetry as an expression of emotion and the Modernists' rejection of affective poetry.

9 lanl.gov/discover/news-release-archive/2020/June/0608-artificial-brains.pdf

10 lanl.gov/discover/news-release-archive/2020/June/0608-artificial-brains.pdf

11 lanl.gov/discover/news-release-archive/2020/June/0608-artificial-brains.pdf

References

Alberini C (2010) Long-term memories: the good, the bad, and the ugly. Cerebrum 2010:21, http://dana.org/news/cerebrum/detail.aspx?id=29272

Amoore L (2009) Algorithmic war: everyday geographies of the war on terror. Antipode 41:49–69

Apaydin E (2016) Machine learning. The new AI. MIT Press

Baldwin R (2016) The great convergence: information technology and the new globalization. Harvard University Press

Benedetti F (2010) The patient's brain. Oxford University Press

Blass R, Carmeli Z (2007) The case against neuropsychoanalysis: on fallacies underlying psychoanalysis' latest scientific trend and its negative impact on psychoanalytic discourse. Int J Psychoanal 88:19–40

Bolter D (1986) Turing's man. Western culture in the computer age. Penguin Books, London

Bostrom N (2016) Superintelligence: paths, dangers, strategies. Oxford University Press

Bruineberg J, Dewhurst J, Dolega K, Baltieri M (2020) The Emperor's new Markov blankets. http://philsci-archive.pitt.edu/18467/

Coeckelberg M (2020) Introduction to the philosophy of technology. Oxford University Press

Collins G, Ortony A, Clore A (1994) The cognitive structures of the emotions. Cambridge University Press

Colvin G (2015) Humans are underrated: what high achievers know that brilliant machines never will. Penguin, New York

Damasio A (1994) Descartes' error. Putnam, New York

Damasio A (1999) The strange order of things. Pantheon, New York

Damasio A (2003) Looking for Spinoza. Heinemann, London

Damasio A (2010) Self comes to mind: constructing the conscious brain. Random House, New York

Davis K, Montag CH (2019) Selected principles of panksepian affective neuroscience. Front Neurosci 12:1025

Decety J, Ickes WJ (2009) The social neuroscience of empathy. MIT Press

Dietrich D, Fodor G, Kastner W, Ulieru M (2007) Considering a technical realization of a neuro-psychoanalytical model of the mind - A theoretical framework. 5th IEEE International Conference on Industrial Informatics. https://doi.org/10.1109/INDIN.2007.4384954

Dietrich D, Fodor G, Zucker G, Bruckner D (eds) (2008) Simulating the mind: a technical neuropsychoanalytical approach. Springer, Berlin

Dreyfus HL (1972) What computers can't do. Harper & Row, New York

Dreyfus HL (1992) What computers still can't do. MIT Press

Dreyfus HL, Dreyfus SE (1986) Mind over machine. Basil Blackwell, Oxford

Dyson G (2012) Turing's cathedral. Random House, New York

Edelson M (1986) The convergence of psychoanalysis and neuroscience: illusion and reality. Contemp Psychoanal 22:479–519

El-Nasr MS, Yen J, Ioerger TR (2000) FLAME: fuzzy logic adaptive model of emotions. Autonom Agent Multi-Agents Syst 3:219–257

Elliott A (2018) AI culture: everyday life and the digital revolution. Routledge, London-New York

Erol B, Majumdar A, Benavidez P, Rad P, Choo KR, Jamshidi M (2019) Toward artificial emotional intelligence for cooperative social human–machine interaction. IEEE Trans Computat Soc Syst 7(1):234–246

Findeli A (2010) Searching for design research questions: some conceptual clarifications. In:Chow R, Jonas W, Joost G (eds) Questions, hypotheses, and conjectures: discussions on projects by early stage and senior design researchers. IUniverse, Bloomington, pp. 34–48

Fjelland R (2020) Why general artificial intelligence will not be realized. Humanit Soc Sci Commun 7:1–9

Fogel A, Kvedar J (2018) Artificial intelligence powers digital medicine. Digital Med 1(5):23–45

Friston K (2009) Causal modelling and brain connectivity in functional magnetic resonance imaging. PLoS Biol 7(2):220–225

Friston K (2013) Life as we know it. J R Soc Interface 10:20130475

Gallese V (2009) The two sides of mimesis: Girard's mimetic theory, embodied simulation and social identification. J Conscious Stud 16(4):21–44

Halliwell S (2017) The poetics of emotional expression. Steiner, Stuttgart

Hitchcock C (2018) Probabilistic Causation. Stanford Encyclopedia of Philosophy

Hobson JA (2007) Wake up or dream on? Six questions for Turnbull and Solms. Cortex 43:1113–1115

Hobson JA, Friston K (2012) Waking and dreaming consciousness: neurobiological and functional considerations. Prog Neurobiol 98(1):82–98

Johnson M, Horn G (1986) Dissociation of recognition memory and associative learning by a restricted lesion of the chick forebrain. Neuropsychologia 24:329–340

Johnson M, Horn G (1988) Development of filial preferences in dark-reared chicks. Anim Behav 36:675–683

Kandel ER (1979) Psychotherapy and the single synapse. New Engl J Med 301 (19):1028–1037

Kandel ER (1983) From metapsychology to molecular biology: explorations into the nature of anxiety. Am J Psychiatry 140(10):1277–1293

Kahneman D (2011) Thinking fast and slow. Penguin Books, New York

Kaplan K, Solms M (2000) Clinical studies in neuro-psychanalysis. International Universities Press, Madison

Kirchhoff M, Parr T, Ensor P, Friston K, Kiverstein J (2018) The Markov blankets of life: autonomy, active inference, and the free energy principle J R Soc Interface 15:20170792

Kriegman S, Blackiston D, Levin M, Bongard J (2019) A scalable pipeline for designing reconfigurable organisms Proc Natl Acad Sci USA 117 (4):1853–1859

Le Cun Y (2019) Quand la machine apprend. La revolution des neurons artificiels et de l'apprentissage profond. Odile Jacob, Paris

LeDoux J (1996) The emotional brain. Simon & Schuster, New York

Lewis D (1973) Counterfactuals. Wiley&Sons, New York

Luria AR (1976) The working brain. Basic Books, New York

Montag C, Widenhorn-Müller K, Panksepp J, Kiefer M (2017) Individual differences in Affective Neuroscience Personality Scale (ANPS) primary emotional traits and depressive tendencies. Comp Psychiatry 73:136–142

Panksepp J (1982) Toward a general psychobiological theory of emotions. Behav Brain Sci 5:407–467

Panksepp J (1998) Affective neuroscience: the foundations of human and animal emotions. Oxford University Press

Panksepp J (2008) Simulating the primal affective mentalities of the mammalian brain: a fugue on the emotional feelings of mental life and implications for AI-Robotics. In: Dietrich D, Fodor G, Zucker G, Bruckner D (eds) Simulating the mind: a technical neuropsychoanalytical approach. Springer, Berlin
Panksepp J, Biven L (2012) The archeology of mind: neuroevolutionary origins of human emotions. W. W. Norton, New York
Pearl J, Glymour M, Jewell PR (2016) Causal inference in statistics. Wiley, New York
Pearl J, Mackenzie D (2018) The book of why: the new science of cause and effect. Random, New York
Penrose R (1989) The Emperor's new mind: concerning computers, minds, and the laws of physics. Oxford University Press
Penrose R (1994) Shadows of the mind: a search for the missing science of consciousness. Oxford University Press
Picard R (1997) Affective computing. MIT Press
Possati LM (2021) The algorithmic unconscious. how psychoanalysis helps in understanding AI. Routledge, London
Prescott T J, Lepora N (2018) Living machines: a handbook of research in biomimetics and biohybrid systems. Oxford University Press
Pulver SE (2003) On the astonishing clinical irrelevance of neuroscience. J Am Psychoanal Assoc 51:755–772
Rolls ET (1999) The brain and emotion. Oxford University Press
Rolls ET (2005) Emotion explained. Oxford University Press
Russell S (2019) Human compatible. AI and the problem of control. Random, New York
Russell S, Norvig P (2016) Artificial intelligence: a modern approach. Pearson, London
Shanahan M (2015) The technological singularity. MIT Press
Shibata T, Yoshida M, Yamato J (1997) Artificial emotional creature for human-machine interaction. In: 1997 IEEE International Conference on Systems, Man, and Cybernetics. Computational Cybernetics and Simulation, Orlando, pp. 2269–2274
Schuller D, Schuller BW (2018) The age of artificial emotional intelligence. Computer 51(9):38–46
Solms M (1996) Towards an anatomy of the unconscious. J Clin Psychoanal 5 (3):331–367
Solms M (2000) Freud, Luria and the clinical method. Psychoanal History 2:76–109
Solms M (2008) Repression: a neuropsychoanalytic hypothesis. www.veoh.com/watch/v6319112tnjW7EJH
Solms M (2013) The conscious Id. Neuropsychoanalysis 15(1):5–19
Solms M, Saling M (eds) (1990) A moment of transition. Two neuroscientific articles by Sigmund Freud. The Institute of Psychoanalysis, London
Solms M, Friston K (2018) How and why consciousness arises: some considerations from physics and physiology. J Conscious Stud 25(5–6):202–238
Solms M, Turnbull O (2002) The brain and the inner world: an introduction to the neuroscience of subjective experience. Other Pr. Llc
Sulloway F (1979) Freud: biologist of the mind. Harvard University Press
Tucci R (2013) Introduction to Judea Pearl's Do-Calculus. https://www.researchgate.net/publication/236887179_Introduction_to_Judea_Pearl's_Do-Calculus
Vial S (2013) L'être et l'écran. Puf, Paris
Yonck R (2017) Hearth of the machine: our future in a world of artificial emotional intelligence. Arcade, New York
Yovell Y, Bar G, Mashiah M, Baruch Y, Briskman I, Asherov J (2016) Ultra-low-dose buprenorphine as a time-limited treatment for severe suicidal ideation: a randomized controlled trial. Am J Psychiatry 173:491–498
Watkins Y, Kim E, Sornborger A, Kenyon GT (2020) Using Sinusoidally-Modulated Noise as a Surrogate for Slow-Wave Sleep to Accomplish Stable Unsupervised Dictionary Learning in a Spike- Based Sparse Coding Model. Working paper, Computer Vision Foundation. https://openaccess.thecvf.com/content_CVPRW_2020/papers/Watkins_Using_Sinusoidally-Modulated_Noise_as_a_Surrogate_for_Slow-Wave_Sleep_to_CVPRW_2020_paper.pdf
Winnicott D (1988) Human nature. The Winnicott Trust

Competing interests

The author declares no competing interests.

Additional information

Correspondence and requests for materials should be addressed to L.M.P.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.