Summary of "Painful Intelligence"

Aapo Hyvärinen

6th September 2024

Abstract

This document is a short 12-page summary of my book titled *"Painful Intelligence: What AI can tell us about human suffering"*, in particular, the Second Edition published in 2024.

CH. 1: INTRODUCTION. One of the most important problems in science and philosophy is the problem of **human suffering**. This book approaches the problem by modelling the human mind and brain using the modern methods of AI, which can also be called computational neuroscience. The brain and the modern computer may not be that different after all: both receive input data, process information, and take some kind of actions based on that. Importantly, both humans and modern AI **learn** from input data: while modern AI is based on machine learning, human evolution can also be seen as a very long learning process. It is true that a computer does not (most likely) have consciousness, so it cannot experience suffering in the same way as humans. But, the central claim of this book is that we can still learn a lot about human suffering by looking at the **information processing behind the conscious experience** of suffering. In particular, based on such a computational theory, we can design **interventions** by which an agent, typically human, can reduce its own suffering.

PART I : SUFFERING AS ERROR SIGNALLING

CH. 2: DEFINING SUFFERING. In the literature of philosophy and neuroscience, two definitions are prominent. One definition, particularly favored by Buddhist and Stoic philosophy, says that suffering is about **frustration**, "not getting what one wants". Another definition, often associated with the contemporary philosopher Eric Cassell, says that suffering is about the perception of a **threat**, in particular, a threat to the "intactness of the person". These two definitions show how to approach suffering from the viewpoint of information processing, and later chapters will show how frustration and threat are both modelled in AI theory. These definitions can be seen as specific instances of the general principle of **error signalling**, which is ubiquitous in AI. Based on these definitions, any advanced AI or robot could be said to suffer as well. Most importantly, these definitions open the road to different interventions to reduce suffering, especially by finding ways to reduce frustration and threats.

CH. 3: FRUSTRATION DUE TO A FAILED PLAN. The branch of AI that is the basis of this book is the theory of **intelligent agents**. What is special about this theory is that it models agents that take **actions**, where every action changes the state of the world in some way. The fundamental computational problem in this theory is **planning**, which means finding a sequence of actions that takes the agent from the current state of the world to a state defined as the goal. A well-known fact in AI is that planning is computationally extremely demanding since the number of plans grows exponentially with the number of action steps considered. Moreover, there is often some randomness in the world: for example, unexpected obstacles may appear. For these reasons, sometimes a **plan fails**, which is precisely what we define as frustration. The goal state has to be chosen by some mechanism in the first place; that mechanism is defined as **desire**. With these definitions, we have formalized the idea that **frustration means the agent did not get what it wanted**.

CH. 4: MACHINE LEARNING AS MINIMIZATION OF ERRORS. Initial attempts at AI were based on programming intelligence explicitly; planning is one example. However, that approach turned out to be very difficult, and research shifted to

learning intelligence, which turned out to be much more successful. Modern AI is thus primarily based on machine learning, where the system learns regularities from **big data sets**. Such learning typically consists of minimizing some kind of **error measure** computed from the input data. For example, a system classifying objects in photographs would minimize classification error. Another problem with older AI was that it used logical operations ("and", "or", "not") on symbolic data that can be expressed as words ("cat", "dog", " black", "white"). In contrast, modern AI uses **neural networks**. Instead of logical operations, neural networks work on continuous-valued numbers, efficiently approximating complex functions even in high-dimensional spaces. Learning in neural networks is also typically **incremental**, which means that the learning happens by making small adjustments to the parameters of the network, often so that each input data point is used to compute one tiny adjustment.

CH. 5: FRUSTRATION DUE TO REWARD PREDICTION ERROR. Machine learning can be applied to the problem of action selection, thus avoiding the computational explosion inherent in planning. If the agent repeatedly takes actions in the same environment, it can learn the best actions for each state, thus circumventing any need for planning. The leads to the theory of **reinforcement learning**. In this theory, the principle of reaching the goal state is also generalized so that different states of the world contain different amounts of **reward**, and the agent should choose its actions to maximize the obtained reward. Reinforcement learning theory leads to its own definition of frustration. It is based on a quantity called **reward loss**: the difference between the reward the agent expected to get and the reward it actually got. If the **agent expected more than it obtained**, the agent is frustrated. This is closely related to a quantity called **reward prediction error (RPE)**: minimization of RPE alone is sufficient for reinforcement learning, which shows the importance of frustration for learning. Crucially, reward loss and RPE are based on **expectations or predictions**, which are thus necessary for producing frustration in this theory. An important implication of this theory is the **insatiability** of an agent: if it gets a lot of reward, it will update its prediction and expect even more in the future; thus, it will never be satisfied. A final point here is that the rewards for humans come mainly from evolution, leading to what we call **evolutionary obsessions**: evolution may force us to crave things that are neither good for us nor give real pleasure.

CH. 6: SUFFERING DUE TO SELF-NEEDS. Among the different desires, those related to the **self** are particularly important. One aspect of the multi-faceted concept of self is the **self-evaluation** of the agent's performance. The utility of such a self-evaluation is that the internal parameters governing the learning system can be modified to improve reward performance. This is called "learning to learn" and can be implemented by giving the agent internal, meta-level rewards based on its long-term performance. Unfortunately, such a self-evaluation system can lead to **meta-level frustration**, which occurs if the agent does not get the amount of reward that was expected in a longer-term evaluation. Another aspect of self is **self-preservation** or survival. It can easily be programmed in an AI, say a robot, but great care has to be taken so that this will not lead to risks for humans. Again, survival can be programmed as an internal reward system that gives negative rewards when the agent is anywhere close to destruction or death, which is a special case of the concept of a threat that will be defined next.

CH. 7: THREAT AS ANTICIPATION OF POSSIBLE FRUSTRATION. **Threat** is another source of suffering. The key to understanding threat is the concept of **risk aversion**. It is widely appreciated in economic theory that humans tend to avoid risks: we prefer a reward that is certain, as opposed to a gamble with some randomness, even if the gamble has the same expected reward. A threat is here considered the appearance of a risk that entails the possibility that the agent might get a very small (possibly strongly negative) reward. As always, the reward has to be compared with its expectation, so the threat is about a possible reward loss. Mathematically, the crucial point here is to consider the whole probability distribution of future reward, in contrast to the basic theory of reinforcement learning, which only computes expected (average) rewards. Thus, we define threat as the prediction of a **sufficiently probable and large frustration in the future**. The crucial difference between threat and frustration is that a threat creates suffering now merely based on a prediction of frustration in the future; threat means a possibility of future frustration as opposed to actual frustration. Nevertheless, an important implication of this theory is that threat depends on frustration; if there were no frustration, there could be no threat either.

CH. 8: FAST AND SLOW INTELLIGENCE AND THEIR PROBLEMS. Chapter 4 already introduced two different kinds of information processing, which is called the **dual-process** or dual-system theory in the case of human agents. On the one hand, there are **neural networks** that are fast, and in humans, mainly unconscious. The other system is called **"Good Old-Fashioned AI"**; it uses symbolic and logical processing, which is typically conscious and slow in humans. This distinction is seen in action selection as planning vs. reinforcement learning, which in humans corresponds to deliberate action vs. habits. It is important to understand that both systems have their advantages and disadvantages. Neural networks need **learning**, which is **slow, data-hungry, and inflexible**; but the resulting processing is **very fast** and powerful, especially in processing sensory data. Good old-fashioned AI is very bad at analyzing sensory data; it needs data with well-defined **categories**, while categories in real life tend to be **fuzzy and uncertain**. Nevertheless, good old-fashioned AI offers the potential for a more **flexible** intelligence whose reasoning can further be **communicated** to other agents. Optimally, the two systems should work together. For example, neural networks may analyze visual input and choose promising goal states, implementing a desire system; the planning system can then plan how to reach one of those goals.

CH. 9: SUMMARIZING THE MECHANISMS OF SUFFERING. To summarize, we have first, frustration defined as reward loss, and second, threat defined as the prediction of possible frustration in the future. Threat is secondary in the sense that it is dependent on frustration; if frustration were eliminated from the system, threat would be too. We have actually seen **two different definitions of frustration** above, one based on not getting what one wants and another based on not getting what one expects. These two definitions are closely related, if not the same. Another important point is that frustration works at **different time scales:** sometimes the brain can compute a reward loss in less than a second, while a long-term frustration of, say, life goals can be computed on a time scale of years. We also saw that in addition to the frustration of ordinary desires, the self-related (meta-level) desires can also be frustrated. It seems that the **self-related desires produce the strongest suffering.**

PART II : UNCONTROLLABILITY AND UNCERTAINTY

CH. 10: EMOTIONS AND DESIRES AS INTERRUPTS. Part II considers different information-processing mechanisms that amplify suffering. This chapter explains how emotions such as fear, disgust, or anger can be considered as **interrupts**, i.e., alarm systems that interrupt ongoing processing. For example, if a threat is perceived, processing unrelated to that threat is suspended, and all resources are directed to processing the threat. This may also entail stereotypical behavior, as in the famous fight-or-flight reaction. Even desire can be considered such an interrupt when it is strong, "burning", and captures all the agent's attention. Interrupts are essentially **computational shortcuts** that are needed because of limited computational resources. While rational thinking and emotions are often contrasted, this logic shows how emotions can sometimes increase rationality, improving reward optimization when computational constraints limit the amount of planning, for example. Yet, such interrupts also **decrease the control** that the conscious thinking system has over our behavior and information processing, and can lead to short-sighted behavior. Furthermore, such interrupts may be triggered too easily, since optimally tuning an alarm system is extremely difficult.

CH. 11: THOUGHTS WANDERING BY DEFAULT. Another case where the mind is uncontrollable is encountered with **wandering thoughts**. If you try to focus on a monotonous, boring task, thoughts with all kinds of unrelated content tend to invade your mind. Wandering thoughts mostly fall into two categories: planning the future, or recalling past events. Paradoxically, such wandering thoughts can be computationally useful in terms of future behavior. Since **planning** is computationally highly demanding, it makes sense to direct computational resources to planning when the current environment does not look particularly interesting or threatening. Likewise, it is useful to **replay** past events because of the incremental nature of learning in neural networks. Each time an event is replayed or re-input to the learning system, the parameters of the system are improved by just a tiny amount, so many repetitions are needed. However, planning and replay are **simulations** where any suffering is also repeated. Thus, wandering thoughts are computationally useful, but they increase suffering.

CH. 12: PERCEPTION AS CONSTRUCTION OF THE WORLD. **Uncertainty** is ubiquitous in any sophisticated sensory system, for example, when trying to recognize objects in a visual scene. The data available to a camera, or our eyes, is fundamentally insufficient. For example, the camera only measures a 2D projection of the real 3D world, and there are obstacles limiting our line of sight. To cope with this problem, the visual system in humans and sophisticated AI uses **prior information**, that is, an internal model of the surrounding world. Combining incoming data with prior information, such a system can compute the **probabilities** of possible explanations of the world, such as what objects are in an image. Yet, such inference can only give probabilities, not certain inferences. Optical illusions are simple illustrations of the idea that the human visual system can go seriously wrong. It is important to realize that many of the quantities related to suffering are based on perceptions and similar uncertain inferences. The uncertainty will necessarily increase frustration since it means the **predictions of reward often go wrong**. More fundamentally, the computed **reward loss** can be **uncertain, even illusory**.

CH. 13: DISTRIBUTED PROCESSING AND NO-SELF PHILOSOPHY. There is a fundamental, general reason for the uncontrollability of the human mind that we have seen in Chs. 10 and 11: it is the **parallel and distributed** nature of information processing in the brain. The processing happens in many neurons simultaneously; this is the parallel aspect. The neurons each have their own incoming input; this is the distributed aspect. Together, this leads to the problem of how the function of all the 10 billion neurons in the brain can be coordinated or managed. One metaphor talks about the **society of mind**, where small groups of neurons are compared to humans in a society; we can then discuss different ways of organizing such a society. One school of thought claims that such a society of mind has no leader, or that the brain has no central executive that would control the rest. This brings us close to the philosophy of **no-self** which is a central part of early Buddhist philosophy. People usually feel they have control over their actions, and they may even think they have free will, but should we take such perceptions seriously? The perception of control is just a perception and might be mistaken; our feeling of **free will could be an illusion**.

CH. 14: CONSCIOUSNESS AS THE ULTIMATE ILLUSION. It may seem futile to talk about suffering without referring to **conscious experience**; however, that is how most neuroscience works. Visual perception, for example, is intuitively the same as a conscious experience, but perception relies on a vast amount of unconscious processing that can largely be modelled in a computer. We already claimed that it is the same for suffering: consciousness is only one aspect, while information processing is another. This raises the question: **Why is there consciousness** in the first place? Does it have some important evolutionary function, or some computational utility? This chapter reviews several proposals, but none is found satisfactory since all the proposed functions can be programmed into a computer. In fact, it is even very **difficult to know if an agent is conscious**: whether animals are conscious has been a matter of debate for centuries, and now that question is being posed about AI. Yet, there is no denying that consciousness exists. One of the most radical claims in Buddhist philosophy is that **all we can ever know is the contents of our consciousness**, and we have no certainty if anything else really exists. However, the actual point in this claim may be that if the reality of the contents of the consciousness were not taken too seriously—if they were seen as mere mental constructs—conscious suffering would be reduced.

PART III : LIBERATION FROM SUFFERING

CH. 15: OVERVIEW OF THE CAUSES AND MECHANISMS. In this chapter, we first recapitulate the theory, based on Fig. 1. We define the **root causes of suffering** as having insufficient data and insufficient computation in a complex world, where our behavior aims to maximize rewards that we didn't choose ourselves. These imply the **three information-processing problems** of uncertainty, uncontrollability, and unsatisfactoriness (an umbrella term grouping insatiability and obsessions). Under such conditions, frustration cannot be avoided, and threats are perceived. There are also further information-processing **mechanisms that amplify suffering**: simulations, interrupts, self-needs, and treating the contents of the simulations as real. As the first step towards interventions, this chapter summarizes the mechanisms leading to frustration as an informal **frustration equation**. In this equation, the amount of frustration is the product of four factors: perceived reward loss, the certainty attributed to that perception, the amount of attention paid to it, and the times it is simulated or perceived. Each of these four factors can be the target of an intervention, and can be reduced to reduce frustration. Another viewpoint looks at the **dynamics of cognition**: the information processing inside an agent can be summarized as a cognitive cycle, starting from incoming sensory data and leading to frustration via planning and plan execution. Stopping the execution of this cycle will prevent frustration.

CH. 16: REPROGRAMMING THE BRAIN TO REDUCE SUFFERING . The first thing the frustration equation suggests is **reducing the expected reward**, since a lower expectation will reduce reward loss. However, reducing expectations is not easy, partly because they are often computed on an unconscious level. One way to achieve this is by a recognition of the uncontrollability of the world, and even of one's own mind. Such recognition will lead to lower expectations since not much reward can be expected in an uncontrollable world. The same applies to recognizing unsatisfactoriness, which will also reduce expectations and, thus, frustration. Another term in the frustration equation that can be reduced is the **certainty attributed to the perception** of reward loss: if the agent recognizes that it does not know exactly how much reward loss was incurred, it should not create a strong frustration signal. Developing this point further, Buddhist philo-

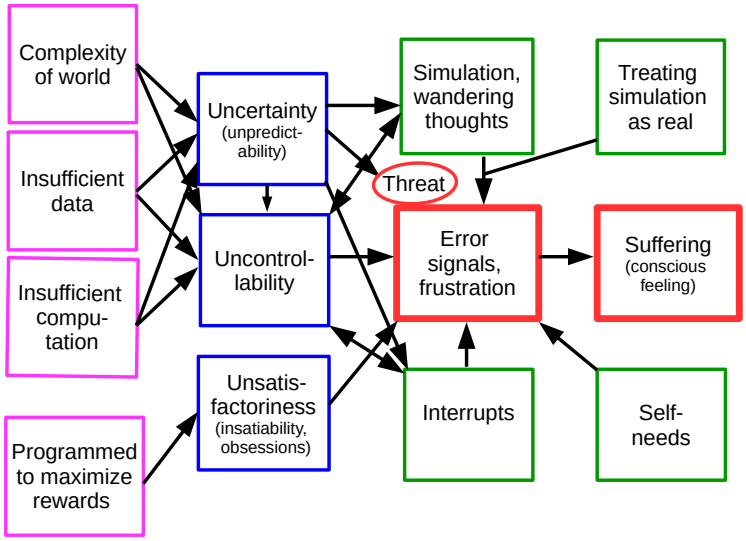

Figure 1: Recapitulation of the causes and mechanisms of suffering,

sophy emphasizes the concept of **emptiness**: the uncertainty of perception necessarily extends to all our thinking and, in particular, our mental usage of categories. Another approach to interventions is **reducing desires**: if there is no desire, no desire can be frustrated. In particular, reducing self-related desires is important here since they produce the strongest frustration. As a concrete intervention, the recognition of uncontrollability etc. works here just as well; furthermore, a **simple lifestyle** is beneficial.

CH. 17: RETRAINING NEURAL NETWORKS BY MEDITATION. **Mindfulness meditation** can be seen as one way of **directly teaching neural networks** to achieve the learning proposed in the preceding chapter. The problem is that because of the dual-process nature of the mind, whatever we think on an intellectual (symbolic-logical) may not have a lot of influence on the neural networks that largely control our expectations and desires. Meditation helps here since it is based on **direct observation** of the world and the mind. The meditator will **see how everything is uncertain and uncontrollable**; how thoughts and feelings

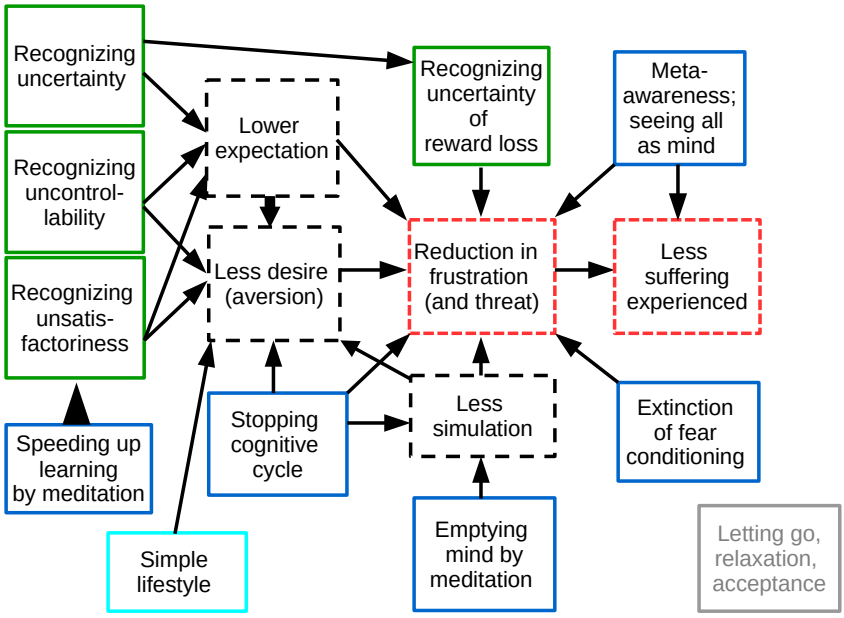

Figure 2: Recapitulation of the mechanisms of the interventions.

just come and go. This data is directly input to the neural networks to train them. Still, a lot of practice is needed because of the incremental nature of learning in neural networks: a large number of data points need to be fed into the system. Meditation has further benefits as well. It empties the mind, thus **reducing simulation** that tends to increase suffering. Another very different benefit of meditation is that by **observing the cognitive cycle** from sensory input to suffering, and learning to discriminate between different parts of this cycle, the meditator will be able to stop the cycle. In particular, the meditator can realize when a desire or planning is being created and, thus, decide to stop it. This reduces desires and the launching of plans that could be frustrated. Finally, meditation develops general **meta-cognitive abilities** which enable the meditator to take some distance from any mental phenomena, and ultimately see how frustration and threats are just mental phenomena produced by the brain; there is no need to take them, or indeed any contents of the consciousness, too seriously or personally. (The interventions of Ch. 16–18 are illustrated in Fig. 2.)

CH. 18: RECAPITULATING AND UNIFYING INTERVENTIONS. While the previous chapters approached the reduction of suffering from a "negative" viewpoint, in terms of reducing expectations, reducing desires and simulation, etc., it is possible to see the effect from a more "positive" viewpoint. Reducing desires and the like leads to **freedom**; it leads to **contentment** where one can even feel **gratitude** for what one has, instead of always wanting more. In fact, it is possible and useful to explicitly cultivate such attitudes by special meditation methods. Buddhist training further emphasizes that proper meditation training should include the attitudes of **acceptance** and **letting go**. Acceptance of all mental phenomena is important to avoid the pitfall of being averse to desire, or wanting not to want, which eventually would create a new kind of desire and suffering. Even suffering itself should be accepted at some level because aggressively fighting against it creates new suffering. Letting go is a related concept that can be considered to summarize the whole training in a Buddhist context. Ultimately, Buddhist training is about letting go of everything: a total **relaxation** of the mind.

CH. 19: EPILOGUE. This book was about modelling human suffering by AI. While a model can never be the real thing, the crucial point is that **this model allows for the design of interventions**. The conscious aspect of suffering is often intuitively considered paramount, but the point here is that interventions work better on the level of information processing. If one can prevent the information processing that causes suffering, it is not necessary to understand the conscious aspect. The interventions developed here are **very similar to Buddhist and Stoic** philosophy. Combining such philosophies with scientific theory, the interventions can hopefully be optimized and new ones developed. It should be noted that a scientific theory can only describe causal connections and mechanisms, not what you should do; whether an individual should apply these interventions depends on their preferences and life goals. This book also omitted several aspects of suffering; perhaps most crucially, anything related to social interaction and society. Later Buddhist thinkers criticized early Buddhist philosophy on the same ground, and proposed that the ultimate way to reduce one's suffering is, slightly paradoxically, to work to **reduce the suffering of others**.