MENTAL PERTURBANCE: AN INTEGRATIVE DESIGN-ORIENTED CONCEPT FOR UNDERSTANDING REPETITIVE THOUGHT, EMOTIONS AND RELATED PHENOMENA INVOLVING A LOSS OF CONTROL OF EXECUTIVE FUNCTIONS

LUC BEAUDOIN

Simon Fraser University

MONIKA PUDŁO

University of Warsaw

SYLWIA HYNIEWSKA

University College London

Abstract

Understanding intrusive mentation, rumination, obsession, and worry, known also as "repetitive thought" (RT), is important for understanding cognitive and affective processes in general. RT is of transdiagnostic significance—for example obsessive-compulsive disorder, insomnia and addictions involve counterproductive RT. It is also a key but under-acknowledged feature of emotional episodes. We argue that RT cannot be understood in isolation but must rather be considered within models of whole minds and for this purpose we suggest an integrative design-oriented (IDO) approach. This approach involves the design stance of theoretical Artificial Intelligence (the central discipline of cognitive science), augmented by systematic conceptual analysis, aimed at explaining how autonomous agency is possible. This requires developing, exploring and implementing cognitive-affective-conative information-processing architectures. Empirical research on RT and emotions needs to be driven by such theories, and theorizing about RT needs to consider such data. Mental perturbance is an IDO concept that, we argue, can help characterize, explain, and theoretically ground the concept of RT. Briefly, perturbance is a mental state in which motivators tend to disrupt, or otherwise influence, executive processes even if reflective processes were to try to prevent or minimize the motivators' influence. We draw attention to an IDO architecture of mind, H-CogAff, to illustrate the IDO approach to perturbance. We claim, further, that the intrusive mentation of some affective states— including

grief and limerence (the attraction phase of romantic love) — should be conceptualized in terms of perturbance and the IDO architectures that support perturbance. We call for new taxonomies of RT and emotion in terms of IDO architectures such as H-CogAff. We point to areas of research in psychology that would benefit from the concept of perturbance.

Keywords: repetitive thought, emotions, executive functions, cognitive architectures, autonomous agents, affective computing

Mental Perturbance: An Integrative Design-Oriented Concept for Understanding Repetitive Thought, Emotions and Related Phenomena Involving a Loss of Control of Executive Functions

> I hope that moving toward a general theory of motivation will help psychology as a whole acknowledge and embrace the fundamental importance of motivation in the grand scheme of integrative psychological theory. (Baumeister, 2015, p. 9)

This paper discusses an important type of human mental state, dubbed *perturbance* (Beaudoin, 1994), which is defined in *integrative design-oriented* (IDO) terms. Perturbance is a mental state in which *insistent* motivators or alarms distract or otherwise influence executive processes in a manner that is difficult for reflective processes to suppress or control. The concept of perturbance provides a rich, design-oriented way of understanding some of the attentional aspects of emotion-like states, wherein an autonomous agent, with a certain type of computational architecture, is subject to loss of control of its deliberative processes. We claim that the concept of perturbance can theoretically unify many important mental phenomena that are characterized by repetitive thought (RT), such as worry (Watkins, 2008), obsessions (Macatee et al., 2015), and emotional episodes involving intrusive mentation (Sloman, 1987). This paper also claims and subsequently illustrates our claim that the IDO approach can shed light on multiple phenomena, which is indeed that necessary for a complete understanding of the minds of autonomous agents, be they natural or artificial.

Sloman and Croucher (1981a, 1981b) claimed that future robots will exhibit human-like emotional mentation not because emotional mechanisms were explicitly implemented in them, but as a necessary emergent *biproduct* of interacting information-processing mechanisms that are designed to meet requirements that would later be referred to as requirements of autonomous agents (Beaudoin & Sloman, 1993; Beaudoin, 1994; Thórisson & Helgasson, 2012). The Cognition and Affect (CogAff) project was launched in 1991 to better understand the requirements of autonomous agents, and the space of real and possible minds that meet, or would meet, these requirements. See Sloman (2008a) for a review. This paper builds on that project, extending and adapting its methodology and theory.

The concept of perturbance does not stand alone. It is grounded in the specification of information-processing architectures resulting from an IDO approach to understanding possible and actual minds. This means that one cannot specify the concept of perturbance, or adequately study it empirically, without familiarity with IDO. This approach, as we shall see, contrasts with what Watkins (2008) claims is the scientific allure of the concept of repetitive thought, namely that it is an *atheoretical* concept. Physicists acknowledge that even empirical constructs are deeply theoretical (Lakatos, 1980) —even speed is a theoretical concept specified in relation to other concepts. The theoretical richness of the concept of perturbance, the difficulty of the IDO methodology, and the fact that few researchers pursue the IDO approach might explain why the concept of perturbance has largely been overlooked in psychology.

One of the main objectives of this paper is to whet its readers' appetite for the IDO approach by making them curious about its potential for unifying many literatures with the concept of perturbance and the theory on which it depends. However, the ambitiousness of the IDO approach also presents the chief difficulty of this paper: to concisely explain a complex, old (in relation to the history of computational psychology) yet still nascent, computational research program.

Accordingly, we begin by describing the IDO approach and a class of IDO agent architectures (H-CogAff) that were developed with the aim of supporting the requirements of autonomous agency. We then summarize an argument according to which certain classes of agents, natural or robotic, will *necessarily* be subject to perturbance as an *emergent* phenomenon. We then describe two major classes of 'emotional' phenomena that may be understood as involving perturbance, namely grief and limerence. We then do a quick survey of other psychological phenomena which, we argue, need to be understood in terms of perturbance.

The integrative design-oriented (IDO) Research approach

The IDO approach recognizes Artificial General Intelligence as the general science of intelligence, proceeding primarily from the 'designer stance' (McCarthy, 2008; Sloman, 2008b; Sloman, 1993). From the *designer stance*, one seeks to understand the environmental niches in which the systems one seeks to explain will operate. One specifies the requirements said systems will satisfy. Then, one explores a set of possible designs that are intended to satisfy the requirements. One then seeks to implement the designs in working systems (simulated and real environments), minding the possibility of different implementations. The result of each stage should be analysed in relation to the previous stage, such as the extent to which the implementations matches the design. The entire procedure is iterative. The designer stance is more concerned with the specification and explanation of competence than with prediction. One should resist the urge to jump prematurely to predictions. IDO theories can be more or less agentic, i.e., deal more or less specifically with the requirements of autonomous agency. For instance, the theory presented here is quite agentic. The somnolent information processing theory (Beaudoin, 2014c; Beaudoin et al. 2019; Lemyre, Belzile, Landry, Bastien, & Beaudoin, 2020), while addressing the sleep onset control system in an IDO manner is less agentic: it deals with specific functions which, while grounded in a broader, agentic IDO theory, are essential to autonomous agency (adaptively controlling the onset of sleep).

The definition and requirements specification of autonomous agency are themselves theoretical. Following Sloman and Croucher (1981a, b), Beaudoin (1994) and Hawes (2011), we posit that autonomous agents have multiple top level complex motives; they operate under real-time and (physical and processing) resource constraints in a rapidly changing and partially unpredictable world that they cannot fully control, and which is not necessarily friendly to their motives. They can generate their own top-level and derivative motives, and are capable of pursuing them. From these abstract specifications of autonomous agency many implications

follow, such as limited parallelism of high level 'management' functions (Beaudoin,1994; Simon, 1967) and the possibility of perturbance, the main topic of this paper.

IDO theories are integrative in two main ways. First, fundamental IDO theories must specify a broad collection of information processing functions, towards the design of relatively complete agents. This means that the theories will specify many 'cognitive', 'conative' (motivational), 'affective', 'executive' and ancillary functions. Whereas it is often assumed that there is a sharp boundary between cognitive and affective functions, which at most *interact*, in IDO systems mechanisms can be both cognitive and affective (Beaudoin, 1994, 2014a; Pessoa, 2008, 2013; Sloman & Croucher, 1981; Sloman, 1989; Todd, 2020). It is noteworthy that recent arguments in favor of modularity of vision (Firestone & Scholl, 2016) is based largely on criticisms of 'top down' theories of perception and criticizing empirical paradigms that were purported to produce illusions, biases and errors that do not replicate. We would agree with those criticisms. However, a third design class of designs (apart from sharp 'top down' vs. 'bottom up' modular designs) is possible: Sloman (1989) argues that perception is not modular but labyrinthine, with many inputs and outputs. Beaudoin (1994) discusses several types of *valenced* perception and knowledge, including the perception of threats and opportunities as such. The perception and computation of valence may be blended.

This type of integration, which we call *functional* integration, typically calls for information processing (computational) architectures. The expression 'computational architecture' seems to have been introduced in the Artificial Intelligence (AI) literature by Sloman (1978). The computational architectures proposed in cognitive science are typically *cognitive* architectures (Cooper, 2007; Newell, 1990; Rosenbloom, Demski, & Ustun, 2016), which are not concerned with the requirements of autonomous agency. For example, they do not necessarily deal with affective considerations and multiple sources of motivation with real-time constraints. While purely cognitive architectures are not truly IDO models, they are an important starting point in understanding computational architectures, particularly since their simpler requirements facilitate computational implementation and analysis. Below we briefly present H-CogAff, which is an IDO architecture developed by Sloman and colleagues (Sloman, 2003, 2011).

Secondly, IDO theories will typically be integrative in the more traditional sense that they combine multiple theories. Moors (2017) presents such an integrative theory, which combines theories and proposes a simple architecture. Not fully an IDO theory as it is exclusively developed from an empirical perspective rather than from the design-stance, it is nevertheless relevant to agentic IDO research.

In the IDO approach one aims to understand real and possible minds in an authentically interdisciplinary manner. This involves the disciplines traditionally associated with cognitive science (computer science and AI, philosophy, psychology, neuroscience, biology, linguistics, anthropology and education). The IDO approach aligns with the grand program of cognitive science as "the interdisciplinary study of mind, informed by theoretical concepts drawn from computer science and control theory" (Boden, 2008, p. 12).

It is important to emphasize a particularly important set of techniques drawn from philosophy, namely conceptual analysis, that aim to make explicit and exploit the rich knowledge built into human language and conception. Conceptual analysis is not to be confused with the factor analytic approach (Osgood, Suci, & Tannenbaum, 1957) which is central to many empirical theories of affect, such as the component process model (Fontaine, Scherer, & Soriano, 2013) and core affect psychological construction theory (Russell, 2003). Those theories capture some of the actual usage of terms, i.e., the *logical geography* of conceptual space, whereas conceptual analysis may go beyond actual usage to explore the space of possible concepts, i.e., *logical topology* (Sloman, 2010). As Ortony, Clore and Foss (1987) suggest, conceptual analysis should be done before factor analysis is performed; but it often is not; and in fact, conceptual analysis is not traditionally taught in education or psychology programs. Albert Einstein used conceptual analysis of space and time in developing the theory of relativity (Disalle, 2006). We claim he could not have produced his theory based on factor analysis. Several authors have articulated the need for conceptual analysis in understanding actual and possible minds, and provided tips for this process (Sloman, 1978; Ortony, Clore, & Foss, 1987; Beaudoin, 1994, 2014). A conceptual analysis of motivators and goals presented by Beaudoin (1994) underpins our theory of autonomous agency and perturbance. That analysis, which to our knowledge is the most detailed theoretical specification of goals and motives, illustrates that the lines between engineering, philosophy and science are blurred — the conception of goals presented there includes insights from all three approaches. For other specifications of the concept of goal, see Boden (1978), Moors & Fischer (2018), Pervin (1989), Higgins (2011), Huang & Bargh (2015), and Toomey (1992).

The IDO approach ultimately requires specifying its models in terms of virtual machinery (Sloman, 2002). However, this paper does not delve into that aspect of mind. Readers who do not understand or are not convinced by the relevance of a design-oriented approach to understanding real and possible minds might not be sufficiently illuminated by the brief defense of this approach that this paper provide. We would like at least to single out one of the purposes of this approach, which is also an argument for the pertinence of AI to psychology, namely that "The problem is not that we do not know which theory is correct, but rather that we cannot construct any theory at all which explains the basic facts" (Power, 1979 pp. 109). For instance, one can select a random theory of emotion and ask oneself: can this theory be used as a design for a working system that explains behavior? If the answer is 'no', then the theory is incomplete or incorrect. To answer the essential question requires taking the design stance. One of the earliest and still pertinent books on the relevance of AI to explaining autonomous agency is Boden (1978). The approach is also helpfully explained and justified in Boden (1987, 1988, 1989, 2006), Dennett (1994), Marcus & Davis (2019), Minksy (1985) and Sloman (1978, 1993).

H-CogAff: An Autonomous Agent Architecture

The concept of perturbance emanated from a design-oriented research program that proposed a class of mental architectures (CogAff schema) whose subclass, H-CogAff, is the

backdrop of this paper (Sloman, 2008a). H-CogAff is designed to meet the human autonomous agency requirements as specified above. In Figure 2, the CogAff schema is depicted based on (Sloman, 2008a).

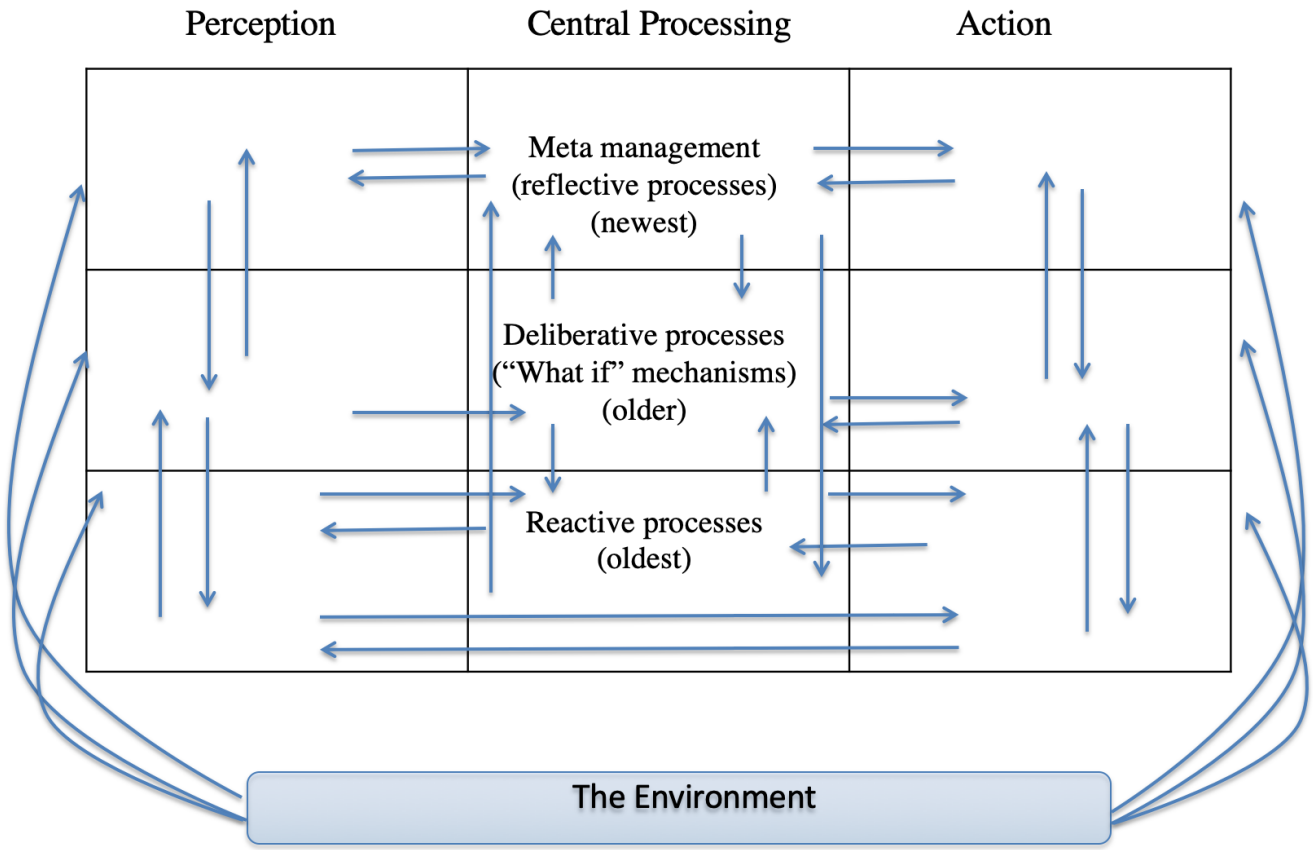

Figure 1. CogAff schema adapted from Sloman (2008a).

In Figure 2, a sketch of H-CogAff is presented, again based on Sloman (2008a). The middle layer in this diagram is dubbed 'management processes', in line with Beaudoin (1994) and Wright, Sloman & Beaudoin (1996), and its functions are slightly generalized compared to Sloman (2008a).

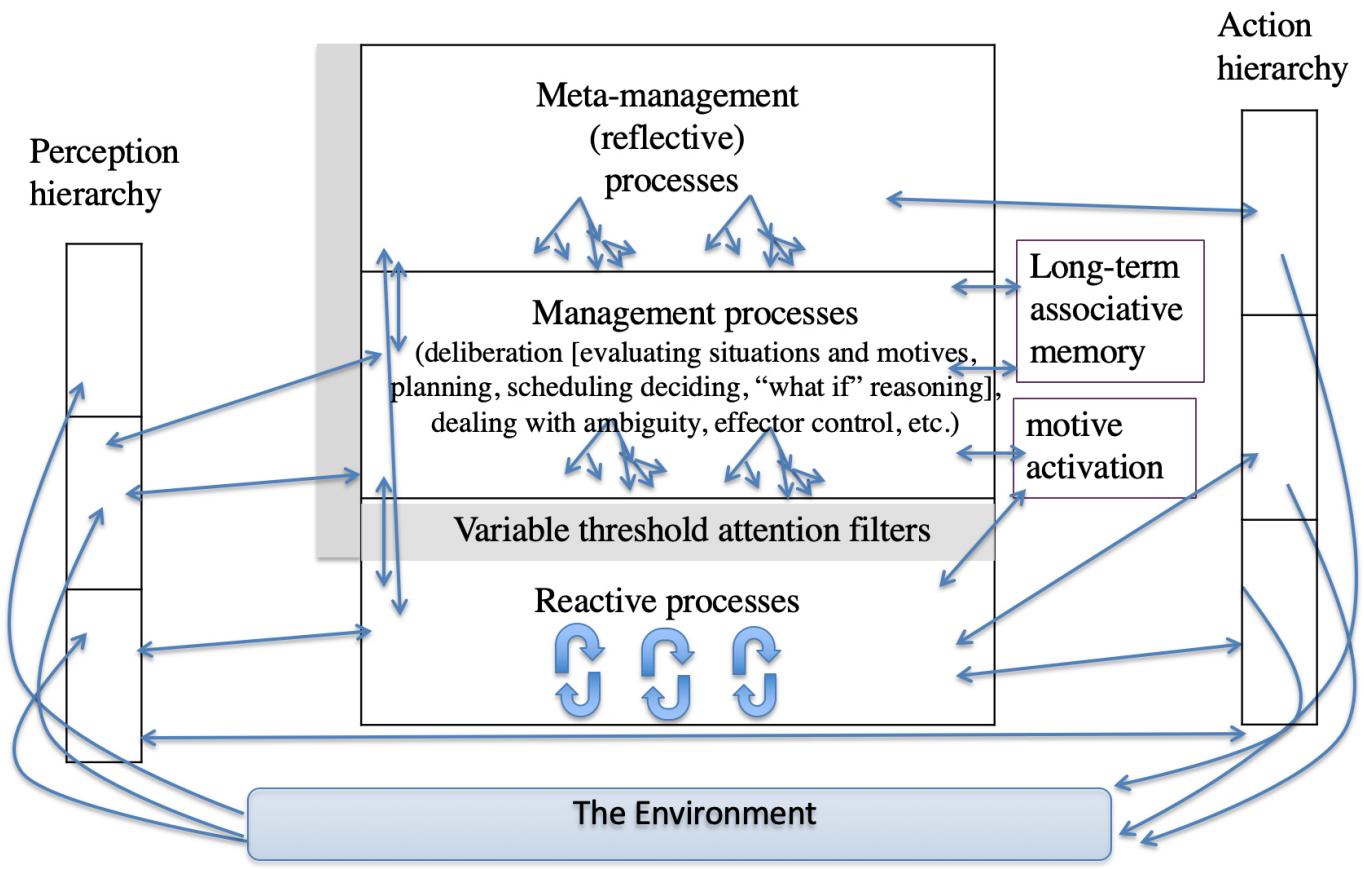

Figure 2. H-CogAff architecture diagram adapted from (Sloman, 2008a).

The highly internally connected H-CogAff architecture includes *reactive* mechanisms for perceiving and affecting the environment, creating and activating motivators in real-time, and generating alarms, all of which happens asynchronously from executive processes. The simplified architecture diagram is neither meant to imply sharp discontinuities between functions nor correspondence between function and biological layers.

We define motivators as an extension of Sloman (1992) definition of affective states, namely as (a) dispositional control states (long term and short term) that: (b) exist at various levels in a control hierarchy, (c) include positive or negative evaluations of something, (d) have at least a tendency to distract executive functions, (e) produce or trigger explicit motives, which in turn, (f) have a tendency to influence behaviour. We view the three forms of subjective value discussed in Ortony, Clore & Collins (1988) (goals, standards, and attitudes) as motivators. Here we use goals and motives interchangeably. In this paper motivators are states of a virtual machine (Sloman, 2002), rather than the external objects (such as foods) that may indirectly trigger them and to which they might refer. We realize these recursive concepts make communication difficult, but software can produce and process recursive representations, and so can the minds we are trying to understand.

In this article, we are chiefly concerned with motives, which specify or imply states towards which the agent has a motivational attitude (to make true, make false; make true faster, etc.), which are described in more detail below and great detail in Beaudoin (1994). Still, more precise and numerous concepts implemented in software will be required.

The H-CogAff architecture supposes two layers of higher-order mental processes which, to more closely align this model with psychology, we refer to as *executive* processes (Diamond, 2013). In so doing, we provide a way to understand some executive functions in terms of H-CogAff in particular and autonomous agent architectures more generally. The two executive layers are (1) *management* processes for interpreting situations and deliberating (e.g., evaluating motives, planning, scheduling, deciding, reasoning, problem solving, etc.), dealing with ambiguity, and various forms of motor control; and (2) *meta-management* processes (reflection and control of management processes). The meta-management layer could, for instance, postpone the consideration of a newly activated goal till some juncture, an example of deliberation scheduling. The reactive layer is more closely coupled to the environment than executive layers are; the latter can reason with contents of sensory memory, working-memory, short-term memory and long-term memory (Donald, 2001). The reactive layer is also more modular and more capable of parallel processing than the executive layer. Some reactive processes, however, can also respond to working memory contents.

Insistence Assignment and Motive Filtering

Given their limited parallelism, not every activated motive can be considered simultaneously by deliberative processes. Therefore, when a motive is generated or re-activated, there must be mechanisms, with similar computational constraints as reactive mechanisms have, that determine whether the deliberative processes may be interrupted (or otherwise influenced) by the motive. Therefore, H-CogAff architecture includes (a) insistence assignment mechanisms that *heuristically* assess the importance and urgency of motives as they are activated; (b) variable threshold filtering mechanisms which only allow a motive to *surface* (i.e., be considered by deliberative processes) if they are sufficiently insistent. For example, if a hungry autonomous agent that implements this architecture detects a rare opportunity to consume an energy source, a new motive to approach the source may be triggered. This motive will be assigned an insistence value that heuristically reflects its importance and urgency. However, for this motive to even be considered, it needs to be sufficiently insistent to penetrate the attention filter and interrupt current executive processing (and potentially behaviour). If the agent is under attack, its executive processes might not even notice its motive to approach the source of energy because the filter threshold will have been raised higher than the insistence level of the motive to approach the source of energy. Designing complex systems always involves trade-offs. Thus, it is impossible to design perfect insistence and filtering mechanisms. Because the purpose of insistence assignment is to protect the precious, resource-limited, deliberative processes, insistence mechanisms use rough and ready heuristics that do not involve deliberation. Sometimes, the agent will tend to be distracted by its own insistent motives even though it has

previously rejected them (e.g., to approach an appealing agent the pursuit of whom would violate its norms or other motives—conflicted robot or natural love.)

Whereas for simplicity the foregoing described insistence as a quantitative value and interrupt filtering as simply doing a numeric comparison between an insistence value and a global filter threshold, different forms of motivator filtering and attention switching or allocation are possible; insistence may be implicit rather than explicitly represented (Beaudoin, 1994). There could be different filtering criteria or rules for different objects and situations. For instance, one might learn to perceive certain situations as inherently dangerous to one's child, and implicitly perceived threats to one's child might become inherently capable of garnering management resources. Moreover, as discussed in the next section, interruption is not the only way in which motivators may *influence* executive functions; for instance motivators may consume executive resources, which some refer to as "attentional resources" (Pessoa 2013, Todd et al, 2020).

Computational Alarm Systems

Further addressing requirements of autonomous agency, and further accounting for psychological phenomena (such as aspects of 'emotional' and stress reactions), H-CogAff assumes mechanisms for generating and processing alarms. Alarms are control information-processing signals that have global effects in the architecture. At a physiological level, alarms can activate the sympathetic nervous system (Buck, 2014). We assume they can parameterize executive processes, such as leading to more vigilance, changing the level of abstraction of thinking, or make deliberation more or less careful. They may have other effects on management processes and "action readiness" that more precise formulations of the theory may specify.

The H-CogAff architecture distinguishes between 1) alarms triggered by perceptual information, such as an angry glare or the unexpected appearance of the object of one's infatuation; and 2) alarms triggered by noticing significant issues in executive layer content e.g., suddenly realizing a plan of action may have a disastrous side-effect (Sloman, 2003; Sloman, Chrisley & Scheutz, 2005).

Selye originally described stress as an *alarm reaction* (1936)—an idea that before the current paper had not been linked to computational alarms. Alarms have also (briefly) been posited in theories of consciousness (Baars & Franklin, 2009), emotion (Oatley, 1992) and pain (Eisenberger & Lieberman, 2004). We believe the IDO conception of alarms, modernizing Selye's concept (1936), is worthy of future IDO and empirical research.

Perturbance

Before specifying the concept of perturbance, it is relevant to consider its historical background. The *term* perturbance was introduced to the literature on emotion by Beaudoin (1994) to refer to the concept of emotion that was introduced by Sloman and Croucher (1981a, 1981b) and Sloman (1987, 1992. There was so much confusion and fruitless debate in emotion research about the proper meaning of 'emotion', Beaudoin proposed the term *perturbance* so that

researchers could focus not on what the term 'emotion' ought to mean in psychology and AI, but on the concept of perturbance and the theory that makes the concept relevant (Sloman, Beaudoin & Wright, 1994).

Sloman and Beaudoin used the term 'perturbance' in several publications (Sloman, Beaudoin & Wright, 1994; Wright, Sloman & Beaudoin, 1996; Beaudoin, 1994). But after Beaudoin left the field for a period of time, Sloman defined two new types of control states (named 'primary emotions and 'secondary emotions') and labeled 'perturbance' as 'tertiary emotion'. We are reintroducing the term 'perturbance' for the same reasons as before: a technical term better suites this unique and important concept. Similarly, we reject parlance of 'primary emotions' in favor of 'alarms'. Names do matter.

We are not alone to express concern about the plethora of concepts of emotion and to propose solutions. For instance, Izard (2010) survey of 34 emotion researchers found a wide variety of definitions of emotion. He suggested that "the topic of an abstract information-processing architecture for all mental functions [...] may be quite appealing to the growing number of scientists who postulate continuous interaction of emotion and cognition" (p. 368). The key idea of the concept of perturbance is that even if the reflective layer were to postpone consideration of an insistent motivator, the motivator still tends to penetrate the filter, consume some management resources, potentially distract management processes and or otherwise maintain control of executive processes. Perturbance is an emergent phenomenon. In fact, Sloman & Croucher (1981a, 1981b) claimed perturbance (which they called 'emotion') will emerge as side-effects in minds designed to meet the requirements of autonomous agency. They drew an analogy with thrashing in a computer operating system. One does not design a computer operating system to thrash. Thrashing is something that can emerge as a side-effect of needing to handle too many tasks with insufficient computational resources. Adaptiveness and function are attributes of the architecture and its constituent mechanisms. What *does* need to be designed into the system (by evolution, learning and or a designer) are mechanisms specified in the architecture (motive generators, insistence determiners, filters, executive processes, etc.) In that respect, perturbance is different from most concepts of emotion which assume that emotions serve a function.

The Component Process Model (Scherer, 2009) is another major computationally inspired model that claims emotion-like episodes are emergent. Its concept of emotion episodes differ from perturbance in several ways, one of which is that it necessarily involves a functional synchronization of major components (motivation, cognition, communication, experience, and physiological). Perturbance, in contrast, is an afunctional concept. Moreover, like Moors (2017), the CogAff model does not assume that in emotions (which we call 'perturbance') the agent enters in a stimulus-driven mode. Perturbance is a state in which executive functions are biased by and towards particular insistent motivators, though we allow for the possibility of alarms, discussed below, to be generated before a motivator is activated.

However, while perturbance is emergent and afunctional, it is of considerable adaptive significance because it is a modulation of executive processes. That which controls executive

functions controls the agent. Our focus as researchers and designers of autonomous agents interested in perturbance needs to be on the IDO of the mechanisms that give rise to adaptively meeting requirements of autonomous agents, how they may lead to perturbance, and how perturbance can be detected and dealt with. .

Simon (1967) developed the first influential computational (nearly IDO) theory of emotion known as the 'interrupt theory of emotion'. Noting the similarities and differences between perturbance and Simon's theory may help one to better understand Simon's theory and the concept of perturbance. This is particularly relevant because some emotion theorists, such as Scherer (1984), have summarily rejected Simon's interrupt theory without discussing the richer perturbance theory that improves upon Simon's theory. Simon's theory, like the perturbance theory, is based upon an analysis of the requirements of autonomous agency. They both emphasize the ability to activate, prioritize and pursue multiple motives. Simon (1967) assumes a highly serial central processor, whereas H-CogAff assumes more parallelism (e.g., between reflective and management processes). Simon's theory identifies emotions with interrupts of a central processor, whereas interruption is just one of the forms of perturbance. We envision that a CogAff design could be specified that includes continuously varying resources (Kruglanski et al, 2012; Pessoa, 2013), where deliberative and reflective processes could independently be consumed by different motivators. This means there are forms of perturbance (modulation of executive processes) that do not entail interruption. For instance, deliberation may be directed towards a certain goal, G, while another asynchronously activated motivator may consume some of the executive resources. This may affect deliberation about G without outright interrupting it. Insistent motivators may cause executive processes to proceed in a careful mode, more slowly or more quickly, as discussed in Sloman & Croucher (1981 a, b), or thinking may become more concrete or abstract. More generally, in perturbance a motive may *parameterize* executive functions. For instance, an asynchronously activated 'off task' motive may cause the agent to engage in social signaling (for instance to impress a potential mate or express sadness in grief). Simon's theory did not include the notion of insistence or dispositional control states. In comparing and contrasting Simon's work on motivation and emotion with ours, it is also worth noting that Simon (1967) was not explicit about computational architecture; moreover, his theory was focused on human information processing, rather than examining the space of possible minds.

Our characterization of perturbance has emphasized insistent *motivators*. However, there are two special cases that need to be considered with respect to perturbance. First, a motivator may be very insistent without being objectively or subjectively important, urgent (temporally pressing) or intense (driving behaviour). At the limit, a motivator could have zero importance, zero urgency and zero intensity and yet still be insistent. These dimensions are specified by Beaudoin (1994). An earworm would be an extreme example of this. This illustrates our claim that the distinction between cognition, emotion and affect is not sharp. Secondly, the concept of perturbance can be extended to apply to 'tertiary alarms' (Sloman, 2003), i.e., control signals that

disturb executive processes but unlike motivators do not necessarily contain semantic content (Oatley, 1992).

Moors (2017) described two sets of theories (psychological constructionism and dimensional appraisal) and her own, each of which deny the usefulness of the concept of emotion as a control *mechanism* while maintaining the concept of emotional *episodes*. She proposes that to understand emotional episodes one must provide an architecture-based theory of all kinds of behavior that involves both motivated and stimulus-driven mechanisms, where the architecture is biased towards goal-directed behavior, and where emotional episodes involve a goal-directed mechanism. Whereas Moors (2017) did not mention the CoffAff project, those postulates were also central to the theory of perturbance Sloman, 1981, 1987, 1992; Sloman, Beaudoin & Wright, 1994; Wright, Sloman & Beaudoin, 1996). For instance, the theory of perturbance also originates in an attempt to explain behavior. Mental perturbance, also, is not a mechanism but an emergent (episodic) state of a mental architecture involving insistent motivators. Reactive mechanisms in H-CogAff parallel 'stimulus' driven ones in her model. In addition to other similarities that space precludes us discussing here there are also several differences between Moors (2017) and the theory of perturbance. Moors (2017) has related her theory in more detail to psychology with a focus on experimentation and prediction, whereas work on perturbance and CogAff more generally has been more concerned with accounting for a broad spectrum of motivated competence.

The potential of a theory of perturbance for psychology derives partly from the IDO research approach that gave rise to it. This stance can help address a deep issue that surrounds psychology's "replication crisis" (Maxwell, Lau & Howard, 2015; Muthukrishna & Henrich, 2019), which is focusing too narrowly on *predicting* behaviour rather than *explaining* competence (Sloman, 2008; McCarthy, 2008). We call for (1) a better explicit characterization of human capabilities (competence), an exploration of mental architectures (designs), and implementations (Sloman, 1993); and (2) empirical research driven by unified theories of mind (Newell, 1990; Wells & Mathews, 1994). Cognitive architectures, still not prominent enough in psychology, require more attention, while motivational and affective processes require more consideration in computational architectures in psychology.

Two Common Classes of Perturbance

Let us briefly consider two types of perturbance that can, even without pathology, last for long periods of time and that have been overlooked by leading general theories of affect (Russell, 2009 ; Scherer, 2005 ; Moors, 2017), namely grief and limerence. These two states do not fit neatly in psychological theories that assume emotions are brief, lasting at most a few hours (Scherer, 2005; Verduyn & Lavrijsen, 2014; Verduyn et al, 2015). In contrast, grief and limerence (like many other perturbances) can last for weeks and months, without being pathological. As these examples illustrate (and the specification of the concept makes clear), perturbances are not moods, affect dispositions, preferences or interpersonal stances (the other categories described in emotion theories (Scherer, 2005). They involve insistent mental content

that tends to come to mind, even without proximal evaluations assumed by appraisal theories (activation and triggering of prior motives is often a better conceptualization than appraisal), regardless of our decisions to postpone their consideration.

Grief. When grieving, one tends to be assailed by memories and motives pertaining to the lost one. Wright, Sloman and Beaudoin (1996) offered a design-oriented reinterpretation of experienced episodes in terms of perturbance which was illustrated by a case study of grief. They claimed grief is (often) "an extended process of cognitive reorganization characterized by the occurrence of negatively valenced perturbant states caused by an attachment structure reacting to news of the death." (Wright, Sloman & Beaudoin, 1996, pp. 31). That theory addresses important questions such as: Why does grief consume the mourner? Reasons could be that executive processes have limited capacity and become swamped by highly insistent motives generated by a structure of attachment to a highly valued individual; in addition, re-learning and detachment require extensive rumination, which can maintain perturbance.

Limerence. The nearly universal attraction phase of romantic love is technically known as *limerence* (Reynolds, 1983; Tennov, 1979). It is noteworthy that whereas psychologists, as mentioned above, cannot agree on how emotion should be construed scientifically (Moors, 2017), let alone that it involves perturbance, those who study romantic love seem to agree that a necessary and defining feature of limerence is repetitive and intrusive thinking about the limerent object (Fisher, 1998; Reynolds, 1983; Tennov, 1979).

Limerence is of evolutionary significance as it enhances the likelihood of mating—and, in m dost cultures, of attaching to the limerent object, which helps offspring survive (Fisher, 1998). While it may be tempting to cast limerence as a pathological form of romantic love (Reynaud, Karila, Blecha & Benyamina, 2010; Wakin & Vo, 2008), this would distort the original and common academic conception of limerence (van Steenbergen, Langeslag, Band & Hommel, 2013). This would also overlook the near universality and evolutionary significance of the experience. Like other long-term affective states, limerence involves several continua, including intensity (Hatfield & Sprecher, 1986), and may or may not be pathological. We believe the casting of limerence as pathological should be resisted by scholars; instead other terms should be used to describe pathological limerence. We also recommend that scientific literature on the intrusive mentation aspect of attraction converge on the term 'limerence', to help focus research attention, and conceptualization, and to help shape popular psychology.

Perturbance, more generally, is diminishment of the already limited human capacity to control one's own attention with respect to a particular cluster of motives. Consider a limerent's diary entry "This obsession has infected my brain. I cannot shake those constantly intruding thoughts of you. Every thought winds back to you no matter how hard I try to direct its course in other directions." (Tennov, 1979 p. 49). Thus, a key feature of limerence is that meta-management processes cannot easily suppress motives nor prevent them from holding one's attention once they surface. Deliberation scheduling fails systematically in perturbance. Many, perhaps most, limerent minds are aware of this lack of self-control. This awareness is only

possible because (unlike most species) humans can, to a limited extent, monitor and voluntarily control their management processes (i.e., execute meta-management functions).

The H-CogAff framework seems to be at least as promising for limerence as it is for grief—two types of perturbance that normally involve attachment structures changing in opposite ways. Limerence, the attraction phase of romance (Fisher, 2004; Fisher, Aron & Brown, 2006), involves establishing attachment structures: motives, motive generators, insistence assignment rules, other reactive processes, plans, etc. Grief is an extended process of dismantling attachment structures. Limerence and grief overlap in heartbreak and lovelornness, which all require the dismantling of attachment structures. Also, like grief, limerence can loosen prior attachment (facilitating the abandonment of one's current partner for a new one, or forgetting a prior love). Accounting for attachment processes is important given that emotions seem to have evolved in large part to enable individuals to indirectly manage each other via commitments and attachments (Aubé, 2009). Perturbance has been examined in relation to attachment (Petters, 2016; Petters & Beaudoin, 2017).

Understanding limerence as perturbance allows the obsessive nature of limerence to be characterized in IDO terms, in a way that can account for similar (potentially long lasting) states. It encourages questions to be raised progressively about mental states in terms of whole-mind design (motive generators, attachment structures, etc.), leading to further requirement and design specification.

The perturbance theory of limerence can also be used to extend, in IDO terms, Miller's (2001) influential theory of human evolution through sexual selection. Producing limerence *qua* perturbance in a desired mate is an advantageous strategy. That is, it is advantageous to trigger the creation and activation of motive generators in the other mate that produce insistent attraction-related motivators towards oneself. Whether the mating motivators are triggered in the other is by one's socially signaling intelligence (Miller, 2001) or other forms of fitness (wealth, pro-social attitudes, etc., Simler & Hanson, 2017), the motivators in limerence hijack the other person's mind. Conversely, signaling that one is in a limerent state (which may be hard to fake) implicitly tells the potential mate that she or he is so valuable, because it indicates that one is dedicating (and, crucially, perhaps *committing*, Aubé, 2009) to him or her one's most precious resources: one's executive resources. For these and other reasons, the ability to signal and interpret perturbance in others is of evolutionary significance, whether the perturbance underpins limerence, grief or other conditions.

Emotion theorists in psychology have not considered loss of control of executive functions, and related attentional processes, as centrally pertinent to emotion, let alone from the designer stance. For instance, while it flirts with concepts of attention and is integrative, the Component Process Model (Scherer, 2005, 2009) does not deal with perturbance. This might partly be because this model views emotion as a special reactive mode of functioning, as argued by Moors (2017), which is relatively short term. Ironically, it is in a *biological* theory of emotion that a related disturbance is highlighted, in what Panksepp and Biven (2012), as well as Sloman (2003), call *tertiary* emotions. If the concept of emotional episode is to be retained in

psychology, we suggest that theoretical psychologists inquire as to why and how perturbance is possible in emotion (not simply whether they empirically tend to co-occur).

Repetitive and Intrusive Mentation Involve Perturbance

Watkins suggested that an important attentional phenomenon should be conceptualized as "repetitive thought" (RT). He echoed a definition of RT as a "process of thinking attentively, repetitively or frequently about one's self and one's world [forming] the core of a number of different models of adjustment and maladjustment" (Watkins, 2008, p. 163). Under the banner of RT, Watkins included such varied phenomena as cognitive and emotional processing of persistent intrusions, depressive rumination, perseverative cognition, rumination, worry, planning, problem solving, and mental simulation, mind wandering, counterfactual thinking, post-event rumination, defensive pessimism, positive rumination, reflection, habitual negative self-thinking. To this list we would add obsessive and compulsive mentation, cravings and preoccupation. Watkins (2008) notes that worry, for instance, was defined as "a chain of thoughts and images, negatively affect-laden and relatively uncontrollable" and as "an attempt to engage in mental problem-solving on an issue whose outcome is uncertain but contains the possibility of one or more negative outcomes" (p. 164). Watkins's reasons for favouring RT as the overarching concept were that it is more inclusive than the alternatives, atheoretical, clearer, highly correlated with measures of worry and rumination, and non-evaluative (constructive or unconstructive).

We agree that RT phenomena are scientifically significant. RT is a feature of normal self-regulation—everyone experiences intrusive mentation. Furthermore, some forms of RT are transdiagnostic (Harvey, Watkins, Mansell, & Shafran, 2004). In other words, they represent a common feature across a number of diagnostic categories of mental health dysfunction. For instance, high levels of rumination are associated with depression and anxiety (Nolen-Hoeksema, Wisco, & Lyubomirsky, 2008). Below, we briefly discuss insomnia which often involves bedtime RT and is itself of transdiagnostic significance (Dolsen, Asarnow, & Harvey, 2014).

However, there is room for amelioration in Watkins' (2008) conceptualization of RT. Firstly, whereas the expression "RT" suggests that the repetitive content is cognitive in the traditional sense (thinking and imagining), it often involves affectively charged motives and it often triggers motive-processing (e.g., assessing and deciding.) 'Repetitive mentation' would be a more inclusive expression. Further, the criterion of being atheoretical is unrealistic and counterproductive (as suggested in the discussion of IDO above); it also runs against Watkins's other criterion of being conceptually clear. One needs a general theory, beyond folk psychology, in relation to which intrusions and the executive processes that respond to them are specified.

Whether or not authors are explicit and clear about their theory, the concepts at play when RT is discussed scientifically require grounding in a functional architecture. Something must be generating motives; something must be interrupting when there are intrusions; something must be considering motives; something must be prioritizing them; etc. These mechanisms need to be named and specified in relation to an architecture. The theory ought to

"cut nature at its joints" and be amenable to a progressive research program of simulation, further theoretical development and cumulative empirical research (Cooper, 2007). Furthermore, the all-inclusive RT conceptualization comes at the cost of papering over significant differences, for instance between reflection and rumination. The farrago of RT concepts requires conceptual analysis and functional specification, which we expect will lead to much pruning and reclassification. In addition, the phenomena of RT are too global, involving too many diverse wide-ranging mechanisms of mind, to be understood without reference to a computational architecture. Moreover, one must understand the *how* of normal information processing (IP) to assess mentation as constructive or unconstructive.

Unfortunately, the RT literature has failed to adopt or develop architectural models of mind. For instance, in describing a highly studied phenomenon of RM, affective biases, Mathews, Mackintosh & Fulcher (1997) invoke interrupt signals, attentional vigilance, effortful suppression and intrusions. The concepts of cognitive and attentional 'biases' are currently cast mainly in terms of 'external and internal stimuli' (Mathews et al., 1997; Todd, Cunningham, Anderson & Thompson, 2012) and 'affective salience' (Schweizer et al., 2019) rather than in terms of motivators, insistence or motive processing, i.e., the mechanisms that are being 'biased' and that process them. The attentional bias and RT literatures fail to invoke an overall model of mind which, for instance generates motives, filters them, prioritizes, them and acts upon them, i.e., that addresses the types of capabilities with which H-CogAff is concerned.

Watkins (2008) and others point to control theory as an explanatory framework for RT and self-regulation. While some of these models are promising (Nafcha, Higgins & Eitam, 2016), they too need to be integrated within an IDO approach. They need to address rich qualitative control states and mechanisms that follow from the requirements of autonomous agency (Sloman, 1995).

H-CogAff provides a theoretical framework in relation to which classification and modelling of RT may proceed. This framework has the advantage of being constructed to explore how human minds might solve real world problems of autonomous agency. It is by no means a complete or detailed specification; but it has proven to be useful for generating and exploring models, many of which have already been implemented (Sloman, 2008a).

H-CogAff offers a path towards a deeper conceptualization of RT. According to Watkins (2008), intrusive thought (IT) is not considered a category of RT, likely because it is an essential aspect of RT. IT is better, and more generally, conceived as intrusive mentation (IM), and more deeply as perturbance. The concept of perturbance is based on the dispositional concept of insistence of mental content: a motive may be insistent and yet not disrupt processing. To understand IM as perturbance we must specify in terms of an architecture (like H-CogAff) the ways in which insistence assignment, interrupt filtering and attention switching are effected.

This may also help address the need in the RT literature for a design-oriented taxonomy of patterns of executive processes. Beaudoin (1994) and Wright (1997) put forth several categories, such as oscillation between decisions, manifest perturbance, digressions and maundering. Several other patterns have been identified in the CogAffect project (e.g., Petters,

2014 ; Wright, 1997). These, and several types of phenomena labelled by Watkins as RT (such as worry and rumination) need to be systematically characterized in terms of patterns of interaction between management, reflective and reactive processes in H-CogAff

Insomnia

Various forms of repetitive thought at bedtime (such as "racing thoughts" and worry) seem to delay sleep onset (Lemyre et al., 2020). In a review of the literature on pre-sleep cognition, Lemyre et al. (2020) concluded "Importantly, better characterizing cognitive activity in insomnia might help to develop more effective pre-sleep cognitive strategies to facilitate sleep onset. While research on such strategies is still scarce, it remains a promising avenue to help patients who are resistant to the conventional cognitive and behavioral therapy for insomnia" (p. 10). Dominant cognitive theories of insomnia (Espie, 2007; Harvey, 2005) invoke affective terminology, such as 'arousal', without commitment to theories to interpret the terms (e.g., Russell, 2003), and do not appeal to fundamental IDO theories. Beaudoin (2014) and Beaudoin et al. (2019) have put forth a prolegomenon towards an IDO theory of sleep onset and insomnia based on H-CogAff, dubbed *the somnolent information processing* (SIP) theory, which attempts to reverse engineer the human sleep-onset control system. The theory postulates that perturbance is insomnolent, meaning that it tends to delay sleep onset.

According to SIP theory, insistent motivators can trigger deliberative processing with respect to the motivators. Controlling one's deliberative processes in bed can be particularly difficult: when there are no other distractors, insistent motivators can loom large. Moreover, it supposes that fatigue (due to homeostatic sleep drive and circadian factors, (Borbély, Daan, Wirz-Justice & DeBoer, 2016) can make deliberation scheduling more difficult. This can make it difficult to postpone consider of insistent motivators. In SIP, insistent motivators are deemed to be insomnolent (a signal to the sleep onset control system to delay the onset of sleep). Executive processing of motivators can maintain the insistence of motivators. Moreover, the theory assumes that the imagery rich, diverse, fluid mentation that is characteristic of a successful sleep-onset period (Nielsen, 2017) is not merely a consequence of sleep onset, it is pro-somnolent (a signal to the sleep onset control system that progression towards sleep is appropriate). During perturbance, insistent motivators capture executive processing, and thus prevent such (presumably) pro-somnolent mentation.

From this theory, Beaudoin (2014c) derived serial diverse imagining, a 'cognitive shuffling' technique, that aims to facilitate sleep onset. This involves deliberate mentation with features of sleep onset (e.g., imagining diverse scenes and/or oneself moving, drawing on diverse episodic memory, incoherent mentation). The various forms of cognitive shuffle, including serial diverse imagining, are meant to work partly by interfering with bedtime perturbance, i.e., being counter-insomnolent, as proposed by Beaudoin (2014c) and Beaudoin, Digdon, O'Neill, and Rachor (2016). It is also meant to be pro-somnolent, partly by emulating sleep-onset like mentation. Whether or not this technique stands the test of thorough experiments, it illustrates the potential to understand ancillary brain mechanisms (here the sleep-onset control system) that

integrate motivational information from reactive and deliberative layers, that involve relatively cognitive processes (such as imagining scenes), and also involve meta-management processes. It also illustrates how from an IDO theory of mind and theory of perturbance one can derive techniques for self-help (regarding insomnolence) and clinical concerns (insomnia).

Other Psychological Phenomena

Several research problems need to be reinterpreted specifically with architecture-based models of autonomous minds that can support perturbance. In this section, we consider a wide variety of them.

Motivation tends to be conceived in psychology simply as directing and energizing behaviour (Danziger, 1997) (determining the goals people choose; and when, why, and how intensely they pursue them), rather than in terms of motive processing (how motives can be processed and pursued by autonomous agents). For instance, none of the peer responses to the Selfish Goal theory in *Behavior & Brain Sciences* (Huang & Bargh, 2015) noted its lack of explicit architecture nor that its goal specification and processes are bare (e.g., what about motive generators and insistence?) Higgins (1997, 2011) notes that pleasure and avoidance of pain are still normally assumed to be *the* final ends, or more generally that behaviour seeks to maximize expected value, while the deeper, more subtle and generative possibility of architecture-based motivation (Beaudoin, 2014b; Sloman, 2009, 2019) is often ignored. In architecture-based motivation, through innate mechanisms, ontogenesis or learning — though not necessarily through reward-based mechanisms, nor hedonic mechanisms, nor means-ends analysis —minds can produce new motivator generators and new motivators. Hence, many of the 'hidden motives' described in Simler & Hanson (2017) as fundamental to human nature, are not, and need not be, explicitly represented at all, not even unconsciously. The concept of architecture-based motivation, which follows from H-CogAff and related designs, can help bridge the intentional stance (Dennett, 1987, where from the outside one ascribes representations that are not implemented in the observed agent) and the design stance. It also helps to understand the incommensurability of motivators (Beaudoin, 1994; Sloman, 2009, 2019).

Stanovich (2011) developed a promising theory to explain successes and failure in rationality, and to improve rationality. It contains a three-level architecture which refers to H-CogAff. Perturbance theory is also meant to account for apparent breakdowns in rationality (Sloman & Croucher, 1981). We think there is potential to combine Stanovich's framework with H-CogAff to better understand success and failure of rationality. For instance, Stanovich's framework could be augmented by affective constructs, such as motive generators and alarms. Meanwhile, the recent theory of cognitive energetics (Kruglanski et al., 2012), which is meant to explain all instances of goal-directed thinking in a quantitative way, also lacks an architecture. The related, quantitative, concept of economy of mind (Wright, 1997) was developed from the designer stance.

Given that perturbance is an underlying construct to explain RT, and some forms of RT are transdiagnostic, it stands to reason that the concept of perturbance is relevant to

transdiagnostic approaches. For instance, addictions involve motives that are both insistent (tend to capture attention) and intense (control behaviour). Obsessions and compulsions must also involve perturbance at their core. More generally, a design-oriented approach is required for transdiagnostic understanding (Hudlicka, 2017). Even more generally, to understand abnormal psychology and apparent breakdowns in rationality we must understand normal psychology in design-oriented terms.

Perturbance is also quite relevant to human memory. Following Anderson's (1991) adaptive explanation of memory, Beaudoin (2014a) proposed the heuristic relevance-signaling hypothesis from the designer stance. On a daily basis, humans process enormous amounts of information. The brain cannot deeply interpret it all, nor store all of its interpretations. Nor can the cortex explicitly signal relevance as a top down command to the hippocampus. (The direct command "I shall remember this phone number" does not work.) The brain needs implicitly to answer the question: what information should be persisted in memory? Testing effects are among the most well documented findings in empirical psychology: repeatedly recalling information potentiates memory of it (Roediger & Karpicke, 2006). The heuristic relevance-signaling hypothesis states that deliberative layer recall attempts are implicit cues to the brain's heuristic memory indexing mechanisms to prioritize access to information (memories) related to the perturbance—information (interpretations, narratives, etc.) that the deliberative layer has at least attempted to recall (reconstruct). Perturbances are hijackings of these mechanisms by insistent motives, potentiating memories related to the perturbant objects (e.g., the limerent object).

On another note, psychology has struggled with the question: in what respect can the experience of music in particular and art more generally be affective (Juslin & Vastfjall, 2008). From the designer stance we might similarly ask how can great art rivet us and reverberate within us, from catchy ear worms to more? It has been argued that a great story is one that holds one's attention (Boyd, 2009). This brings us close to the mark. The *architecture-based* concepts of insistence and perturbance suggest ways of deepening such explanations. We speculate that music and fiction can trigger an illusion of perturbance: the reflective-layer impression that the agent is experiencing a genuine perturbance (as if self-generated motives were insistently being activated, captivating management processes). More obviously, art likely often operates by increasing the insistence of one's own latent motives (triggering limerence and grief, for instance). Among the many reasons that limerence and grief are two of the most popular themes of art is that they are implicitly about perturbance and they trigger perturbance. Furthermore, for a work of art to have a social impact, it must affect individuals over periods of time, taking hold of their executive processes, and prompt them to think in its terms and to communicate about it. One way to explore these hypotheses would be to model responses to high-caliber, multi-modal art depicting limerence and grief that uses repetition in provocative ways, as is common in musical theatre.

We also believe a design-oriented theory of autonomous agency whose architectures can support perturbance can be applied to positive psychology and self-help. For example, focusing and flow are arguably essential to cognitive productivity and hence to knowledge economies.

Distraction is largely a motivational phenomenon —i.e., executive functions are captured not just by facts, but motives. Yet theories of attention —and knowledge translation on the subject (Gallagher, 2009; Levitin, 2014)— do not deal with motive processing and fail to consider, let alone account for, perturbance. Theories of learning, expertise and productive practice need to explain how humans can deliberately develop their mental architectures, e.g., creating new motive generators (Beaudoin, 2014a, 2014b).

In short, a broad range of previously studied phenomena and problems can systematically be revisited from the designer stance as involving perturbance.

Conclusion

In this paper, we have argued that perturbance is a major feature of the human mind that deserves to be thoroughly investigated. This concept has the advantage of being firmly rooted in AI and of involving a flexible, extensible architectural framework meant to account for requirements of autonomous agency. This enables research problems to be considered in terms of models of entire minds.

Many areas of interdisciplinary research on perturbance and autonomous agency more generally can fruitfully be pursued. Some have already been alluded to in this document. The concept of perturbance has the potential to unify several areas of study, including attention, emotional episodes and self-regulation, repetitive mentation, and psychopathological conditions such as depressive rumination, obsessive worrying and addictions. There is a rapidly growing number of instruments to automatically recognize emotions and to measure emotion perception (Adolphs, R., 2017). It is no surprise that there has yet to be research on whether or how humans tacitly perceive perturbance or how machines could do so, both of which would be challenging tasks that could advance theory. It may be helpful to integrate Moors' (2017) two-level architecture with H-CogAff, drawing on their respective strengths. There is a need for detailed modeling of mental processing in insomnia, for which the somnolent information processing theory provides a framework. Beaudoin (2014a) has argued in detail that the important concept of 'effectance', proposed by White (1959), which roughly means motivation for competence, needs to be modernized in terms of architecture-based motivation. Detailed IDO models of grieving and limerence as prolonged perturbance could be developed.

We urge resisting the temptation of *assimilating* the concept of perturbance to related concepts, such as obsession, rumination, infatuation, repetitive thought, or even emotion. Perturbance is not a phenomenological or descriptive concept, though the theory behind it is meant to also account for experience. What ultimately makes perturbance of interest are the IDO theories and approach in relation to which perturbance is to be understood.

The IDO approach is directly relevant to the education of educators, psychologists and cognitive scientists. In this paragraph we focus on psychology since it is, or should be, a requirement for the training of educators and cognitive scientists. There was a day when psychology students were virtually guaranteed to graduate knowing an overall model of the human mind, though they did not tend to believe it or use it. That model was based on the wrong

metaphor, hydraulic systems, as computers had not yet been invented. We are referring of course to Freud's id, ego, superego model of mind. In rejecting the model, psychology threw out the baby with the bathwater (Minsky, 2013). Fortunately, psychology students are trained to apply many theories to the same phenomena. Unfortunately, they are not yet typically trained to think about themselves, other humans and possible (AI) minds in terms of an IDO information-processing architecture with multiple interacting virtual machines — let alone, as they should, multiple such theories. Yet this is teachable and important (Borsboom et al., in press; Sloman, 1993; Beaudoin, 1994). Here we have focused on H-CogAff, but there are other relevant IDO models, such as Baars & Franklin (2009). We also recommend students be trained in conceptual analysis (Ortony, Clore, & Foss, 1987; Sloman, 1978), which is part of the IDO approach, as they are in empirical research methods. We are not suggesting a one-way flow of influence from a design-oriented perspective to phenomena-based methods. Instead, we advocate a progressive theory-driven research program to propose and improve IDO models. There is a need for more AI researchers to consider broad, integrative, multi-layered, affective autonomous agency. We believe psychology and AI researchers need to work more closely together, not only on purely cognitive problems but affective ones as well. AI and psychology must blend more (Reisenzein et al., 2013).

Acknowledgments

We would like to thank Dr. Eva Hudlicka for her contributions to an earlier draft of this paper.

References

Adolphs, R. (2017). Emotion perception from face, voice, and touch: comparisons and convergence. *Trends in Cognitive Sciences, 21*(3), p. 216–228. doi:10.1016/j.tics.2017.01.001

Anderson, J. R. (1991). Is human cognition adaptive. *Behavioral and Brain Sciences*, *14*(3), p. 471–485. doi:10.1017/S0140525X00070801

Aubé, M. (2009). Unfolding commitments management: A systemic view of emotions. In J. Vallverdú & D. Casacuberta (Eds.), *Handbook of research on synthetic emotions and sociable robotics: New applications in affective computing and artificial intelligence*, p. 198–277. New York, NY: IGI Global.

Baumeister, R. F. (2015). Toward a general theory of motivation: Problems, challenges, opportunities, and the big picture. *Motivation and Emotion*, *40*(1), 1–10.

Baars, B. J., & Franklin, S. (2009). Consciousness is computational: The LIDA model of global workspace theory. *International Journal of Machine Consciousness*, *1*(1), 23–32. doi:10.1142/S1793843009000050

Beaudoin, L. P. (1994). *Goal processing in autonomous agents* (Doctoral dissertation). Birmingham, England. Retrieved from http://www.cs.bham.ac.uk/research/projects/cogaff/Luc.Beaudoin_thesis.pdf

Beaudoin, L. P. (2014a). *Cognitive productivity: Using knowledge to become profoundly effective*. Pitt Meadows, BC: CogZest. Retrieved from https://leanpub.com/cognitiveproductivity/

Beaudoin, L. P. (2014b). Developing expertise with objective knowledge: Motive generators and productive practice. In J. Wyatt & D. Petters, *From Robots to Animals and Back* (pp. 161–189). Springer.

Beaudoin (2014c). A design-based approach to sleep-onset and insomnia: super-somnolent mentation, the cognitive shuffle and serial diverse imagining. Retrieved from https://summit.sfu.ca/item/17237.

Beaudoin, L. P., Digdon, N., O'Neill, K., & Rachor, G. (2016). Serial diverse imagining task: A new remedy for bedtime complaints of worrying and other sleep-disruptive mental activity (pp. A209). Denver. Retrieved from http://summit.sfu.ca/item/16196

Beaudoin, L. P., Lemyre, A., Pudlo, M. & Bastien, C. (2019). *Towards an integrative design-oriented theory of sleep-onset and insomnolence from which a new cognitive treatment for insomnolence (serial diverse kinesthetic imagining, a form of cognitive shuffling) is proposed*. World Sleep Congress 2019, Vancouver, BC.

Beaudoin, L. P., & Sloman, A. (1993). A study of motive processing and attention. In A. Sloman, D. Hogg, G. Humphreys, A. Ramsay, & D. Partridge (Eds.), *Prospects for Artificial Intelligence* (Proceedings AISB–93), (pp. 229–238). Birmingham: IOS Press.

Boden, M. A. (1978). *Purposive explanation in psychology*. Cambridge, MA: Harvard University Press.

Boden, M. A. (1988). *Computer models of mind*. Cambridge University Press.

Boden, M. A. (1989). *Artificial Intelligence in psychology*. Bradford Books.

Boden, M. A. (2006). *Mind as machine: A history of cognitive science* (2 volumes). Oxford University Press.

Boden, M. A. (2008). *Mind as machine: A history of cognitive science*. Oxford University Press.

Borbély, A. A., Daan, S., Wirz-Justice, A., & DeBoer, T. (2016). The two-process model of sleep regulation: a reappraisal. *Journal of Sleep Research*, *25*(2), 131–143. doi :10.1111/jsr.12371

Borsboom, D., van der Maas, H., Dalege, J., Kievit, R., & Haig, B. (2020). Theory Construction Methodology: A practical framework for theory formation in psychology. https://psyarxiv.com/w5tp8/

Boyd, B. (2009). *The origin of stories*. Harvard University Press.

Buck, R. (2014*). Emotion: A biosocial synthesis*. Cambridge University Press.

Cooper, R. P. (2007). The role of falsification in the development of cognitive architectures: Insights from a Lakatosian analysis. *Cognitive Science*, *31*, 509–533.

Danziger, K. (1997). *Naming the mind*. SAGE.

Dennett, D. C. (1971). Intentional systems. *The Journal of Philosophy*, *68*(4), 87–106.

Dennett, D. C. (1987). *The intentional stance*. Cambridge, MA: The MIT Press.

Dennett, D. C. (1994). Cognitive science as reverse engineering. In D. Prawitz, B. Skyrms, & D. Westerstahl (Eds.), *Proceedings of the 9th International Congress of Logic, Methodology and Philosophy of Science*.

Diamond, A. (2013). Executive Functions. *Annual Review of Psychology*, *64*(1), 135–168. doi :10.1146/annurev-psych-113011-143750

Disalle, R. (2006). *Understanding space-time: the philosophical development of physics from Newton to Einstein*. Cambridge University Press.

Dolsen, M. R., Asarnow, L. D., & Harvey, A. G. (2014). Insomnia as a transdiagnostic process in psychiatric disorders. *Current Psychiatry Reports, 16*(9), 471. doi:10.1007/s11920-014-0471-y

Donald, M. (2001). *A mind so rare: The evolution of human consciousness*. W. W. Norton & Company.

Espie, C. A. (2007). Understanding insomnia through cognitive modelling. *Sleep Medicine*, 8, S3–S8. doi:10.1016/S1389-9457(08)70002-9

Eisenberger, N. I., & Lieberman, M. D. (2004). Why rejection hurts: a common neural alarm system for physical and social pain. *Trends in Cognitive Sciences*, *8*(7), 294–300. doi:10.1016/j.tics.2004.05.010

Firestone, C., & Scholl, B. J. (2016). Cognition does not affect perception: Evaluating the evidence for "top-down" effects. *Behavioral and Brain Sciences*, 39.

Fisher, H. (1998). Lust, attraction, and attachment in mammalian reproduction. *Human Nature*, *9*(1), 23–52. doi:10.1007/s12110-998-1010-5

Fisher, H. (2004). *Why we love: The nature and chemistry of romantic love*. New York, N.Y.: Henry Holt & Company.

Fisher, H., Aron, A., & Brown, L. L. (2006). Romantic love: A mammalian brain system for mate choice. *Philosophical Transactions of the Royal Society of London*. Series B, Biological Sciences, 361(1476), 2173–2186. doi:10.1098/rstb.2006.1938

Fontaine, J. R. J., Scherer, K. R., & Soriano, C. (2013). *Components of emotional meaning*. Oxford University Press.

Gallagher, W. (2009). *Rapt*: *Attention and the focused life.* The Penguin Press.

Harvey, A. G., Watkins, E., Mansell, W., & Shafran, R. (2004). *Cognitive behavioural processes across psychological disorders*. Oxford University Press, USA.

Harvey, A. G. (2005). A cognitive theory and therapy for chronic insomnia. *Journal of Cognitive Psychotherapy, 19*(1), 41–59.

Hatfield, E., & Sprecher, S. (1986). Measuring passionate love in intimate relationships. *Journal of Adolescence*, *9*(4), 383–419.

Hawes, N. (2011). A survey of motivation frameworks for intelligent systems. *Artificial Intelligence*, *175*(5-6), 1020–1036. doi :10.1016/j.artint.2011.02.002

Higgins, E. T. (1997). Beyond pleasure and pain. *American Psychologist 52*(12). 1280-1300.

Higgins, E. T. (2011). *Beyond pleasure and pain*. OUP USA.

Huang, J. Y., & Bargh, J. A. (2015). The Selfish Goal: Autonomously operating motivational structures as the proximate cause of human judgment and behavior. *Behavioral and Brain Sciences*, *38*(01), 121–135. doi :10.1017/S0140525X13000290

Hudlicka, E. (2017). Computational modeling of cognition-emotion interactions: Theoretical and practical relevance for behavioral healthcare. In M. P. Jeon (Ed.), *Handbook of affective sciences in human factors and HCI* (pp. 1–62). Waltham, MA : Elsevier.

Izard, C. E. (2010). The many meanings/aspects of emotion: Definitions, functions, activation, and regulation. *Emotion Review*, *2*(4), 363–370. doi :10.1177/1754073910374661

Juslin, P. N., & Vastfjall, D. (2008). Emotional responses to music: The need to consider underlying mechanisms. *Behavioral and Brain Sciences*, *31*(05), 65–63. doi :10.1017/S0140525X08005293

Kruglanski, A. W., Bélanger, J. J., Chen, X., Köpetz, C., Pierro, A., & Mannetti, L. (2012). The energetics of motivated cognition: A force-field analysis. *Psychological Review*, *119*(1), 1–20. doi :10.1037/a0025488

Lakatos, I. (1980). The methodology of scientific research programmes. *Philosophical papers*, 1. Cambridge University Press.

Lemyre, A., Belzile, F., Landry, M., Bastien, C. H., & Beaudoin, L. P. (2020). Pre-sleep cognitive activity in adults: A systematic review. *Sleep Medicine Reviews*, *50*, 101253. doi :10.1016/j.smrv.2019.101253

Levitin, D. J. (2014). *The organized mind: Thinking straight in the age of information overload*. Dutton Penguin.

Macatee, R. J., Allan, N. P., Gajewska, A., Norr, A. M., Raines, A. M., Albanese, B. J., et al. (2015). Shared and distinct cognitive/affective mechanisms in intrusive cognition: An

examination of worry and obsessions. *Cognitive Therapy and Research*, 40(1), 80–91. doi:10.1007/s10608-015-9714-4

Marcus, G., & Davis, E. (2019). *Rebooting AI: Building artificial intelligence we can trust.* Pantheon.

Mathews, A., Mackintosh, B., & Fulcher, E. P. (1997). Cognitive biases in anxiety and attention to threat. *Trends in Cognitive Sciences*, *1*(9), 340–345.

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does « failure to replicate »" really mean? *The American psychologist*, *70*(6), 487–498. doi:10.1037/a0039400

McCarthy, J. (2008). The well-designed child. *Artificial Intelligence*, *172*(18), 2003–2014. doi:10.1016/j.artint.2008.10.001

Miller, G. (2001). *The mating mind: How sexual choice shaped the evolution of human nature*. Anchor.

Minsky, M. L. (1985). *The society of mind*. New York, NY: Simon & Schuster.

Minsky, M. (2013). Why Freud was the first good AI theorist. In M. More & N. VitaMore (Eds.), *The transhumanist reader* (pp. 167–176). Oxford: John Wiley & Sons. doi:10.1002/9781118555927.ch16

Minsky, M., Singh, P., & Sloman, A. (2004). The St. Thomas Common Sense Symposium: for Human-Level Intelligence. *AI Magazine*, *25*(2), 113–124.

Moors, A. (2017). Integration of two skeptical emotion theories: Dimensional appraisal theory and Russell's psychological construction theory. *Psychological Inquiry*, *28*(1), 1–19. doi:10.1080/1047840X.2017.1235900

Moors, A., & Fischer, M. (2018). Demystifying the role of emotion in behaviour: toward a goal-directed account. *Cognition & Emotion*, 1–7. doi:10.1080/02699931.2018.1510381

McCarthy, J. (2008). The well-designed child. *Artificial Intelligence*, *172*(18), 2003–2014. doi:10.1016/j.artint.2008.10.001

Muthukrishna, M., & Henrich, J. (2019). A problem in theory. *Nature Human Behaviour*, *3*(3), 221-229.

Nafcha, O., Higgins, E. T., & Eitam, B. (2016). Control feedback as the motivational force behind habitual behavior. In *Motivation - theory, neurobiology and applications*, *229* (pp. 49–68). Elsevier. doi :10.1016/bs.pbr.2016.06.008

Newell, A. (1990). *Unified theories of cognition*. Cambridge, MA : Harvard University Press.

Nielsen, T. (2017). Microdream neurophenomenology. *Neuroscience of Consciousness*, *3*(1), 1–18. doi:10.1093/nc/nix001

Nolen-Hoeksema, S., Wisco, B. E., & Lyubomirsky, S. (2008). Rethinking rumination. *Perspectives on Psychological Science*, *3*(5), 400-424.

Oatley, K. (1992). *Best Laid Schemes*. Cambridge: Cambridge University Press.

Ortony, A., Clore, G. L., & Foss, M. A. (1987). The referential structure of the affective lexicon. *Cognitive Science: A Multidisciplinary Journal*, *11*(3), 341–364.

Ortony, A., *Clore*, G. L., & *Collins*, A. (*1988*). *The cognitive structure of emotions.* Cambridge University Press.

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). *The measurement of meaning*. University of Illinois Press.

Panksepp, J., & Biven, L. (2012). *The archaeology of mind: neuroevolutionary origins of human emotions (Norton series on interpersonal neurobiology)*. WW Norton & Company.

Pervin, L. A. (1989). Goals concepts: themes, issues, and questions. In L.A. Pervin (Ed.), *Goal concepts in personality and social psychology*, pp. 473–479. Lawrence Erlbaum Associates, Inc.

Pessoa, L. (2008). On the relationship between emotion and cognition. *Nature Reviews Neuroscience*, 9(2), 148-158.

Pessoa, L. (2013*). The cognitive-emotional brain*. MIT Press.

Petters, D. (2014). Losing Control Within the H-Cogaff Architecture. In D. Petters (Ed.), *From animals to robots and back: Reflections on hard problems in the study of cognition*, *22* (pp. 31–50). Springer International Publishing. doi :10.1007/978-3-319-06614-13

Petters, D. (2016). An encounter between 4e cognition and attachment theory. *Connection Science*, *28*(4), 387–409. doi:10.1080/09540091.2016.1214947

Petters, D., & Beaudoin, L. P. (2017). Attachment modelling: From observations to scenarios to designs. In P. Erdi, B. S. Bhattacharya, & A. Cochran (Eds.), (pp. 227-271). *Computational Neurology and Psychiatry*.

Power, R. (1979). The organisation of purposeful dialogues. *Linguistics*, *17*, 107–152.

Reisenzein, R., Hudlicka, E., Dastani, M., Gratch, J., Hindriks, K., Lorini, E., & Meyer, J.-J. C. (2013). Computational modeling of emotion: toward improving the inter- and intradisciplinary exchange. *IEEE Transactions on Affective Computing*, *4*(3), 246–266. doi:10.1109/T-AFFC.2013.14

Reynaud, M., Karila, L., Blecha, L., & Benyamina, A. (2010). Is love passion an addictive disorder? *The American Journal of Drug and Alcohol Abuse*, *36*(5), 261–267. doi:10.3109/00952990.2010.495183

Reynolds, S. E. (1983). « Limerence »: A new word and concept. *Psychotherapy: Theory, Research and Practice*, *20*(1), 107–111.

Roediger, H. L., & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1*(3), 181.

Rosenbloom, P. S., Demski, A., & Ustun, V. (2016). The Sigma cognitive architecture and system: Towards functionally elegant grand unification. *Journal of Artificial General Intelligence*, *7*(1), 1–103. doi :10.1515/jagi-2016-0001

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological Review*, *110*(1), 145–172. doi:10.1037/0033-295X.110.1.145

Russell, J. A. (2009). Emotion, core affect, and psychological construction. *Cognition & Emotion*, *23*(7), 1259–1283. doi:10.1080/02699930902809375

Scherer, K. R. (1984). On the nature and function of emotion: A component process approach. In K. R. Scherer, Scherer, & P. Ekman (Eds.), *Approaches to Emotion* (pp. 293–317). Hillsdale, NJ.

Scherer, K. R. (2005). What are emotions? And how can they be measured? *Social Science Information*, *44*(4), 695–729. doi:10.1177/0539018405058216

Scherer, K. R. (2009). Emotions are emergent processes: they require a dynamic computational architecture. *Philosophical Transactions of the Royal Society*. Series B, Biological Sciences, 364(1535), 3459-3474.

Schweizer, S., Satpute, A. B., Atzil, S., Field, A. P., Hitchcock, C., Black, M., et al. (2019). The impact of affective information on working memory: A pair of meta-analytic reviews of behavioral and neuroimaging evidence. *Psychological Bulletin, 145*(6), 566–609. doi:10.1037/bul0000193

Selye, H. (1936). A syndrome produced by diverse nocuous agents. *Nature*, *138*(3479), 32-32.

Simler, K., & Hanson, R. (2017). *The Elephant in the brain : Hidden motives in everyday life*. Oxford University Press.

Simon, H. A. (1967). Motivational and emotional controls of cognition. *Psychological review*, *74*(1), 29.

Sloman, A. (1978). *The computer revolution in philosophy: Philosophy, science and models of mind*. Harvester Press. doi:10.2307/2184449

Sloman, A. (1987). Motives, mechanisms, and emotions. *Cognition and Emotion*, *1*(3), 217-233.

Sloman, A. (1989). On designing a visual system. *Journal of Experimental and Theoretical AI, 1*(4), 289–337.

Sloman, A. (1992). Prolegomena to a theory of communication and affect. In A. Ortony, J. Slack, & O. Stock (Eds.), *Communication from an artificial intelligence perspective* (pp. 229–260). Heidelberg, Germany.

Sloman, A. (1993). Prospects for AI as the general science of intelligence. In A. Sloman, D. Hogg, G. Humphreys, D. Partridge, & A. Ramsay (Eds.), *Prospects for Artificial Intelligence*, (pp. 1–10). IOS Press: Amsterdam. Retrieved from http://www.cs.bham.ac.uk/research/projects/cogaff/Aaron.Sloman_prospects.pdf

Sloman, A. (1995). Beyond Turing equivalence. In P. Millican, A. Clark (Eds.). *Machines and thought: The legacy of Alan Turing*, vol I: 179-219. First published 1990, revised 1995.

Sloman, A. (2002). Architecture-based conceptions of mind. *Synthese Library*, 403–430. Retrieved from https://www.cs.bham.ac.uk/research/projects/cogaff/sloman-lmpsfinal.pdf

Sloman, A. (2003). How many separately evolved emotional beasties live within us? In R. Trappl, P. Petta, & S. Payr (Eds.), *Emotions in humans and artifacts* (pp. 35–114). Cambridge, MA: MIT Press.

Sloman, A. (2008a). *The Cognition and Affect Project: Architectures, architecture-schemas, and the new science of mind* (pp. 1–34).

Sloman, A. (2008b). The well-designed young mathematician. *Artificial Intelligence*, *172*(18), 2015–2034. doi:10.1016/j.artint.2008.09.004

Sloman, A. (2010). *Two notions contrasted: 'Logical geography' and 'logical topography' variations on a theme by Gilbert Ryle: the logical topography of "logical geography"*. Retrieved from https://www.cs.bham.ac.uk/research/projects/cogaff/misc/logical-geography.html.

Sloman, A. (2011). Varieties of meta-cognition in natural and artificial systems. In M. T. Cox & A. Raja (Eds.), *Metareasoning: Thinking about Thinking* (Vol. 5144, pp. 307–322). Boston, MA: The MIT Press. doi:10.1007/978-3-540-85110-3_45

Sloman, A. (2009). Architecture-based motivation vs. reward-based motivation. *Newsletter on Philosophy and Computers* 9.1,10-13. (pp. 1–9).

Sloman, A. (2019). Architecture-based motivation vs. reward-based motivation. Retrieved from https://www.cs.bham.ac.uk/research/projects/cogaff/misc/architecture-based-motivation.html

Sloman, A., Beaudoin, L. P., & Wright, I. (1994). Computational modelling of motive-management processes. In N. H. Frijda (Ed.). *Proceedings of the 1994 Conference of the International Society for Research on Emotions*, Cambridge, UK.

Sloman, A., Chrisley, R., & Scheutz, M. (2005). The architectural basis of affective states and processes. In *Who needs emotions? The brain meets the robot* (pp. 203–244). New York: Oxford University Press.

Sloman, A., & Croucher, M. (1981a). Why robots will have emotions. In *Proceedings IJCAI 1981*, Vancouver. Retrieved from https://www.cs.bham.ac.uk/research/projects/cogaff/81-95.html

Sloman, A., & Croucher, M. (1981b). You don't need a soft skin to have a warm heart: Towards a computational analysis of motives and emotions (No. 004). *Cognitive Science Research Papers*, Sussex University. Retrieved from https://www.cs.bham.ac.uk/research/projects/cogaff/sloman-croucher-warm-heart.html

Stanovich, K. E. (2011). *Rationality and the reflective mind*. Oxford University Press, USA.

Tennov, D. (1979). *Love and limerence*. Scarborough House. doi :10.1037/018287

Thórisson, K., & Helgasson, H. (2012). Cognitive architectures and autonomy: A comparative review. *Journal of Artificial General Intelligence*, *3*(2), 1–30. doi:10.2478/v10229-011-0015-3

Todd, R. M., Cunningham, W. A., Anderson, A. K., & Thompson, E. (2012). Affect-biased attention as emotion regulation. *Trends in Cognitive Sciences*, *16*(7), 365–372. doi:10.1016/j.tics.2012.06.003

Todd, R. M., Miskovic, V., Chikazoe, J., & Anderson, A. K. (2020). Emotional objectivity: Neural representations of emotions and their interaction with cognition. *Annual Review of Psychology, 71*(1), 25–48. doi:10.1146/annurev-psych-010419-051044

Toomey, C. N. (1992). When goals aren't good enough. In J. Hendler (Ed.), *Artificial intelligence planning systems*, (pp. 311-312). College Park, Maryland: Morgan Kaufmann Publishers.

van Steenbergen, H., Langeslag, S. J. E., Band, G. P. H., & Hommel, B. (2013). Reduced cognitive control in passionate lovers. *Motivation and Emotion*, 444–450. doi:10.1007/s11031-013-9380-3

Verduyn, P., Delaveau, P., Rotgé, J.-Y., Fossati, P., & Van Mechelen, I. (2015). Determinants of emotion duration and underlying psychological and neural mechanisms. *Emotion Review*, 7(4), 330–335. doi:10.1177/1754073915590618

Verduyn, P., & Lavrijsen, S. (2014). Which emotions last longest and why: The role of event importance and rumination. *Motivation and Emotion, 39*(1), 119–127. doi:10.1007/s11031-014-9445-y

Wakin, A., & Vo, D. B. (2008). Love-variant: The Wakin-Vo IDR model of limerence. Inter-Disciplinary – Net. 2nd Global Conference: Challenging Intimate Boundaries.

Watkins, E. R. (2008). Constructive and unconstructive repetitive thought. *Psychological Bulletin*, *134*(2), 163–206. doi:10.1037/0033-2909.134.2.163

Wells, A., & Mathews, G. (1994). *Attention and emotion: A clinical perspective*. Hillsdale, NJ: Lawrence Erlbaum Associates Publishers.

White, R. W. (1959). Motivation reconsidered: the concept of competence. *Psychological Review*, 66(5), 297–333.

Wright, I. P. (1997). *Emotional agents* (Doctoral dissertation). University of Birmingham.

Wright, I., Sloman, A., & Beaudoin, L. P. (1996). Towards a design-based analysis of emotional episodes. *Philosophy, Psychiatry & Psychology*, *3*(2), 101–126. doi:10.1353/ppp.1996.0022