AISB
2021

Communication and conversation

8 April 2021

Philosophy after AI

MEANING AND UNDERSTANDING

AISB2021

8 April 2021

Philosophy after AI Symposium - Schedule

13:00 - 13:20 Jonas Bozenhard, *AI and the Future of Philosophy – Philosophy and the Future of AI*
13:20 -13:40 Caterina Moruzzi, *On the Relevance of Understanding for Creativity*
13:40-13:50 Discussion

13:50-14:00 Break

14:00-14:20 Arthur Sullivan, *Non-Literal Usage as Operations on Meaning: What does Philosophical Pragmatics have to offer AI?*
14:20-14:40 Sonia de Jaeger, *Inevitably falling into place*: Intuition as a function of spatial prediction
14:40-14:50 Discussion

14:50-15:00 Break

15:00-15:20 Steve Battle, *Robots with purpose*
15:20-15:40 Rafal Kur, *The Abstractionist Origin of Analytic Concepts*
15:40-15:50 Discussion

15:50-16:00 Break

16:00-16:30 General discussion
16:30-16:45 Closing remarks, Giusy Gallo and Claudia Stancati

16:50-17:00 Break

17:00 Plenary talk: Prof Peter Robinson

Conference website: https://aisb20.wordpress.com/

ORGANISING COMMITTEE
Giusy Gallo, Dept. of Humanities, Università della Calabria, Italy – giusy.gallo at unical.it
Claudia Stancati, Dept. of Humanities, Università della Calabria, Italy – stancaticlaudia at libero.it

AI and the Future of Philosophy – Philosophy and the Future of AI

Jonas Bozenhard[1]

Abstract. Artificial Intelligence transforms many aspects of human life. It also changes the way we gain knowledge since various scientific fields have implemented AI-driven applications in their research practices. In contrast to the sciences, AI has not yet had any noticeable impact on the practice of philosophy. Yet, philosophers with transhumanist tendencies are convinced that future technologies will exceed the cognitive capabilities of humans and be able to find answers to the venerable questions of the philosophical canon. But many other philosophers are less optimistic and hold, for various reasons, that AI bears no substantial value for philosophical inquiry. For decades, philosophy has had a similarly ambiguous standing in the AI community: some AI researchers and practitioners collaborated with philosophers, while others either criticised philosophical interventions in their field or deemed them irrelevant for their work and ignored them. However, the ethical issues surrounding state-of-the-art and potential future AI systems are currently a matter of lively debate – and philosophers have played a key role in this conversation from the very beginning. In view of these constellations, this paper will focus on two questions: first, it will explore what value AI might have for the future of philosophy. This investigation will draw on different and, in part, contrasting metaphilosophical conceptions prevalent today and consider crucial potentials and limits of machine learning and AI more generally. On this basis, different visions of what philosophy after AI might look like shall be outlined. Secondly, the paper will also reflect on how philosophy can productively contribute to future AI research. In other words, the second half of the paper will examine what value philosophy might have for the future of AI.

[1] University of Oxford, UK, email: jonas.bozenhard@queens.ox.ac.uk

On the Relevance of Understanding for Creativity

Dr. Caterina Moruzzi[1]

Abstract. Artificial Intelligence (AI) applications help humans accomplish many everyday tasks and we are increasingly in contact with its always improving technologies. However, research in AI has been hindered by biases and misconceptions that concern its possibilities and by the uncertainty and lack of clarity in respect to what does it mean for a machine to 'learn' or to 'understand'. The aim of this paper is to discuss human and artificial understanding and to assess how state-of-the-art research fares in respect to the development of artificial systems capable of understanding. To do that, I first consider four notions that can help us in having a richer vision of what understanding amounts to: knowledge, grasp, problem-solving, and context sensitivity. In the central section of the paper, I identify two kinds of understanding that are necessary for creativity, a competence that current research in AI strives to reproduce in artificial systems. I then, discuss the impressive, but not completely satisfactory, achievements of Natural Language Processing, a branch of AI that is moving towards gaining a higher level of understanding. I close the paper by arguing how the current program of neurosymbolic AI is showing promising results in the progress toward achieving the desired level of understanding and creativity in machines. This program takes inspiration from the dual-model of the human mind proposed by Daniel Kahneman in his bestselling work *Thinking Fast and Slow* and it parallels the mechanisms that are necessary to develop not only a context-sensitive understanding, but also creative capabilities.

1 INTRODUCTION

While research in philosophy has focused on trying to provide a definition of terms like understanding, knowledge, and belief and on describing the connections that stand between these concepts, what in the literature has not arguably been sufficiently addressed is a discussion on the relation between the notions of understanding and creativity.

In this paper I aim to bridge this gap and to argue for the relevance of understanding for creative processes. After presenting the complexity of the notion of understanding and how it has been addressed in the literature, I will focus on how this term has been used in research on Artificial Intelligence (henceforth AI) (section 2.2). The main aim of the paper, indeed, will be that of discussing how state-of-the-art AI systems fare in respect to understanding and to indicate what I think is one of the most promising AI programs to develop machines capable of a deeper and context-sensitive understanding (section 5). Inspiration for this encouraging research path, the neurosymbolic AI program, is Daniel Kahneman's dual-system model of the mind's processes [1]. I will argue that an essential part of this model is the ability to undertake creative processes and that, in order to develop creative abilities, understanding is needed as well. I will not limit the discussion to the understanding skill that the agent of creative processes should display. Indeed, I claim that for creativity another kind of understanding is needed, namely the understanding from the part of the receiver (section 4). I, thus, argue that alongside the necessary care that is devoted toward the development of machines that are capable of understanding, attention toward the reception of the performance of these machines is equally essential.

2 UNDERSTANDING

2.1 Defining Understanding

Understanding can be enumerated among what Marvin Minsky called 'suitcase words', i.e. words that contain a variety of meanings. The term understanding is, indeed, a compound of different notions and concepts. There is a wide variety of things that can be understood. We can understand *that* we need to push a button to switch the lights on, *how* to bake a loaf of bread, or *why* the sun rises every morning [2]. The faceted nature of this concept started to be investigated already by ancient Greeks. Plato's dialogue *Theaetetus* is entirely concerned with a discussion on the nature of *episteme*. The latter can be translated with 'science', 'understanding', or 'knowledge' but it is toward this last interpretation of the word that the interest of philosophers has mainly shifted [3][2].

Despite the scarcity of debates and the diagonal complexity of the term, what scholars seem to agree on is that understanding is a kind of cognitive accomplishment that involves a discernment of some kind. Four concepts can help us in having a more thorough vision on the nature of understanding. These notions will be relevant also for the discussion on the relation between understanding and creativity: knowledge, grasp, problem-solving, and context-sensitivity.

The view that understanding is a form of knowledge is prevalent in the literature [9], [3]. Like knowledge, understanding is a cognitive achievement or success that can be ascribed to an agent and, like knowledge, it has been at the center of debates regarding the necessity of truth for achieving real understanding [3], [4], [9]. To many, however, understanding seems to be more demanding than knowledge. In order to understand, it is not enough to know: an additional component is needed which can be identified with the concept of grasping.

Understanding requires more than knowing information, it requires also that the agent grasps how different pieces of information are connected to one another [9], [3], [4]: "The act

[1] Dept. of Philosophy, University of Konstanz, 78464 Konstanz, Germany. Email: **caterina.moruzzi@uni-konstanz.de**.

[2] Indeed, there are not many works devoted to understanding in philosophical literature [4] but see [5], [6], [7], [8].

of joining the dots, of establishing relationships between facts is what grasping refers to" [10: 77]. This act of grasping happens often without us consciously being aware of it [10: 78][3]. Grasping a previously unknown connection between two elements or facts is accompanied by an *Aha!* moment [3: 14], a revelation that much has in common with the creative process. A key component of creativity is, indeed, the act of building connections between unrelated fields. Creativity as analogical thinking and the capacity of building metaphors and connections finds wide agreement in the literature [11], [12], [13], [14], [15]. A connection-making process consists in drawing links between apparently unconnected pieces of knowledge and in exploring novel paths toward the successful conclusion of the problem-solving task that is part of the creative process itself [16].

This introduces the next key concept that is relevant for understanding: problem-solving. Problem-solving processes consist in the exploration and analysis of the different paths that can be taken in finding a solution to the problem at hand [16]. Problem-solving is, thus, a more articulate stage than grasp. In order to successfully complete a problem-solving task, it is necessary not only to have a grasp of the relevant connections between the pieces of information provided but also to articulate their dependency and causal relations. A competent problem-solver needs to be aware of both the subject-matter and the context in order to fulfil her role [10].

The consideration of the context of action is an essential element in the evaluation of whether the agent completes her task successfully. Context-sensitivity is, thus, the fourth and last notion that can help us in forming a richer picture of what understanding amounts to. Understanding is a context-sensitive concept and what determines the ascription of understanding is the ascriber's and the agent's context [10]. Context is relevant not only in the consideration of whether it is possible to recognize an understanding capability to the agent but also as a necessary component of the agent's understanding. In order to be successful in grasping and problem-solving, the awareness of the contextual situation in which the agent is inserted is arguably necessary[4]. According to Bachmann [10] this contextualist interpretation of understanding can help us address the gradual nature of the notion. Understanding comes in degrees and the ascription of understanding to an agent is not binary: one can understand a situation more or less properly. The evaluation of one's capacity of understanding lies therefore on a spectrum.

I have here presented understanding as a cognitive accomplishment that requires knowledge of the state of affairs and grasp of the connections between different pieces of information available to the agent. The knowledge and grasp of the situation are necessary to accomplish the problem-solving task that the agent is confronted with and, in order to accomplish it with competence, a sensitiveness to the context is essential as well. So far, when talking about the subject of understanding, I referred to her as the 'agent'. When reading this word, I suspect that many of us will think about a *human* agent. However, the paradigm of understanding I outlined here can as well apply to non-human agents. In particular, in the next section I will quickly review how the notion of understanding can be interpreted in relation to artificial agents.

2.2 Understanding in AI

The first hurdle when addressing the notion of understanding in relation to machines is to distinguish between subjective and objective attributions of understanding. One thing is indeed to believe that machines can display understanding because we interpret their behavior *as* exhibiting understanding, and another thing is to acknowledge that there is a reliable basis for assuming that machines themselves understand [17: 11]. This is part of the well-known Chinese Room debate, namely the discussion on the possibility of deeming a digital computer able to 'understand' or have 'consciousness' if it behaves intelligently or in a human-like fashion[5].

A second hindrance in this discussion is the constant change of goalposts in what understanding for AI means. In this respect the cognitive psychologist and computer scientist Geoff Hinton says: "Before, if an AI system could translate a text, it would have been deemed as having understood the text, or if the system could sustain a conversation or describe the scene on an image. Now, none of these count as proper understanding." [18: 28]. The above-mentioned subjective vs objective attribution of understanding is one of the grounds for the difficulty in finding reliable standard benchmarks. The Other-Minds problem affects our capacity of assessing whether an individual has real understanding of a situation or if she just displays the suitable behavior. An analogous problem is run in relation to machines with the well-known challenge of building artificial systems the performance of which can be transparent and explainable [19].

A consequence of the re-calibration of goalposts for AI is the so-called 'AI-Effect', namely the disappointment toward research in AI when state-of-the-art systems fail to exhibit what is deemed to be true understanding. It is this disappointment which has contributed to the succession of AI summers and winters in the last decades and that, as I will discuss later (section 4.1), is part of a generalized discontent toward AI performance.

As pointed out by Aaron Sloman [17], the use of terms like 'true' understanding denotes a binary interpretation of the notion. Yet, as I mentioned above, understanding should not be assessed in a binary way. Rather, understanding stands on a spectrum and acknowledging this can allow us to have a more fine-grained way of assessing whether a machine can be capable of understanding: "it could 'understand' well enough to be an utterly slavish servant. It could not, however, be entrusted with tasks requiring creativity and drive, like managing a large company or a battle force, or minding children." [17: 8]. The ordinary concept of understanding that we use, as noticed above, includes a variety of different kinds of capabilities that can be variously exhibited by humans, animals, and machines. Trying to reduce this complex cluster to a binary assessment is neither appropriate nor fruitful.

The paradigm of understanding presented above can, with this proviso, be applied to the consideration of the capacity of an artificial agent to understand. The consideration of whether state-of-the-art AI can be capable of this will be the focus of section

[3] I will discuss how this element of unconscious grasp is relevant for the discussion regarding the application of Kahneman's dual-system model to the development of the neurosymbolic AI program (see section 5).

[4] This will prove to be a key point in the consideration of how state-of-the-art AI fares in respect to understanding, see section 4.

[5] See also Searle's observer-independent and observer-relative distinction, at https://www.youtube.com/watch?v=rHKwIYsPXLg.

4.2. Before then, what still needs to be discussed is how the notion of understanding relates to creativity and whether the capacity to understand is a prerequisite of creative processes.

3 UNDERSTANDING AND CREATIVITY

I argue that two kinds of understanding are necessary for creativity: the understanding possessed by the agent (what I will refer to as Agent-Understanding) and the understanding of the receiver (Recipient-Understanding). The receiver is whoever has access, as spectator or observer, to the creative process performed by the agent[6]. I will start discussing this kind of understanding first.

3.1 Recipient-Understanding

The understanding of the receiver is necessary for a process to be recognized as creative. Margaret Boden identifies three dimensions of creativity: combinational, exploratory, and transformational [20]. Transformational creativity "involves the transformation of some (one or more) dimension of the space, so that new structures can be generated which could not have arisen before. The more fundamental the dimension concerned, and the more powerful the transformation, the more surprising the new ideas will be." [20: 348]. Among the three, transformational creativity is the kind of creativity most challenging to achieve. The challenge resides in finding a balance between too little and too much novelty in the creative process that is being undertaken. If it is not novel enough, then it does not comply with the necessary requirements for creativity. But if it is too novel, it runs the risk of not being deemed creative, as it cannot be understood by the observers.

This is one of the arguments used by Sean Dorrance Kelly to argue against the possibility for an artificial system to surpass human creative abilities. Either the performance of the machines would be understandable and not surpass the achievements that could be reached by humans, or it would not be understandable and this would be against the possibility for machines to be deemed creative: "Suppose the best and brightest deep-learning algorithm is set loose and after some time says, 'I've found a proof of a fundamentally new theorem, but it's too complicated for even your best mathematicians to understand.' This isn't actually possible. A proof that not even the best mathematicians can understand doesn't really count as a proof. Proving something implies that you are proving it to *someone*." [21]

A possible objection may be raised here: the necessity of Recipient-Understanding for creativity may apply to domains like science where there is something objective that needs to be understood (like a theorem in Kelly's example). However, it does not apply to other fields, like the arts, where this objectiveness is lacking. In reply to this objection, I argue that, instead, also in the art-domain there is the necessity for a process/product to be understood by the receivers in order for it to be creative. An example is the famous case of Stravinsky's revolutionary ballet *The Rite of Spring* (1913) which caused a riot in the theatre on its premiere, since the audience was not used to such sonorities and harmonies. Many are the examples of unappreciated and not-understood geniuses in the history of art (to cite just a few: Van Gogh, Caravaggio, and Giuseppe Verdi). Even if the acceptance of a process as creative by the audience does not affect the ontological nature of the process itself, I argue that we need to be aware of the fact that creativity is a socially-constructed notion and, for this reason, how the term is used and in which circumstances its use is deemed legitimate constitutes a fundamental aspect of the concept, on a par with its more basic ontological features.

3.2 Agent-Understanding

The second necessary aspect for creativity is the understanding that is displayed by the agent of the creative process. When engaging in a creative process, an agent normally reflects on what she produces and tries to improve according not only to the feedback that she receive from the outside (part of the Recipient-Understanding) but also to the inner feedback she gives to herself. The ability to assess the process and to 'know when to stop' is a relevant element of creativity which elsewhere I referred to as 'evaluation' [16].

Understanding is a key part of this evaluative component and it is crucial in the process of trial and error performed when considering whether the connections established during the creative process work or not [12], 15], [22], [23], [24]. This problem-solving process highlights the relationship between creativity, understanding, and learning [25], [26][7].

In the next section I discuss whether state-of-the-art AI is capable of this second kind of understanding, Agent-Understanding, and if we as recipients are capable of the first kind.

4 STATE-OF-THE-ART AI AND UNDERSTANDING

4.1 Recipient-Understanding

In what follows, I discuss whether humans, as recipients of creative processes achieved by artificial agents, can display understanding toward them. I need to specify that my discussion applies to the domain of artistic creativity and that, therefore, different considerations could be made for the understanding of creative achievements of artificial agents in other fields.

In the domain of artistic creativity, it seems that humans struggle to achieve Recipient-Understanding in respect to the performance of artificial agents. The lack of Recipient-Understanding is not, however, caused by the fact that AI undertakes processes that are so novel that cannot be understood by us. As I will argue in the next section, state-of-the-art AI is still far from achieving this kind of creativity. Instead, the lack of Recipient-Understanding is caused by the challenge that we face in finding a reason why artificial agents should engage with tasks such as creativity that are a mark of human nature[8].

[6] When referring to the creative performance of an agent, I will refer to it a creative *process* instead of as creative *product*, see [16].

[7] This critical thinking stage is usually referred to also as 'selective retention' [12: 24], [27] or as 'convergent thinking' [28]. I will discuss later convergent thinking in the context of the neurosymbolic program, see section 5.

[8] There is also another level of understanding that we lack in respect to AI, namely the understanding related to the issue of transparency, see [19], [29], [30].

A survey I conducted in 2019 on perceptions of human and artificial creativity in the artistic domain revealed a strong opposition against the latter. Respondents felt that the incursion of AI into a paradigmatically human terrain such as creativity was not only improbable, they rejected the possibility with determination[9].

The majority of replies given by participants to the survey expressed skepticism about the possibility for AI to become creative: AI systems cannot be creative, since they lack experience and contextual knowledge, flexibility, feelings, intentionality, and other skills, such as unpredictability and the ability of self-evaluation.

Alongside these features, what participants reported as lacking to state-of-the-art AI, and that would instead be necessary for them to be deemed creative, are aspects that are essentially human, such as charisma, personality, experience, and the property of having and transmitting feelings and emotions. What seems to emerge from their replies is the belief that creativity and art are human-centric domains that are, and should remain, the prerogative of humans.

As I mentioned above, this survey was specifically designed to test the intuitions regarding artistic creativity. There is the possibility that these results would have been different if, instead of creativity in the artistic sector, the survey focused on creativity in the scientific domain or creativity in every-day life[10]. Still, even if the results of this survey do not allow us to draw definite conclusions, a reflection on how we perceive creativity in artificial systems can be initiated not only for having a better grasp of our own relation to machines but also to reflect on how best to design human-computer interfaces to prevent the risk of not taking advantage of the benefits that may derive from the human-machine interaction in the creative sector.

4.2 Agent-Understanding

I defined Agent-Understanding as the capacity to evaluate one's own performance. In the last years, researchers have developed generative algorithms which go in the direction of a more autonomous evaluation of their performance. It is the case of Creative Adversarial Networks (CANs) [32]. CANs are a development of the more famous Generative Adversarial Networks (GANs). These models consist of two different neural networks: a generator and a discriminator. The generator has the role of originating new data instances, while the discriminator evaluates them for authenticity. The feedback loop between generator and discriminator can be interpreted as an evaluation process, in some ways analogous to the human process of trial and error. Unlike other algorithms which use human judges to evaluate the algorithm's performance, in CANs the feedback mechanism happens within the system itself: "The CAN concludes the process when it reaches the right balance between generating works that would be accepted as 'art' and that, at the same time, are sufficiently style-ambiguous. This peculiar structure vouches in favor of attributing a capacity of autonomous evaluation to the system. There is no need for external feedback, the assessment of the generator's performance is carried out by the discriminator within the system: 'This interaction also provides a way for the system to self-assess its products.' [32: 21]." [16: 15]

Moving beyond the artistic sector, also in other domains state-of-the-art research in AI is trying to move toward the acquisition of more autonomous evaluation and understanding capacities. This is evident, in particular, in the field of Natural Language Processing (NLP) with the development of techniques aimed at embedding semantic, and not just syntactical, understanding in the programs.

In October 2019 the Google AI team released the model BERT (Bidirectional Encoder Representations for Transformers) which is supposed to consistently improve the users' experience of browser search in English and in other languages, thanks to an increased responsiveness to the context of the query and to a better 'understanding' of it. BERT was trained though the BooksCorpus (800M words) [33] and English Wikipedia (2,500M words) and, in respect to previous language models, it combines left-to-right and right-to-left training, thus processing words in relation to all the other words in the sentence, not just to the adjacent ones [34].

In 2020 another language and sequence transduction model was presented by OpenAI: GPT-3 (Generative Pretrained Transformer version 3). GPT-3 uses an input in order to make a prediction about the possibility of an output. The difference between GPT-3 and BERT is, that GPT-3 learns from an even bigger data set than BERT does, 175 billion parameters and 45TB of training data. which is why "it can excel in task-agnostic performance without fine tuning." [35].

In favor of these latest achievements in the field of NLP, it must be acknowledged that the mentioned models present an increased sensitivity to the semantics of language. The use of a bidirectional model allows BERT to incorporate contextual information from both directions of the text and, as the authors of the BERT paper say, "to fuse the left and the right context." [34]. Thanks to this innovative feature in the world of NLP, BERT has been by many deemed able to "understand the context of the words in your query" [36] or to "understand many fundamental relationships between words in English language." [37].

However, although BERT passed multiple tests, such as the 'common sense' and 'reading comprehension' tests, its capacity is still far away from that of human's common sense [37]. The same can be argued for GPT-3 which, although having a quantitatively superior dataset, is not arguably qualitatively superior to BERT [38]. Indeed, in order to acquire understanding of utterances, syntax and semantics are not sufficient. Meaning is linked to the general knowledge of the world, which is acquired by humans as we interact with the environment and to a pragmatic understanding of the use of language. In order for a system to understand language, it must create connections between different concepts and actions. Even large language models like the ones mentioned above cannot connect utterances to the environment when pre-trained with form data only [39].

From this short review of state-of-the-art AI systems in the field of understanding, it seems that AI is still at the beginning of the journey towards the achievement of Agent-Understanding. Indeed, AI lacks the necessary context-sensitivity, above identified as one of the key components of understanding, and

[9] For the details regarding the survey methodology and results, see [31].

[10] With the aim of responding to this interrogative, I will conduct a follow-up survey aimed at testing whether similar results would emerge in relation to the application of artificial creativity to the scientific domain or if, instead, the attribution of creativity to non-human entities would meet less resistance in this case.

features such as one-shot learning, namely the capacity to generalize from trained distributions [18]. A task that is trivial for both human adults and children, such as learning to distinguish a Segway from cars, bicycles, motorcycles, and other vehicles just by seeing it once, is a challenge for machines [40].

These are features that were found as lacking in AI also decades ago by Hubert Dreyfus. Despite the improvement in data storage and retrieval, Dreyfus was skeptical of the possibility that digital machines could simulate the "indeterminate needs and goals and the experience of gratification" which guides the determination of human needs [41: 194]. Machines lack the flexibility that humans possess in respect to variations in contextual demands. While human actions are guided by the selection of what is significantly relevant to the situation at hand, "[a] machine table of objectives […] has only an arbitrary relation to the alternatives before the machine, so that it must be explicitly appealed to at predetermined intervals to evaluate the machine's progress and direct its next choice." [41: 187]. In other words, the machine's choices are predetermined and cannot be flexibly adapted to contextual variations. This is apparent not only in the embodied interaction of the agent with the environment [41: 37, 162, 167] but also in linguistic exchanges. Natural Language Processing and Understanding continue to be challenging fields for AI research. In part, as we have seen, the difficulties met are due to the syntactic and semantic vagueness of natural language itself which we, as humans, are able to handle thanks to an extremely large set of information we possess about the world and to a capacity for insight and ambiguity tolerance that machines have not achieved, yet [41: 126-128, 204-301]. In the next section, I suggest that the recent neurosymbolic AI program is a promising path to work toward finding a solution to these weaknesses.

5 A WAY FORWARD

Current research in neurosymbolic AI seeks to combine the capabilities of the symbolic and the sub-symbolic research path in AI to address the weaknesses of both systems and, as a consequence, to achieve programs capable of more understanding and of dealing with uncertainty[11].

Symbolic AI is an approach that has been introduced by Allen Newell and Herbert Simon in the 70s. Also known as rule-based AI or Good Old-Fashioned AI, it involves the explicit embedding of human knowledge and behavior rules into computer programs. It is generally localist and discrete, including planning and nonmonotonic reasoning, and capable of extrapolation and reasoning by analogy.

Sub-symbolic AI, on the other side, is also known as Connectionist AI. Examples of sub-symbolic AI include genetic algorithms, neural networks and deep learning. The origins of sub-symbolic AI come from the attempt to mimic the human brain and its complex neural network. In contrast to symbolic-AI, sub-symbolic AI is not based on the manipulation of symbols to find solutions to problems but instead on finding correlations between input and output variables.

In the design of this state-of-the-art AI program, researchers have taken inspiration from Daniel Kahneman's dual-system theory of the human mind. System 1 is effortless, fast, and automatic and it is the one that we use the majority of the time to interact with the world. System 2, on the other hand, is slower and more rational. It is activated when we need to perform a task that requires our full attention or when we are confronted with an unexpected situation [1].

The way system 1 works is analogous to the mechanisms of sub-symbolic AI, with its ability to identify patterns and connections from a large variety of input data. Still, there are processes that system 1 can perform that instead current machine learning systems are uncapable of. For example, as mentioned above, machine learning systems are not capable of grasping basic notions of causal correlations and of common-sense reasoning, tasks that system 1 can instead achieve. On the other hand, system 2's capability to solve complex problems is more akin to AI techniques based on logic and planning and that employ explicit knowledge and symbols [43].

So far, the AI models that have been applied to creative tasks use mainly sub-symbolic techniques, since they grant programs the capacity to process a large amount of data and to find patterns within it. Yet, sub-symbolic AI has also some downsides: it requires a huge amount of data and computational power, it is highly dependent on the training data, and therefore prone to biases, and it is difficult to interpret.

Researchers have therefore started to integrate neural models with logic-based techniques to develop AI systems capable of explainability and higher-level abstract knowledge [18]. This hybrid approach is, in turn, deemed a viable way to solve some of the challenges presented by symbolic AI, e.g. brittleness, being prone to noise, and computational complexity[12].

Granted that the combination of symbolic and sub-symbolic techniques can be beneficial to build systems that achieve a higher degree of understanding, why should this matter for the discussion on the link between understanding and creativity?

The division of tasks between symbolic and sub-symbolic AI techniques and the dual-system model proposed by Kahneman reflect the two dimensions present also in creative processes and described by Joy Paul Guilford as divergent and convergent thinking [28]. Divergent thinking refers to the intuitive production of a large number of possible alternatives in response to a problem that needs to be solved, without a clear connection among themselves and between them and the original problem. Divergent thinking is, thus, based on associations between remote concepts, a key feature of creativity. Convergent thinking involves instead the constructions of relevant connections between the alternatives and their examination in respect to their logical validity. This second stage focuses on finding a solution through the application of pre-established rules.

Divergent thinking, system 1, and sub-symbolic AI are thus based on the intuitive, and almost unconscious, building of connections between remote concepts. On the other hand, convergent thinking, system 2, and symbolic AI are based on the more rational analysis of the connections found, on linear logic, and on a schematic kind of processing aimed at reaching a firm conclusion.

[11] The development of this program was at the basis of the debate in the Montreal 2019 and AAAI2020 conferences, see https://medium.com/@Montreal.AI/transcript-of-the-ai-debate-1e098eeb8465, and https://aaai.org/Conferences/AAAI-21/aaai21focusareacalls/. See also [42].

[12] There are AI models that are already using both techniques, for example [44], [45], [46], [47], [48].

In order to successfully conclude a creative process, it is necessary that both stages are present: while the association task carried out during the divergent thinking stage is necessary to generate ideas and conceptual combinations even if remotely connected, convergent thinking is necessary for the final selection of the relevant solution to the problem and it represents the evaluation stage that above I connected to the notion of understanding in creativity.

6 CONCLUSIONS

While the development of hybrid systems that combine symbolic with sub-symbolic approaches to AI may be a way forward to develop creative capacities and Agent-Understanding in artificial systems, there is still the need to work toward the achievement of Recipient-Understanding. The acceptance of processes and products achieved by AI, in fact, is dependent not only on the capabilities of the artificial systems themselves but also in the attitude and pre-conceptions that we as humans have when assessing their performance.

The notions of creativity and understanding are deeply interconnected with anthropic features. The anthropocentric interpretation of these concepts could deter us from recognizing AI performance as a result of understanding and creativity, even when it could legitimately be deemed as such. One way to avoid this consequence is to propose definitions of understanding and creativity that are not based on anthropomorphic characters, in order to avoid disciplinary biases and facilitate a broader understanding of these concepts [16].

Undoubtedly, one of the goals that state-of-the-art research is aimed to is for machines to better understand our needs and, consequently, to satisfy them more effectively. Still, I believe that an additional, and probably more relevant aim, is for humans to better understand AI and, as a consequence, to achieve a desirably better transparency and interpretability of artificial systems and their mechanisms.

REFERENCES

[1] D. Kahneman, *Thinking, fast and slow*, Macmillan, (2011).

[2] A. Boylu, How Understanding Makes Knowledge Valuable. *Canadian Journal of Philosophy*, 40: 591-609, (2010).

[3] C. Baumberger, C. Beisbart, G. Brun. What is understanding? An overview of recent debates in epistemology and philosophy of science. In: S. Grimm, C. Baumberger, and S. Ammon (eds). *Explaining Understanding. New Perspectives from Epistemology and Philosophy of Science.* Routledge, New York, (2017).

[4] R. L. Franklin. Knowledge, Belief and Understanding. *The Philosophical Quarterly*, 31: 193-208, (1981).

[5] L. Zagzebski. *On Epistemology*. Wadsworth, Belmont, CA, (2009).

[6] D. Pritchard. The Value of Knowledge. In E. N. Zalta (ed.). The Stanford Encyclopedia of Philosophy. Stanford University, Stanford, CA, (2009).

[7] W. R. Dallmayr. *Understanding and social inquiry*. University of Notre Dame Press, Notre Dame, Ind., (1977).

[8] S. Grimm, C. Baumberger, and S. Ammon (eds). *Explaining Understanding. New Perspectives from Epistemology and Philosophy of Science.* Routledge, New York, (2017).

[9] S. R. Grimm. Is understanding a species of knowledge? *The British Journal for the Philosophy of Science*, 57 (3): 515–535, (2006).

[10] M. Bachmann. The Epistemology of Understanding. A contextualist approach. *Kriterion – Journal of Philosophy*, 34(1): 75-98, (2020).

[11] A. Currie. Existential risk, creativity & well-adapted science. *Studies in History and Philosophy of Science Part A*,76: 39–48, (2019).

[12] J. Glover, R. Royce Ronning, and C. Reynolds (eds.). *Handbook of Creativity*. Plenum Press, New York, (1989).

[13] S. Mednick. The associative basis of the creative process. *Psychological review*, 69 (3): 220, (1962).

[14] A. I. Miller, *Insights of genius: Imagery and creativity in science andart.* Springer Science & Business Media, (2012).

[15] R. K. Sawyer. *Explaining creativity: The science of human innovation*. Oxford University Press, Oxford, (2011).

[16] C. Moruzzi. Measuring Creativity: an account of natural and artificial creativity. *European Journal for Philosophy of Science*, 11(1), 1–20, (2021).

[17] A. Sloman. What Enables a Machine to Understand? *IJCAI*, (1985).

[18] A. D'Avila Garcez, L.C. Lamb. Neurosymbolic AI: The 3rd Wave. arXiv:2012.05876, (2020).

[19] B. Haibe-Kains et al. Transparency and reproducibility in artificial intelligence. *Nature*, 586 (7829): E14–E16, (2020).

[20] M. Boden. Creativity and Artificial Intelligence. *Artificial Intelligence* 103: 347-356, (1998).

[21] S. D. Kelly. A Philosopher Argues That an AI Can't Be an Artist. Accessible at https://www.technologyreview.com/s/612913/a-philosopher-argues-that-an-ai-can-never-be-an-artist/, (2019).

[22] N. Collins. Automatic Composition of Electroacoustic Art Music Utilizing Machine Listening. *Computer Music Journal*, 36: 8-23, (2012).

[23] M. Elton. Artificial Creativity: Enculturing Computers. *Leonardo* 28: 207–213, (1995).

[24] L. Wyse, Lonce. Mechanisms of Artistic Creativity in Deep Learning Neural Networks. arXiv:1907.00321v1, (2019).

[25] H. J. Briegel. On creative machines and the physical origins of freedom. *Scientific reports*, 2 (1): 1–6, (2012).

[26] J. C. Kaufman, R.A. Beghetto. Beyond big and little: The four cmodel of creativity. *Review of general psychology*, 13 (1): 1–12, (2009).

[27] J. Campbell. Knowledge and Understanding. *The Philosophical Quarterly*, 32: 17-34, (1982).

[28] Guilford, Joy Paul. *The Nature of Human Intelligence*. McGraw-Hill, New York, (1967).

[29] R. Goebel et al. Explainable AI: the new 42? *International cross-domain conference for machine learning and knowledge extraction*, 295–303, (2018).

[30] A. Holzinger et al. Current advances, trends and challenges of machine learning and knowledge extraction: from machine learning to explainable AI. *International cross-domain conference for machine learning and knowledge extraction*, 1–8, (2018).

[31] C. Moruzzi. Should human artists fear AI? A report on the perceptionof creative AI. *Proceedings of xCoAx2020*, 170–185, (2020).

[32] A. Elgammal et al. CAN: Creative adversarial networks, generating "art" by learning about styles and deviating from style norms. arXiv:1706.07068, (2017).

[33] Y. Zhu et al. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. *Proceedings of the IEEE international conference on computer vision*, 19–27, (2015).

[34] J. Devlin. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805v2, (2019).

[35] S. Tehranipour. OpenAI GPT-3: Language Models are Few-Shot Learners. *Analytics Vidhya*, (2019).

[36] P. Nayak. Understanding searches better than ever before. Accessible at: https://blog.google/products/search/search-language-understanding- bert/, (2019)

[37] C. Metz. Finally, a machine that can finish your sentence. Accessible at :

https://www.nytimes.com/2018/11/18/technology/artificial-intelligence-language.html, (2018).
[38] W. Wallach, S. Vallor. Moral machines. *Ethics of Artificial Intelligence*, 383, (2020).
[39] E. M. Bender, A. Koller. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of ACL*, (2020).
[40] B. M. Lake et al. One Shot Learning of Simple Visual Concepts. Accessible at https://cims.nyu.edu/~brenden/LakeEtAl2011CogSci.pdf, (2011).
[41] H. L. Dreyfus. *What computers still can't do: A critique of artificial reason*. MIT press, (1992).
[42] J. F. Bonnefon, I Rahwan. Machine thinking, fast and slow. *Trends in Cognitive Sciences*, (2020).
[43] G. Booch et al. Thinking fast and slow in AI. arXiv:2010.06002, (2020).
[44] T. Anthony et al. Thinking Fast and Slow with Deep Learning and Tree Search. arXiv:1705.08439, (2017).
[45] R. Manhaeve et al. Deepproblog: Neural probabilistic logic programming. *Advances in Neural Information Processing Systems*, 31: 3749–3759, (2018).
[46] J. Mao et al. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences fromnatural supervision. arXiv:1904.12584, (2019).
[47] P. Minervini. Differentiable reasoning on large knowledge bases and naturallanguage. *Proceedings of the AAAI conference on artificial intelligence*, 34: 5182–5190, (2020).
[48] L. Serafini, A. Garcez. Logic tensor networks: Deep learn-ing and logical reasoning from data and knowledge, arXiv:1606.04422, (2016).

Non-Literal Usage as Operations on Meaning:
What does Philosophical Pragmatics have to offer AI?

Arthur Sullivan[1]

Abstract. My primary research question is: What does the field of philosophical pragmatics have to offer ongoing work in AI which faces certain sorts of challenges, pertaining to certain aspects of linguistic interpretation? Specifically, what help can philosophy be with designing an NLP (natural language processor) that is as pragmatically competent as a normal human speaker? That is able to 'get' a novel case of metaphor, or irony, say?

In §1, I give an overview of an approach to philosophical pragmatics. Then in §2, I engage with some of the reasons to be skeptical as to whether this philosophical grounding has anything useful to offer AI. Finally, in §3 I sketch a way in which AI can benefit from philosophy's conceptual refinements.

1 PHILOSOPHICAL PRAGMATICS

I begin by stipulating a distinction between two terms, both of which are used rather diversely within the study of language: **'meaning' vs. 'content'**. 'Meaning' will be used narrowly (in the sense of: 'literal linguistic meaning'), to designate what is semantically encoded in the expression tokened. 'Content' will be used more broadly, to designate what gets communicated with an utterance of an expression in context.[i]

Meaning can be vague, ambiguous, or under-specific in several sorts of ways ('Alice is tall' / 'Bill hurried to the bank' / 'It's raining' / 'She'll be back later'/ etc.); but I am not concerned with that sort of complication here. It's rather the **meaning-content relations** which interest me: more specifically, the systematic, nuanced processes governing the communication of contents distinct from, but related to, the meanings expressed.

And so, for example, a host uttering "It's getting late" (or anything reasonably equivalent in meaning) could communicate a desire for the guests to leave soon; or: an utterance of "It's cold in here" could serve to communicate a request to close the window; or: a (flat, understated) utterance of "That went well" could serve to communicate the judgement: This is a disaster. Etc. – **the varieties of relations between meaning and content**, as they are embodied in complex conversational interactions, are vast and diverse.

This meaning/content distinction sets up a way to theoretically orient a lot of work in philosophical pragmatics (as well as adjacent areas across the cognitive sciences): i.e., as attempts to account theoretically for the context-sensitive relations between meaning and content. Early seminal work includes Austin (1962) on perlocutionary effects, Searle's (1969) indirect speech acts, and Grice's (1975) conversational implicatures; and the tradition continues to develop and thrive. I close this introductory section with a run through of an assortment of topics in pragmatics which lend themselves to insightful meaning/content analyses, where **(pragmatic) content** can be seen as the output of a kind of **operation on (semantic) meaning**:

(i) Stern (2000) on **metaphor** (operator M):
[meaning] x **M** = content
(ii) Camp (2012) on **sarcasm** (operator SARC):
[meaning] x **SARC** = content
(iii) Kapogianni (2016), Sullivan (2019) on **verbal irony**:
[meaning] x $\mathbf{I^R}$ = content
(iv) Carpintero (2018) on **indirect speech acts** (ISAs):
[meaning] x **ISA** = content
(v) Liebesman & Magidor (2018) on **meaning transfer**:
[meaning] x **MT** = content[ii]

2 REASONS TO BE SKEPTICAL THAT PHILOSOPHICAL PRAGMATICS IS RELEVANT TO AI

Well, why wouldn't it be? How could it *not* be the case that foundational philosophical groundwork guides, aids, and abets theoretical research projects? That is how it works generally in the sciences (cf. Soames (2019)).

One main reason to be skeptical is: artificially created machines do things differently than organically evolved human minds. AI may seek to replicate human cognitive skills and capacities, but it need not do so by replicating the way in which those skills are implemented by a human mind. In fact, quite the contrary. History suggests that the way to get a pragmatically competent NLP is the way to build a successful chess-playing program: first, you junk any constraint that the machine has to, or ought to, do it in the same way as the mind; second, you identify pertinent things that the machine is better at than then mind, and maximally exploit them.

Evidently, the way to train an NLP to attain pragmatic competence is to feed it a massive annotated corpus. Pragmatic regularities get frozen into idioms, yielding a statistical probability that (say) a token of "That went well" communicates the content: this is a disaster. Essentially, these meaning/content intricacies are one and all taken to be sub-cases of ambiguity, and are solved via the same means that NLPs have for distinguishing bank_1 (river's edge) from bank_2 (financial institution). (Cf., e.g., Ghosh et al. (2015), Peled & Reichart (2017).)

Of course, to philosophers, it just seems plain wrong to say that "That went well" is ambiguous – rather, what you have is a case of **one univocal meaning admitting of multiple uses**. (Ditto for "It's cold in here", "It's getting late", etc.) Or: to focus on the classic case of an ISA, even if 99.9% of the actual tokenings of "Can you pass the salt?" are requests for salt, not for information about the addressee's

[1] Memorial University of Newfoundland, Canada, arthurs@mun.ca

capacities, that fact would not change the meaning of the expression (or of any of its constituents). But what AI might say to the philosopher at this point is: "There, there". While technically correct, the point is practically inert. AI should go with what seems to work best, not with what satisfies all philosophical scruples.

The foil, though, is novelty. That method could not possibly prepare an NLP to 'get' the first ever instance of a metaphor, or a one-off case of sarcasm or irony, in the way that a pragmatically competent speaker is able to. More generally, that rough-and-ready, case-by-case methodology lacks the kind of comprehensive theoretical grounding and integration that marks the most satisfactory of scientific explanations. In this respect, the method outlined below might point to improvements.

I next distinguish three pertinent aspects of the wide-ranging project under discussion here: (i) **conceptual geography**, (ii) **psycho-linguistics**, and (iii) **AI**. While they overlap, each has distinctive goals, methods, and desiderata. Conceptual geography seeks a satisfactory philosophical account of the target phenomenon that is integrated with other related theoretical issues and questions. Psycho-linguistics seeks to explain how the phenomenon is implemented in the human mind. AI seeks to simulate the phenomenon in a technological artifact. This distinction is important, in that AI need not be answerable to either of the other two. However, to the extent that it is not, we must be wary of claims to the effect that AI is able to illuminate philosophical or psychological questions.

3 CONVERSATIONAL MISFIT, MIND-READING ALGORITHMS, AND MEANING-OPERATING ALGORTIHMS

While computational approaches to pragmatics are not exactly in their infancy (cf., e.g., Jurafsky (2003)), still, there remains a lot of work to be done (Cummings (2013: Ch.8)). There has been moderate progress on some specific, circumscribed issues (Bender et al. (2019: Ch.13), Ghosh et al. (2020)), but I want to push the hypothesis that the approach described in §1 might help to abet some more general progress on this front.

3.1 How to interpret non-literal usage

Step 1: recognize **conversational misfit** (despite the presumption of a cooperative interlocutor, there is an incongruity between meaning and the ambient conversational context)

Step 2: determine which pragmatic phenomenon is at work, via the **mind-reading algorithms** (which are defined in terms of specific sorts of meaning/context conversational misfit – compare M vs. SARC vs. I^R vs. ISA vs. MT, etc., from the end of §1)

Step 3: run the appropriate **meaning-operation algorithm**, which outputs the content that the speaker is intending to communicate

3.2 Specific cases

For every particular specific variety of non-literal usage, there will of course be close fit between the mind-reading algorithms (MRAs) and the meaning-operation algorithms (MOAs). Take classic verbal irony[iii], as a relatively straightforward example. The MOA in this case is inversion – i.e., the speaker expresses the **meaning P**, the target output **content is ~P**. So, consider the following scenario: in the midst of an unexpected, unwanted downpour, your interlocutor says: 'What lovely weather". **Step 1**: Conversational misfit is activated, by the incongruity between meaning and context. **Step 2:** Compare the meaning to the context, to triangulate a hypothesis as to the speaker's complex communicative intentions. In this case, the MRA of ironic meaning-inversion would account for the situation. **Step 3:** The operative MOA is: [Meaning] x CVI^R = 'This weather is awful'.

The other four varieties of non-literal usage mentioned at the end of §1 involve more complex MOAs than meaning-inversion. These MOAs are nuanced, flexible, and highly context-specific. The amorphousness of metaphor has long been recognized; cf. Nunberg (1995) and Camp (2012) for the voluminous varieties of meaning-transfer and sarcasm, respectively; and there are no principled bounds on what can (in context) be the content of an ISA. The above point about the fit between MRA and MOA, in each such particular sub-case, will still hold – as would be predictable via more or less any plausible account of the relations between meaning, content, and (cooperative) speakers' communicative intentions. The three-step model will apply across a wide range of MRA-MOA pairings.

Thus, as linguists and philosophers get progressively closer to proper characterization of these MOAs, this will afford MRAs which are available for implementation by AI researchers.

4 CONCLUSION

While I propose to apply a certain common methodology uniformly across a variety of communicative phenomena, I am not assuming here at the outset that they constitute a homogenous psycho-linguistic 'natural kind'. There may be distinct cognitive mechanisms or capacities at work, across these varieties of non-literal usage – let alone when we turn to various others, such as **metonymy** (e.g, 'Hollywood is obsessed with this trend'), **synecdoche** (e.g., 'I can't pick you up because I don't have any wheels'), **personification** (e.g., 'The thermostat wants more information'), etc. It would seem reckless to assume that there is any one general explanation as to exactly how they are all processed by competent speakers. Nonetheless, this approach is pitched at a level at which (despite lots of potential differences in the implementing mechanics) there is a common structure to these diverse communicative phenomenon. Conceptual geography and AI may well line up here, despite their possibly falling out of step with psycholinguistics.

ACKNOWLEDGEMENTS

This paper was prepared in the stimulating environment of the Rotman Institute of Philosophy at Western University, and is based on research supported by an Insight Grant from the Social Science and Humanities Research Council of Canada. Many thanks to Jiangtian Li and Rob Stainton for very helpful conversations.

REFERENCES

[1] Austin, J. L. (1962) How to do Things with Words. Oxford University Press.

[2] Bender, E., Lascarides, A. & Hirst, G. (2019) Linguistic Fundamentals for Natural Language Processing II. Morgan & Claypool.

[3] Carpintero, M. (2018) "Sneaky Assertions". Philosophical Perspectives (32): 188-218.

[4] Camp, E. (2012) "Sarcasm, Pretense, and the Semantics/Pragmatics Distinction". Nous (46): 587-634.

[5] Cicero (55 BCE) De Oratore. Project Gutenberg, NetLibrary.

[6] Cummings, L. (2013) Pragmatics: A Multidisciplinary Perspective. Taylor & Francis.

[7] Ghosh, D., Guo, W., & Muresan, S. (2015) "Sarcastic or not: Word embeddings to predict the literal or sarcastic meaning of words". Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing: 1003-12.

[8] Ghosh, D., Musi, E., Upasani, K., & Muresan, S. (2020) "Interpreting Verbal Irony". Proceedings of the Society for Computation in Linguistics: 76-87.

[9] Grice, H.P. (1975) "Logic and Conversation". In Studies in the Ways of Words. Oxford University Press, 1989.

[10] Jurafsky, M. (2003) "Pragmatics and Computational Linguistics". In Horn & Ward, ed., Handbook of Pragmatics, Blackwell.

[11] Kapogianni, E. (2016) "The Ironic Operation". Journal of Pragmatics (91): 16-28.

[12] Liebesman, D., & Magidor, O. (2018) "Meaning Transfer Revisited". Philosophical Perspectives (32): 254-97.

[13] Nunberg, G. (1995) "Transfers of Meaning". Journal of Semantics (12): 109-32.

[14] Peled, L., & Reichart, R. (2017) "Sarcasm Sign: Interpreting sarcasm with sentiment based monolingual machine translation". Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics: 1690-1700.

[15] Searle, J. (1969) Speech Acts. Cambridge University Press.

[16] Soames, S. (2019) The World Philosophy Made. Princeton University Press.

[17] Sperber, D., & Wilson, D. (1986) Relevance. Cambridge University Press.

[18] Stern, J. (2000) Metaphor in Context. MIT Press.

[19] Sullivan, A. (2019) "The Varieties of Verbal Irony". Lingua (232): 1-22.

[i] There are various other terms that do similar work, in the literature – for example, Grice's (1975) "what is said"; vs. "speaker's meaning", or Sperber & Wilson's (1986) "encoded meaning"/"explicature"/"implicature". I am trying to steer clear of any pre-existing associations and theoretical baggage which these terms will carry.

[ii] Some concrete examples: (i) **metaphor** – 'Juliet is the sun'; (ii) **sarcasm** – 'Like she'd ever go out with you!'; (iii) **irony** – 'What lovely weather!'; (iv) **indirect speech act** – 'Can you pass the salt?; (v) **meaning transfer** – 'The ham sandwich left without paying'. (Note that while both 'sarcasm' and 'irony' tend to be variously used, and overlap significantly on most any usage, there is growing consensus that we should countenance both non-ironic sarcasm (e.g., skater-teen use of 'Fail!' when a rival messes up) and non-sarcastic irony (e.g., a use of 'What awful weather!' on a lovely day) – *cf.* Sullivan (2019: §1.5) for discussion and references.)

[iii] By 'classic verbal irony (CVI)', I mean the trope in which the content is the inverse of the meaning. The first explicit discussion of the notion might be Cicero in 55 BCE ("De Oratore", Bk. II). Following Sperber & Wilson (1986), most contemporary theorists countenance various other (subtler, more nuanced) sorts of verbal irony, beyond meaning-inversion – *cf.* Kapogianni (2016), Sullivan (2019), Ghosh et al. (2020).

Inevitably falling into place: Intuition as a function of spatial prediction

S. de Jager [1]

Abstract. The concept of *intuition* is often yet vaguely employed to signify that which exceeds the bounds of *rationality*, as well as that which plays a functional, regulative role behind effects such as *creativity*, *instinct* or *improvisation*. Often treated as an unconscious or inevitable cognitive activity which stands on the opposite end of the spectrum to (dia)logical reasoning: intuition is usually assumed to be a holistic and improvised reaction to a problem where immediate decision-making is required. In this paper we will argue that the concept of intuition can be understood as a combination of *spatial reasoning* and *predictive processing*: two interrelated; enactive interpretations of perception and action which can be considered as underlying our use of natural language, and hence the use of (dia)logical reasoning as well. We reject the reductionist, dual interpretation of intuition versus non-intuition, and conclude by speculating about the significance of intuition in the use of metaphoric reasoning. In order to trace this conceptual development we propose an account which conceives of predictive cognition as an act of *falling* forward into time as a spatial dimension. The hope is to generate new insights relevant to discussions concerning *common sense* and *explanation* in data science by questioning the divide between reason and intuition.

1 Outline

Before diving into the technicalities of the argumentation, we begin with a an allegorical introduction guided by the question: *What does falling have to do with thinking and computation?* The purpose behind this introduction is to help visualize some of the conceptual mechanisms which relate intuition to data science, as we propose an account which conceives of spatial prediction as an act of *falling*.

In the *Background* section that follows after that we provide some brief background to the concept of intuition in order to contextualize the rest of the arguments that will be presented.

In the third section, we organize our arguments around the rational and the intuitive by considering the difference between making a decision and experiencing the inevitable, the *inevitable* being constraining factors not only in the physiology of agents but also in the structure of their given environments.

Following from this, in the fourth section we explore the case of predictive processing (PP), and its indebtedness to a what can be considered a Hegelian constructivism. We will also analyze specific examples from spatial reasoning which experimentally demonstrate some of the concepts at hand. The speculation we will explore is that underlying the pursuit of new knowledge is the tacit expectation of stability (in PP attributed to free-energy minimization) and the assumption of a socially-distributed commonality, the feeling that "others experience this too," a defining feature of intuition in collective intentionality.

Finally, as further lines of exploration we propose how a concept such as *Absolute Spirit* [Hegel]—whose genesis is aimed at exploring the productive constraints of reason and thus the bounds of the context of discovery—functions as a metaphor for collective intentionality in order to further elucidate the concept of intuition and its key role in the development of explanations. We will also speculate about whether or not to demand a description of intuition—at least at the macro-level: e.g. that of philosophy, as opposed to micro-level of cellular homeostasis—is an effective step towards the pursuit of novel insights into cognition and its explanation.

2 What does falling have to do with thinking and computation?

> However, it is also possible to suppose that becoming is a dimension of being corresponding to a capacity of being to **fall out of phase with itself**, that is, to resolve itself by dephasing itself.[2]

If we take philosophy seriously, *everything has to do with everything*. In what follows, seemingly disparately, we seek to combine and contrast the act of falling with the processes of (intuitive) thought and non-human computation. With the help of predictive processing and spatial reasoning, we seek to provide an account of intuition as essentially indistinguishable from other forms of reasoning (such as propositional or logical knowledge). This consideration is crucial now—in the context of the desire and/or need for artificial intelligence—seeing as one of the most sought after formalizations in current AI is the development of so-called "common sense,"[3] primarily in spatial terms, as in: an agent knowing how an object will behave when picked up, handled, and what the possible uses for it are. But also in terms of understanding natural language and knowing what to do when a human asks the AI agent to do something [26].

"Common sense" is, in fact, very close to what "intuition" is about. According to John McCarthy "[a] program has common sense if it automatically deduces for itself a sufficiently wide class of immediate consequences of anything it is told and what it already knows" [25]. What is "automatically deducing for [one]self" if not the characteristically defining function of intuition? Intuition has been proposed to have a variety of meanings, and we will expand on this later, but for now let's frame it around the sense of looking at a

[1] Erasmus University, Rotterdam, Netherlands, $dejager@esphil.eur.nl$

[2] G. Simondon, *The position of the problem of ontogenesis*, tr. G. Flanders, 2009, p. 6, my emphasis in bold. [31]

[3] Related and crucial to the development of so-called *artificial general intelligence.*

bunch of objects, and then looking at a box, and "feeling" whether all that stuff will or will not fit in that box. When we do this, we are, indeed, *automatically* feeling our way around.[4] This applies to the realm of symbolic-trading as well, since we employ a similar search-strategy when we try to understand the meaning of a never-before-encountered term or phrase, but more on this later as well.

This more or less inevitable feeling—inevitable because it "happens" to us rather than us properly understanding how we bring it about—can be understood a spatial prediction: we are predictively intuiting how something will behave. This happens internally and relates to the future of our relationship with the external world. Following Piaget in his exposure of causal learning [Piaget 1930], most of us would agree that it's a process that has its origins in our early childhood encounters with the world, and grows very slowly as one learns to move about through space and interact with objects. Alan Turing already said, many decades ago, that if we want to understand intelligence we need to understand developmental intelligence [22]. The prediction of space is a crucial aspect in the development of intelligence: the more spatial constraints an agent is able to overcome, or work with, the more 'intelligent' it should be considered (at least according to researchers like McCarthy, as evidenced by the quote above).

Things being handled, molded, stacked, moved, balanced, shuffled, etc.: all of these activities lie at the basis of what we normally consider to be "higher order" cognitive tasks, like doing mathematical calculations or writing a philosophy paper. Developing efficient ways of dealing with a spatially-challenging environment is the first and foremost evolutionary progression of all species: resilience, agility, balance, speed, etc. But first and foremost beginning with the fact that a defining characteristic of what we consider 'life' is that it separates itself off, through an membrane enclosure, from the rest of the world: the genesis of the concept of life is the genesis of the concept of space.[5] Dealing with space in the context of existing on a planet is unavoidably dealing with *gravity*: unintentionally falling, letting yourself fall ever so slightly, or avoiding a fall. What is walking if not a series of gravitational events where one continually falls forward? Life emerges in a universe bounded by gravity, and so are our thoughts and the metaphors we evolve: we fall in love, things fall through the cracks, we fall to pieces, rise and fall, we fall prey to something, we fall asleep, fall behind, fall through, fall apart, fall ill or fall silent. What is free association if not a process where we let the brain fall recursively into itself? Things fall into place by the very nature they possess: a nature bounded by gravity. Gravity is space-time curved, ergo, lo and behold: *everything is space*.

But let's not fall too deep into the poetic zone and return to AI. A prominent algorithmic method used for teaching neural networks how to teach themselves, so to speak, is the process of *gradient descent*. (Stochastic) gradient descent is an optimization algorithm which has the goal of finding an efficient way around a function. Efficiency here means that we get acceptable accuracy—whatever our parameter for this may be—at a relatively low cost, i.e.: the shortest amount of time possible, i.e.: dealing with more space by predictively simplifying it. Picture a graph with a lot of data points and a line running through all of them. The algorithm runs following the depth of the sharpest descending points, and so develops another line that closely fits the pattern produced by the dots. The "easiest way around it" is a line that generalizes an average mid-point of all those dots, ideally. By way of additional intelligent algorithmic design, what we end up with are generative input-output mechanisms that can average stuff: you feed an artificial neural network a bunch of cat pictures and after a while it will be able produce an average of a cat picture based on all the gradient-descending (and backpropagation) it's undergone, based off of a data set of real cat pictures. By falling, by following descending gradients, artificial neural networks can be considered to be engaging in something which becomes comparable to what the brain engages in when it estimates or intuits, according to a lot of research, especially in the field of Bayesian inference and neurocomputation [1]. Following patterns, traveling the path of least resistance, is a way of falling; a way of thinking; a way of computing. Be mindful that I am not saying these things are the same, I am saying there is an interesting comparison to be made here: no two things are the same, but all things are *comparable*.

What is very interesting for our purposes is that a spatial analogy of somebody moving through space is one of the main examples that is employed to explain gradient descent in a less technical way (even though the explanation above also attempted to drastically minimize the technical aspects of the matter).[6] The metaphoric explanation goes like this: let's say someone is happily climbing up on a sunny mountain and suddenly the weather changes, it starts to get foggy and cloudy and there's not much visibility anymore. Space becomes smaller, in a way, the choices one can make are limited. The person, if they want to get down, needs to rely on very little; immediate information in order to start descending: looking left and right trying to find the steepest paths to be able to get down as fast as possible because, let's say, it's gonna rain soon. Literally moving by steepest-gradient-descent, the person would eventually find their way down. Gradient descent in practice is mathematically more complex than this, but we get the point. The philosophically interesting commentary about this is that for *understanding* to happen, for the grasping of a concept to happen, we need to employ a spatial metaphor. One visualizes how going down a mountain is comparable to gradient descent in order to understand how a perceptually-inaccessible algorithmic process happens. Thought unfolds spatially once more, and in more directions than one: literally and metaphorically speaking, which are, of course, interrelated.

Hopefully, by now it will be clear how falling, thinking and computation can be seen as related to each other. Agents—let's call them agents for the sake of simplifying humans and possible artificial intelligences into one category—if they process information in a predictive fashion, need to have internalized knowledge of how they "fit into" space, in order to act. However, it is also very important to mention that spontaneity and randomness are also crucial aspects of thinking and acting,[7] less effectively predictive because they require that there is no (or little) knowledge about what comes next, but this is cognitively; creatively speaking just as fundamentally important. Perhaps we could say that intuition is a cognitive state where prediction is being attempted but there is still quite a lot of uncertainty in the picture, like when we descend down a foggy mountain. The metastable faculty of intuition emerges from the real-life limitation of being born into a world of gravity: there is a certain resistance,

[4] This statement demands, of course, a definition of 'automatic.' We will explore this later as we look into the concept of 'autonomy' with regard to the allocation of agency and the distinction between agent and environment.

[5] And as we will argue later, the genesis of whatever it is we call intelligence is the genesis of dealing with space in terms of time: remembering and projecting.

[6] Coincidentally interesting is the fact that Wikipedia explains this by saying that "The basic intuition behind gradient descent can be illustrated by [the following] hypothetical scenario." A sentence which happily joins intuition, falling, computation and explanation: the very topics of this paper. "Gradient descent" Wikipedia entry, accessed March 1, 2021.

[7] This couldn't be more clear in our stochastic gradient descent example.

a negativity that needs to be overcome, and agents overcome it by positively falling forward, flying, floating, etc. instead of neutrally staying put. An intuitive prediction is the connection between two or more (metastable) spatial models: the one(s) an agent are at, and the one(s) an agent is inclined to be at. Intuitive prediction, what we do when we think and move, is thus driven by the combination and recombination of spatial speculations.

How is this connected with so-called "higher order" cognitive faculties, like "pure thought", or propositional logic, in isolation from spatial events? Well, first of all nothing can be said to be isolated from spatial events, and this is one of the biggest shortcomings in conceptualizing what should or should not be considered as "thought." The naïve-realist inability to appreciate thought in action shows itself most clearly in language model projects that map 1-to-1 correspondence between words and their descriptions [3], or the projects that compile highly complex databases of 'common sense facts' or propositional syntax [23]. These approaches, if we may freely speculate, have certainly been inherited from the categorizing drive of the Enlightenment, which we will discuss later as well. This database or encyclopedic approach, which considers all capital-letter *Thought* proper as happening exclusively in heads and books, is ethically-speaking rather unfortunate. It has, for example, been one of the main reasons why "manual" labor becomes less remunerated than "desk" labor, or why all sorts of spatial activities that leave no paper-trail are not considered work [6]. But that's another subject.

Let us consider propositional thoughts, such as the definitions of abstract concepts. Think about how one acquires concepts. Let's start with a very simple concept: truck. First you probably learned about cars. Then, after seeing a bunch of cars and suddenly seeing a truck, you understand it's a sort of car, but not exactly. It has more space, it carries more stuff. Then, later on in life, you hear someone say "truckload," as in "that paper Sonia wrote was a truckload of garbage." You have never heard the expression "truckload" but you know what a truck is, and that it carries big loads: you understand immediately that we're talking about a lot of garbage. Later on in life, when you hear the expression again, you don't need to think about trucks or loads anymore, the word-container of "truck-load" becomes a summary that quickly points you to an estimation of something being "a lot"—in this case particularly aggressively so: we know trucks do heavy duty work and carry large quantities of stuff. Additionally, the combination of the words "garbage" and "truckload" certainly reminds you of your experiences with both garbage and trucks, and thus we understand we're talking about a whole lot of crap. You see? Another spatial expression which creates a thought, based on your previous life experiences with the spatial stuff of *crap*.

As famously explored by Lakoff and Johnson [20]: the establishment and constant reassociation, at different levels, of these summarized conceptual patterns results in conceptual cognition. Analogical thinking, which includes metaphorical thinking, is an essential aspect of abstraction which emerges from rather simplistic parameters [17]. In a recent analysis by Jeannette Littlemore, specific aspects of metaphors and their emergence from an embodied perspective are explored towards the thesis that "the role played by embodied cognition in metonymy creation is influenced by the presence of **movement and emotion**," [13] my emphasis. In her even more recent book *Metaphors in the Mind: Sources of Variation in Embodied Metaphor* [21], she explores how it can be almost universally stated that the common language that humans use to conceptualize is that of metaphor [p. 19], and how the interpretation of metaphors (and thus concepts) varies across the experiences of different *bodies*, or in our terms: different spatial configurations and thus differently predictive perspectives.

We now continue with an exploration of the concept of intuition, especially in light of how it has often been positioned: as an activity separate from (dia)logical thought, in order to move towards an understanding of both activities as one and the same.

3 Background

What is intuition and why is it important? According to the influential decision-making scholar Robin M. Hogarth "the major challenge facing intuition research is the need for conceptual work to define the nature and scope of different intuitive phenomena" [Hogarth 2010 p. 338]. Hogarth also remarks that in order for it to be useful, the concept should not become too broad [ibid.].[8] In necessarily broad strokes, however: intuition is currently understood as the non-propositional and/or even unconscious capacity to *feel* one's way out of something, and opposed to the daunting task of 'conscious thought' proper where one *thinks* oneself forward by way of logical steps. Interpretations vary, but intuition is sometimes related to instinct, and sometimes differentiated from it by referring to instinct as entirely inaccessible to 'conscious thought' (e.g. muscular reflex).

Notwithstanding the oscillating philosophical meaning of intuition over time, traditional philosophical distinctions between *nous* and *logos*, *ratio* and *intellectus*, *Verstand* and *Vernunft*, could be interpreted as having given way–via the modulation of capital **R** Reason by the categorizing drive of European Enlightenment thought–to 20th century cognitive scientists postulating comparable *dual process theories* [Chaiken and Trope 1999] that distinguish between intuitions and analytical,[9] or dialogical[10] processes. Some examples are: "System 1" and "System 2" [Stanovich and West 2000], "experiential" and "rational" [Epstein 1994], "tacit" and "deliberate" [Hogarth, 2001], thinking "fast" and "slow" [Kahneman 2011], etc. Many scholars have proposed a reconsideration of dual-process theories as overly simplistic or inconsistent, ranging from Gerd Gigerenzer's criticism that "Dual-process theories of reasoning exemplify the backwards development from precise theories to surrogates" [11] to Kahnemann himself, stating early on in his 2011 book that "System 1" and "System 2" are but useful fictions [18]. In the present paper we do not question their fictional character, which we consider to be obvious, but we do question their usefulness.

Despite Herbert Simon's remarkable influence in the realm of decision-making, and his proposal that the dualistic interpretation of 'thought' versus 'intuition' is a fallacy [29], scholars of all fields continue to contrast propositional analysis with intuition, as well as variations on this theme. While we will not provide an account of dual-process theories, they need to be mentioned as the background drive behind the present paper: what drives the need to distinguish between these two modalities as opposed to each other? Some scholars—for example Keith Stanovich himself—acknowledge the critique but propose, for example, that results stemming from experimental cognitive science have become evidence to the usefulness of the distinction in dual-process theories, particularly if the distinction is considered to be 'autonomous' versus 'reflective' response [7]. However,

[8] It is important to already note, however, that the concept carries this broad vagueness about it because it generally denotes an ineffable, impalpable effect of cognition. As we will observe later when analyzing, for example, *abduction* and *Absolute Spirit*, the concept's generality can, in many contexts, be considered its strength.

[9] Relating to logic; truths and contradictions, often not accommodating ambiguity.

[10] Relating to dialogue; discussion; disambiguation, whether social or internal.

if taken to its pragmatic consequences, continued insistence on this distinction would imply dramatic results in all areas relating to the adjudication of decision-making and agency, ranging from psychiatry to the judicial system, education and eventually AI. Together with Simon, we find the distinction between intuition and non-intuition unsatisfactory: it divides the activity of thought into two arbitrary categories which more often than not results in the privileging of one over the other. That is the starting point and main claim of this paper: intuition and rationality are one and the same thing, i.e: spatially-predictive thought, guided by an agent's constraining parameters: the contingency of the environment, and the agent's internally-driven outward exploration hereof.

We will analyze a variety of different definitions of intuition without necessarily searching for a resolute formalization. It is also important to mention that the ambitions of this paper are rather expansive, as it discusses the overlaps between a variety of approaches: from philosophy to discourse analysis; neurobiology; artificial intelligence and beyond. The desire is to offer a comparative perspective that connects similar patterns between these disciplines. This is not only necessary if we wish to move forward multidiciplinarily, but also to clarify concepts and terms across specializations. Despite the comparability of a certain 'duality-bias' across domains, there is a large gap between philosophical and cognitive science contemplation on intuition. For example, the *Stanford Encyclopedia of Philosophy*'s entry on intuition opens with a few examples where the faculty is said to play a role, such as "it is impossible for a square to have five sides," where one should intuitively understand said impossibility. In our consideration this is an analytic statement, which necessarily employs intuition, but it cannot be said to produce intuitive cognitions in isolation from (dia)logical propositional knowledge.[11] The author of the entry acknowledges that the "clear that ordinary usage [of the word] includes more in the extension of the term "intuition" than such states," [ibid.] but does not deem this type of intuition to contain epistemic, distinctive philosophical relevance. For clarity: the *SEP* defines intuitions as "mental states or events in which a proposition *seems true in the manner of these propositions*." The *seeming* part is interesting to us here; as it signifies a certain instability in these knowledge claims. However, for our purposes, this *seeming*, especially in collectively intentional contexts, plays a much more prominent role than we tend to acknowledge. Peter van Inwagen, mentioned in the *SEP* article, explains that intuitions might be better understood as inclinations or tendencies that make beliefs attractive, "that 'move' us in the direction of accepting certain propositions without taking us all the way to acceptance" [32]. This interpretation still separates belief from intuition, but does get closer to our interpretation, which highlights the spatial dimension of 'moving towards,' or indeed: *falling* into place by way of intuiting. As we will continue to argue, there is not very much to gained from separating beliefs from intuitions, given that they are too interrelated to be categorically distinguished from each other.

The present paper hopes to expose a philosophical interpretation of the contemporary "clear and ordinary usage" of intuition, observing its importance across various scales, from the homeostatic processes of amoeba all the way up to the linguistic evolution of social groups: an ambitious taxonomy made possible by framing (dia)logical thought and intuition as mutually-constituting spatial predictions. In the case of the 'five-sided square' we find a striking example of our capacity to spatially-predict, as we elicit the imagination of said square, and we intuit or "picture" a hexagon instead. Depending on how one chooses to imagine—or interpret—said statement, this interpreting activity can thus be a variety of things contingent upon how much time one chooses to spend pondering over it and how many times one has performed similar actions in the past. It is thus both a quickly-grasped analytic statement, because—as an adult—one need not picture a square or a hexagon to understand the statement (an example of (dia)logical reasoning), but also a spatially-predictive process (in case we do picture it, as we are internally contrasting squares with hexagons, or numerical quantities with each other). Both of these are an inevitable search for resolution (as we 'feel' our way through the propositional spatial prediction in search of an outcome). What is important to note is that these activities do not oppose each other or occur as 'either or,' they are modulations of the same thing–i.e.: finding one's way about space–and they can certainly overlap [for a similar claim about self-deception and multiple-beliefs happening at once [28]]. This proposal is indebted to *4E* accounts of cognition, where propositional knowledge is said to emerge from embodied knowledge.[12] Our proposal is that these modalities *need be epistemically circular*, much in the same way that one hand washes the other; there is no primacy in either hand, only a self-referential, recursive event. Whether the content of intuitions is "correct" or "erroneous" is not of interest to us, and beyond the scope of our inquiry as we would require a definition of *truth* and *error*.

Some additional notes must be made with regard to the history of the term *intuition*, albeit brief ones. Immanuel Kant famously emphasizes in the *Critique of Pure Reason* (CPR) that our intuitions (the unfortunate English correlate to *Anschauungen*) are not just received, but are co-created by the faculty of intuition itself (*Anschauungsvermögen*).[13] According to predictive processing scholars Wanja Wiese and Thomas Metzinger, Kant is credited as the first philosopher to have proposed the idea that objects *conform* to our perception, and not the other way around. As *Anschauung*, intuition refers to *immediate sense-perception*, at its most fundamental: the experience of spacetime. Even if *Anschauungen* in Kant refer to sensory contemplations, we could still draw their connection to the present-day definition of intuition[14] as "an ability to understand or know something immediately based on your feelings rather than facts,"[15] provided that the term implies immediate recourse to one's present situation, given sensorial state, and access to short-term memory [8].[16] Since Kant's distinction between thinking and sensing in the CPR [A 51/B 75] was an elementary foundation of his entire philosophical edifice, this brief consideration of his concept of intuition is necessary in order to elucidate some aspects of how the categorizing drive of the Enlightenment is perhaps what led to the aforementioned dual-process approach to cognition.

Kant's categorical approach to the bounds of our rational capacity

[11] Where one converses internally with oneself producing or exploring 'justified true belief', the opposite of an intuition, according to the *SEP* entry.

[12] *4E*: Where all cognition is interpreted as being "materially **e**mbodied, culturally/ecologically **e**mbedded, naturalistically grounded, affect-based, dialogically coordinated, and socially **e**nacted. (Thibault, 2011: 211). The additional **e** is attributed to mind as *extended* across the physical realm beyond the body, as exposed in the "extended mind hypothesis" by Andy Clark and David Chalmers, 1998.

[13] "Wenn die Anschauung sich nach der Beschaffenheit der Gegenstände richten müßte, so sehe ich nicht ein, wie man a priori von ihr etwas wissen könne; richtet sich aber der Gegenstand (als Objekt der Sinne) nach der Beschaffenheit unseres Anschauungsvermögens, so kann ich mir diese Möglichkeit ganz wohl vorstellen." (Kant 1998[1781/87], B XVII) [19]

[14] Which stems from the mid-fifteenth century, *intuicioun*, denoting "insight, direct or immediate cognition, spiritual perception," an originally theological term from Late Latin *intuitionem*: "a looking at, **consideration**," from *intueri* "look at, **consider**." [Etymonline.com, accessed 02 July 2020], my emphasis in bold.

[15] *Cambridge English Dictionary*, online version, accessed 02 July 2020.

[16] And, not unimportant: when is one ever *outside* said state?

was expanded by Hegel's insistence that the rational exceeds the confines of the human membrane and exists *out there*: contradiction—i.e. dialectical process—is the very stuff of reality, and does not simply reside in the finite cognitive capacities of—wealthy, male, white, etc.—human beings.[17] This observation is implicit throughout his philosophy and most famously expressed by his laconic dictum: "What is rational is real/actual; And what is real/actual is rational."[18] This major conceptual event serves to illustrate the 4E thesis of this paper: given a cognitive agent's dependence on its physiological and environmental configuration in order to make decisions, the effects of cognition—whether intuitive or rational, if so polarized—cannot be said to reside within a finite command and control unit (i.e. skull-bound rationality) but are better understood as distributed over social, symbolic and material networks of constraints. Hence the author's literary decision that these developments can be poetically allegorized as a process of things *inevitably falling into place*: the cognitive is a recursive fractal cascade resisting the pull of gravity, but inevitably driven towards it.

To recapitulate, we acknowledge the contemporary use of the term 'intuition' as the one conceived by Hogarth in his 2001 and 2010 papers: "the essence of intuition or intuitive responses is that they are reached with little apparent effort, and typically without conscious awareness" [Hogarth 2001, 2010]. However, whereas Hogarth notes intuitions "involve little or no conscious deliberation" [ibid.], we propose to explore them as *metaconscious, hyperdeliberative* phenomena. *Metaconscious* because these processes refer consciousness to itself (as we just saw in the case of the square-hexagon: prediction is a matter of self-relation, by eliciting the power of the imagination/memory), and thus go beyond what is commonly understood as consciousness,[19] but also because consciousness does not "start" where the unconscious begins (the autonomic nervous system is just as much a part of consciousness as is the consideration of the square-hexagon). *Hyperdeliberative* because they involve an organism, which, minimally defined as something which reacts to the environment from the perspective of its enveloped condition; becomes something which therefore necessarily, constantly *chooses*. This eternal enveloping condition of choice—and the equally important inability to choose—is what we choose to metaphorize as the inevitability of *falling into place*. It does not matter whether we give primacy to the agent's own drives or the constraints presented by the environment, both represent the effects of gravity conditioning the evolution of agents in space. The relevance of all this to evolving our understanding of developments in data science is that it explores the supposed distinction between rule-following (what computers running algorithms supposedly do) and creative thinking, or: rule-inventing (what humans designing algorithms supposedly do).[20] The comparability between these two as "opposites" is what is at stake in a definition of intelligence (as Alan Turing showed in "Computing Machinery and Intelligence" in 1950 [22]), and this is precisely what we are putting into question.

[17] Though Hegel failed to understand certain intersectional powers of privilege in his account.

[18] 'Vorrede: Was vernünftig ist, das ist Wirklich; und was wirklich ist, das ist vernünftig." G. W. F. Hegel, *Elements of the Philosophy of Right*, 1821 [14].

[19] An awareness of one's own spatiotemporality, sometimes coupled with an inner (logical or abstract sort of) dialogue. [Thomas Metzinger (2015). M-Autonomy. Journal of Consciousness Studies, 22 (11-12).]. As we will observe later, we have to assume consciousness much in the same way that we have to assume existence: the questioning thereof seems to lead to cul-de-sacs and dead ends rather than bifurcations or at least roundabouts.

[20] And relatedly: the P/NP problem (more on this later).

3.1 Herbert Simon's productive constraints

The 20th century quest for (machine) rationality has revealed something to both philosophers and data scientists alike: that the search for unbounded, frictionless decision-making is rather elusive, and that human (and other organic) beings are actually good decision-makers because of the fact that their 'computing' power is constrained, restricted. Herbert Simon famously introduced his concept of *bounded rationality* as the inability to resort to 'real' rationality, since agents are unavoidably bound to spatiotemporal constraints [1940]:

> "It is impossible for the behaviour of a single, isolated individual to reach any high degree of rationality. The number of alternatives he must explore is so great, the information he would need to evaluate them so vast that even an approximation to objective rationality is hard to conceive."[21]

Simon answered to the 'problem' of boundedness with purported human capacity to *satisfice*,[22] which approximates the function we often assign to intuition, but not exactly. While an agent may not be able to offer any grounds for their intuitions, *satisficing* implies that it is preceded by a logical assessment. Simon did consider intuition extensively as well, providing one of the most influential accounts in the field of (economic) decision-making. In his book *Reason in human affairs* he defined intuition as subconscious information-processing in the service of pattern recognition [1983]. His impetus was to demonstrate how the given experience of a decision-maker could effectively lead to intuitive insights, the more extensive said experience was. 'Adequate knowledge' would thus arise over time, based on habituation, and certain structures would persist depending on whether they prove successful or not [Prietula and Simon 1989]. In a Spinozan fashion: for Simon, intuition and analytical judgment are not two distinct types of information-processing, they are rather "frozen into habit and into the capacity for rapid response through recognition" [29], According to Simon: "intuition is not a process that operates independently of analysis; rather the two processes are essential complementary components of effective decision-making systems" [30].

Herbert Simon's work inspired plenty of new heuristics in decision-making, among which Naturalistic Decision-Making (NDM). NDM scholar Gary Klein states that the NDM community defines intuition—almost in the same words as Simon—"as based on large numbers of patterns gained through experience, resulting in different forms of tacit knowledge" [Klein 2015]. Another interpretation in decision-making is the Heuristics and Bias approach (HB), according to which agents usually rely on irrational beliefs or biases in order to satisfice. While rigid in our interpretation—because they seek a formalization—a combination of these perspectives does approximate the picture of intuition we are after, with the added connotation that the concept in everyday language also implies a certain quickness, volition, urgency. Indeed, as Hogarth states: "intuitive judgments are typically—but not always—correlated with speed and often a sense of confidence" [2002]—or as Gerd Gigerenzer[23] would have it, they are 'fast and frugal.' Intuitions can build up over time, but the experience of having an intuition is phenomenally automatic, inevitable. As an example: I may not be able to access the exact reason why I do not trust someone or something, but the intuition of said distrust feels as if this happens automatically, and it influences my decision-making around that person or thing. Relatedly, it

[21] Simon 1940.

[22] Satisfy and suffice.

[23] Bounded rationality scholar and colleague of Simon's.

should be said that an intuition differs from an emotion in the sense that an emotion can sometimes be a reaction with no external consequences, whereas an intuition is a reaction geared towards future (short or long-term) actions. In Peircian terminology: while an emotion would be firstness, an intuition would be secondness, and very often also thirdness, as we're trying to argue.

Following Simon, intuition as the capacity to receive knowledge that cannot be directly traced back to a specific observation or proposition, does not mean, however, that intuition is an independent source of knowledge without ties to observations and propositions. Intuition is an amalgamation of a large variety of general impressions and previously encountered propositions (inferences, previously embodied states: gathered habits). For the purposes of this paper: an intuition is an intuition in the etymological sense of it being a contemplative act; in the (originally Spinozan) sense of it existing on a spectrum of deterministic cognitive phenomena; in the Simonian sense of it being an accumulation of patterns and habits through time; and in the PP sense of it being an inference biased by our predictive spatial condition (inherited from the Kantian sense of it being always-already affected by our perceptual capacities). We will focus specifically on the latter two senses, but all these meanings should loom in the background connecting experiential dots into intuitive patterns.

4 Making decisions or experiencing the inevitable

"What is done and what is undergone are... reciprocally, cumulatively, and continuously instrumental in each other."[24]

It would be strange to say that we rarely engage with intuitive scenarios. As a matter of fact we could dare say that, on a daily basis, we entertain intuitions more often than we do not. So-called *gut-feelings*, for example, are efficient: decisions can be made faster in an *umwelt* that demands fast decision-making. Organic beings resort to this *unmovable mover* type of action, lest they become metastable forever before two equal stacks of hay.[25] Despite the prominence of this capacity, however, when grappling with questions about consciousness, rationality, knowledge, and various other formulations of *intentional cognitive action*,[26] philosophy and data science alike often take the ample repertoire of underlying conditions and behaviors that what we would normally deem 'intuitive' for granted. This is not only because they pretend to deal with the descriptions of 'ideal,' symbol-trading, non-pathological decision-makers, but also—as Henri Bergson famously observed—because of the mirage generated by a specific finite metaphysics and a tendency towards perceptual transcendentality. To name a few of these assumptions: an agent's implicit self-sustaining mechanisms (i.e. homeostasis, physiological set-up), the explicit persistence or stability of their drives through time (i.e. self-perception, opinions, etc.), their capacity to perceive both an inner and an outer frame of reference, and many more. It seems 'an agent' is often treated as an empty shell, a vessel through which the aforementioned assumptions run like standard software on a computer. Behind these assumptions exist a variety of ideological, habitual and preferential biases, all of which could be categorized in the realms of: transcendentality, stability, continuity and sameness. As we will see later, the organic tendency to reduce unpredictable, surprising elements on the long run (in order to persist over time), could be said to effectuate said desire for sameness.

However, another defining characteristic behind processes such as evolution, learning and cognition is, in fact: constant change. More specifically, constant spatially-predictive change: where an agent is an agent because they displace themselves through space and sometimes—though certainly not always—learn how to predict their own displacement. Considering organic beings as metastabilizing through constantly individuating processes reveals *satisficing* as the driving motor of action, and thus all reasoning, as Simon noted. In terms of efficiency, the mind is indeed complex, but not as complicated as it is generally made out to be.[27] It is an *integrative model* which serves the purpose of a map or working definition that assists the organism in the pursuit of its own persistence, or rare annulment thereof (in the case of e.g. suicide). *Integrative* not only because it exists within an envelope, but also because it is most often experienced as a unifying coherence (admitting various crucial pathological and pharmacological exceptions). *Model* because its coherence is not a 1-to-1 mapping, it does not align with the informational milieu which exists alongside it: there always exists entropic difference between the two. This certainly implies a kind of dualism, but not the kind that separates body from mind or the Kantian kind that separates *phenomena* from *nuomena*. It is dualistic in the same way that a two-dimensional photograph can be interpreted as a three-dimensional representation. The third dimension is 'hallucinatory,' yet clearly there. The organism is a model of its environment, a model which can be literally said to project its own hallucination over space, about space, thereby altering space. The stronger the fidelity—for example in the case of a very successful and quickly replicating virus, or an incredibly precise and agile surgeon—the higher the capacity to alter space.

4.1 Autonomy

In a paper by Bertschinger, Olbrich, Ay and Jost, the authors attempt to develop an information-theoretic account of *autonomy*, where autonomy requires that "(1) an autonomous system should not be determined by its environment and that (2) an autonomous system should determine its own goals" [2]. The paper's main interest is to define a measure for the investigation of artificial life models. Its key observation is that autonomy becomes, in some cases, undefinable, due to the fact that information shared between the system and its environment exerts influence on both, and the definition of a system's autonomy thus depends on what causes it (the environment or the system itself) [ibid]. In our view, the desire for a distinction between system and environment is too simplistic. In the *Science of Logic* G. W. F. Hegel, whose work we will deal with more extensively later on, regards such *either-or* causal problems as problems of the *understanding* (*Verstand*), which tends towards categorical limits, as opposed to *reason* (*Vernunft*) which tends towards synthesis and infinity. Causality, in Hegel's view, is a reciprocal bond between cause and effect, where the two are co-defined and thus one and the same: because we understand one in terms of the other one and we often attribute a cause only after we have witnessed an effect [*The Relation of Causality*, §1242]. Primacy should be given to *potentiality* over actuality, and not to cause over effect. The important theoretical difference here is that 'potentiality' is an active category, it can shift temporal dimen-

[24] John Dewey, *Art as Experience*, 1934.

[25] Metastability is the *haystack paradox* in computers. More on metastability in the context of Simondon later on.

[26] Intentional both in the phenomenological sense of "aboutness" as well as in *deliberately*.

[27] "The most complex thing in the universe," "a black box," "the biggest mystery," "the (meta)hard problem," "a riddle," "a puzzle," etc.

sions and encompass multiple modalities of thought.[28] This is not a relativistic reductionism (or amplificationism) but rather an attempt to draw attention to the fact that the search for *cause-and-effect* explanations often seems to reduce relationships to dualisms, which offer arbitrary interpretations of experience. A necessary requirement for any duality is that it is composed of two singular things. The quality of something being 'singular' (such as the experience of being a self-aware agent, or picking up *a* stone and not two) is a matter of *attention*. It is perspectival focus which singularizes, but everything which exists is necessarily plural. All experiences are at least dual: self-awareness implies an outside, and any perspectival contemplation implies at least two, if not multiple, perspectives. As Hegel understood, we exist in a reality where everything contradictingly plural (and infinite).

Recapitulating: we'd like to take the basic condition of self-referential consciousness for granted. According to cognitive and computer scientist Joscha Bach, attention, short-term and long-term biographical memory, all form a simulation of the agent's own regulation: the self. Recursively joking, jokingly recursively: the self itself is thus the ultimate *technique of the self* [cf. Foucault]. As Bach explains: "This model [the self] is **not identical with the regulation**, but it allows the organism to explain, predict, and evaluate its own behavior, and thereby improve the regulation." [Bach 2019, his emphasis]. The self-model and the agent's current spatiotemporal limitations "constitute an *ego centric* local perceptual space of the organism. The agent is also able to create counterfactual world states (imagined or remembered mental states that don't conform to the present state of the environment or self). This *mental stage* is crucial for planning, learning and reasoning." [ibid, my emphasis]. Without this map, perceptual chaos would ensue, as in William James' description of what young infants probably undergo: a kaleidoscopic experience, or as we observe in agents who lack specific parts of the brain, or suffer from conditions such as depersonalization or schizophrenia.[29]

Going down a few levels, below selves, to unicellular organisms and their organelles: a bacterium's DNA structure changes, and, if the mutation happens to provide an advantage in its capacity to respond to the environment and endure, the organism is 'reborn' under previously disadvantageous spatial constraints which have now been answered to.[30] Going back up a few levels: an organ—whether organic or simulated by organic beings—learns and specializes by inevitably developing models which answer to (an ever larger diversity of) spatial problems. The model can be said to be the organ itself, as would be the case with a stomach or an adversarial neural network; while the affordances it engages with—perceivable spatial configurations and the traces they define—shape all future interactions.

At an even higher level, if we wish to continue with this hierarchical stacking: a person thinks by constantly reshuffling beliefs and motivations, based on earlier input. In all three cases, the initial input is the environmental, spatial configuration, information about which is imprinted on by way of interaction and predictive modeling. Responding to stimuli by way of anticipation is a key evolutive mechanism. A simple example to illustrate this would be an everyday experience such as picking up a large suitcase: if I know it is full, my grip and upward force will be stronger, resulting in a more or less stable resistance against gravity. If I think it's full when it is in fact empty, the force and grip employed will be inadequate, resulting in an unexpected—slapstick—scene. This phenomenon occurs at a variety of levels and it is referred to as 'unconscious inference' in the PP literature.[31] Information-processing such as the effects of optical illusions occur at levels which are harder—but not impossible—to modulate than, say, symbolic-trading in words and gestures. These effects prompt the agent to intuit its way around space, leading to the formation of habits, such as the effects of illusions, leading to new habits, *ad infinitum*: here we observe a hand washing the other in perfect recursivity. The importance of understanding this, of giving intuition more space than it is often granted, is understanding that what we call *rationality* is perhaps more elusive than we often tend to assume. Let's take the inclining Shepard tables below as an example:

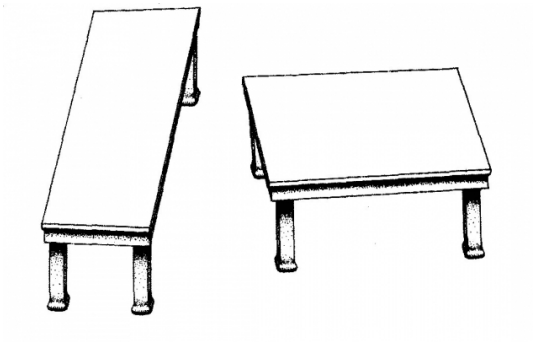

Figure 1. Shepard tables optical illusion, Roger Shepard, from *Turning the Tables.* (1990)

Because you and I have perceived the optical effects of spatial depth a myriad of times, we see the left table as leaner and longer than the right table. These two surfaces are, however, exactly the same. The prediction that they are different emerges from an accumulation of habits over time, your ability to expect objects to behave in a certain spatial manner toward your interests has been prompted by the many three-dimnesional interactions you had, where it was adequate to understand depth in such a way. (Intuitive) thought is amassed unconscious inference converging towards a decision: the affordances of the tables matter in your *umwelt*. It is **inevitably both a decision and a reaction** to the environment and the given somatic state of an agent: a biased state of *inclination*.[32] Thus, repeating what has already been said: what evolutive, cognitive systems do is eternally *fall into place* by way of their inclinations or biases. The catchphrase works on a few conceptual levels, *falling* relates to the gravitational, physical inclination that inevitably constrains organisms, *into* to their predictive capacities, and *place* relates to the affordances/restrictions of the environment (spatial reasoning).

[28] It can shift temporal dimensions because, as a concept, it speculates, it it essentially the category of: *what if*. It also modulates thought because it engages with multiple-realizability scenarios, and thus impulsing spatial reasoning. We will explore this in more detail when discussing "multiple working hypotheses."

[29] For an in-depth analysis into the 'dysconnection hypothesis' see Friston et al 2016 [10].

[30] These spatial constraints may be anything from antibiotic resistance; to a robuster cell wall; to higher efficiency in the ability to synthesize food. We call them spatial because they are molecular interactions that depend on the displacement of elements in one way or another, to recombine things into changing structures.

[31] *Unbewusster Schluss*, first referred to by Hermann von Helmholtz in 1867 to describe an involuntary, automatic mechanism in visual experience. Optical illusions and our inability to 'snap out' of them is an obvious example of unconscious inference.

[32] In an earlier paper we mention the fact that *bias* is, etymologically speaking, also an inclination.

When speaking of the infamous Müller-Lyer illusion, developmental psychologist Alison Gopnik refers to illusions as examples of vision in which our conceptual system overrides our perceptual system [4], which is perhaps another way of separating intuition from belief. However: illusions do not stop happening once we understand that they are happening, we simply understand that they are happening in the same way, perhaps, that we understand even though we cannot watch a tree growing, that it does indeed grow over time. This is not a way of one 'system' overriding another, but a moment during which an organism learns about the availability of affordances by way of prediction. Both experiences require repeated exposure to spatially-challenging events, where reflexive thought and intuitive reaction develop in tandem, and not in isolation from each other. Or perhaps to illustrate it with an even simpler example: a tree which we see at a relative distance from us will look different as we get closer to it: depth perception in general is just as much an 'illusion' as the Müller-Lyer illusion. Our perceptual-conceptual inclination drives action forward: we need to get closer to something to understand it better. The Müller-Lyer illusion stops working if we get very close to it with our eyes, and so do many other illusions. We should not fall prey to a naïve realism which understands illusions as perceptual mistakes, but we should follow the PP dictum and understand all perception as (controlled) hallucination.

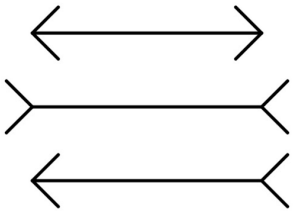

Figure 2. Müller-Lyer illusion, F. C. Müller-Lyer (1889).

From the reactive nature of chemicals all the way up to unconscious inference, there exists a trajectory marked by a stochastic combinatorics, the traces of which remain when things bond, as in chemicals, or when they provide an advantage, as in the evolution of a bacterium. In the case of cognition, this spatial, predictive interaction is not a static organization or structure, but is intimately bound up with the capacity for action. This is why every memory is an adaptation of that memory, every utterance of a word: a new word. This urges us, as proposed earlier, to reconsider imagination along a spatial vector. Imaginatory exploration is better understood as a recapitulation of space in the imagined, predictive dimension. The intensive experience of an agent is not concerned, as we have explained earlier, with the formal diagram of its lived experience but rather with a much more 'vague' stability or persistence. The multi-level probabilistic[33] models that guide the perception-action axis issue predictions for a specific type of balance between the desired act (the internal commitment to a certain representation of action) and the action carried out (which feeds back and adjusts the commitment to that internal representation). This specific balance is a decision, an inference, in accordance with the homeostatic models of the agent, and the affordances of the environment.

What remains thus, in the face of constant change, is a border constituting the 'outer shell' of an organism,[34] and its tendency to persevere in a state of (intermittent) self-coherence: metastability. Perhaps the modern distinction between reason and intuition exists due to a confusion of spatiotemporal scales: the long-term expectation of coherence is often deemed 'rational' (i.e. I will pursue a PhD in philosophy because I wish to explore ideas which will reveal a trajectory of coherence through the years), whereas short-term coherence is often deemed 'intuitive' (i.e. I will combine these two authors because I sense parallelisms I cannot yet explain in long-term terms).[35] Even though the short-term situation is in service, it is a subroutine of, the long-term situation, we tend to prioritize the long-term one as 'rational.' The ability to speculate across this spatiotemporal dimension is, in fact, what Hegel crucially demonstrated to be a manifold more complex than just 'causes' and 'effects.' Retroactive justification—*A* happened because *B*—signals a connection between two events which can only happen because of how one chooses to connect said events, pragmatically speaking.

An *neither-here-nor-there* engagement with this spatiotemporal manifold, such as the one we're developing here, might provide a new perspective on what we mean by 'rational,' or at least shift and confuse the traditional focus on what our priorities are when we employ this term. The making of decisions is based on the availability of *umweltic* information, and whatever is available as information is a result of prior decisions. Being alive means being stuck with this from the get go: neither here nor there, but produced by and co-producing one's environment. In the sections below we continue to explore these topics in the context of predictive processing and its links to Hegel.

5 Predictive processing

> All men dream, but not equally. Those who dream by night in the dusty recesses of their minds wake in the day to find that it was vanity: but the dreamers of the day are dangerous men, for they may act their dream with open eyes, to make it possible.[36]
>
> Though we do not wholly believe it yet, the interior life is a real life, and the intangible dreams of people have a tangible effect on the world.[37]

The Lawrence and the Baldwin quotes above, disparate in original context as they may be, are meant to bespeak the following impression, elegantly expressed by both writers: **we dream ourselves into existence**. What Lawrence intends as social commentary is a distinction between stereotypical cowards and heroes, heroes being decidedly more volitional than cowards. Indeed, those whose character makes them more action-oriented are certainly different than passive personalities who recede into smaller loops of entropic influence. In

[33] Or, according to Bach [2020], "possibilistic," even if something if highly improbable it must remain *within* the possible.

[34] In predictive processing a *Markov blanket*, more on this later.

[35] The observations here presented, linking the immanent yet determinisctic nature of the chemical world and the world of organisms, are not unlike those of Gilles Dezeuze in *Difference and Repetition*: "We are made of contracted water, earth, light and air - not merely prior to the recognition or representation of these, but prior to their being sensed. Every organism, in its receptive and perceptual elements, but also in its viscera, is a sum of contractions, of retentions and expectations. At the level of this primary vital sensibility, the lived present constitutes a past and a future in time. Need is the manner in which this future appears, as the organic form of expectation. The retained past appears in the form of cellular heredity. Furthermore, by combining with the perceptual syntheses built upon them, these organic syntheses are redeployed in the active syntheses of a psycho-organic memory and intelligence (instinct and learning)." [Deleuze, Difference and Repetition, 1968, p. 73]

[36] T.E. Lawrence, *Seven Pillars of Wisdom*, 1926.

[37] James Baldwin, *Nobody Knows My Name*, 1961.

general, we would also certainly agree that not all dream equally, for the aforementioned reason and other reasons. Baldwin, similarly—alluding to much more than just an inner apperceptive world: an entrenched historical, social condition—refers to the palpable, high degree of inertia inner-world-proposals may gather. Setting the tone with these two quotes above, in what follows we will try to prove that we all dream, at night and also during the day. This observation is certainly not new: from René Descartes through to Friedrich Nietzsche, all the way to Douglas Hofstadter, Daniel Dennet, Thomas Metzinger, Andy Clark, Joscha Bach and others, all have tried to push variations of what could be summarized as an *illusionist* agenda. The follow up move we wish to make with our exploration of this is set by the tone of the opening quotes: by **taking the illusion-dimension seriously**—as a matter of fact, as the only dimension we have access to—instead of dismissing it as 'perceptual mistake,' we expand the realm of the possibilistic, in predictive processing terms: or models accommodate more, and surprise becomes thus more manageable.

Descartes is often credited with posing 'the dreaming question' [5], but Al-Ghazālī (ca. 1058 – 1111) inquired much earlier: "[W]]hat assurance have you that you may not suddenly experience a state which would have the same relation to your waking state as the latter has to your dreaming, and your waking state would be dreaming in relation to that new and further state? [...I]n such a state, you would be sure that all your rational beliefs were unsubstantial fancies." Later, much later, in *Human, All Too Human*, Nietzsche exposes a further level of inception what pertains dreams (and the imagination) where, at the end of section 13, he speculates:

> [T]he imagination is continually interposing its images inasmuch as it participates in the production of the impressions made through the senses day by day: and the dream-fancy does exactly the same thing–that is, the presumed cause is determined from the effect and *after* the effect: all this, too, with extraordinary rapidity, so that in this matter, as in a matter of jugglery or sleight-of-hand, a confusion of the mind is produced and an after effect is made to appear a simultaneous action, an inverted succession of events, even.

This observation comes very close to what Hegel observed about cause and effect, as we saw earlier, though Nietzsche's musings are of a more psycho-physiological nature.[38] One's—somatic, symbolic and semantic—exchanges with the world feed back into one's inner states of what one thinks exists, and this feeds back, once again—somatically, symbolically and semantically—into the world *beyond* oneself. This interrelation, or (Markov) chain-reaction results in matter falling within or outside a thermodynamically dynamic boundary. The dynamics of the chemical exchanges coupling internal with external states (causes with effects) can be considered a Markov blanket process. A Markov blanket, in predictive processing and machine learning, is a dynamic probabilistic delineating boundary between multiple states. Coined by Judea Pearl, it can be defined in very simple terms as an interrelated network of dynamic conditions which depend on each other for their interrelating succession. A real-life example can be the skin of an organism, where all cells have a disposition towards remaining closed. In the face of injury the state of the injured cells changes, and chemical reactions tend towards the closure of the injured point. The skin cells seek to return to their previous state, and their previous state depends on their model of themselves. Beyond its skin, an agent perseveres by searching for model evidence of its own existence, by seeking out a certain extent of surprise—even risking the disintegrating effects of injury—in the service of reducing long-term uncertainty (i.e. doing a PhD).

Think of the Heideggerian *zuhandenheit*, in its *un-readiness-to-hand* configuration, that is: you don't notice the tool until it's broken. Everything that 'silently' works, right now, is a product of one's learned, coupled integration of the body and the immediate environment, a *companion-contraption always-already* at play, if you will. *Companion* because it implicates the hallucinatory self, non-identical to itself, a self which allows for self-reference and is thus multiple, as we mentioned earlier; a passenger. *Contraption* because it is a (technological) construction, as well as a trap: it's a limit, a (productive) constraint, a Markov blanket. *Always-already* because we do not have access to our infant selves, our not-yet-selves, in the inchoate stages of self-development. We don't notice precisely because those are the mechanisms that have become evolutionarily integrated into the cohorts we are right now. The nervous system spends the least amount of energy possible in processing new information because it is energetically cheaper to generate a simulation based on what it supposes it will encounter. Imagine having to think anew about every step, every object you pick up, every letter in this text. One's incorporation—called *in*-corporation for a reason—of the types of engagements possible with things is an effect of constant prediction, implicitly and explicitly. It is the result of existing within a Markov blanket, of being an enclosed agent expecting to persevere within that enclosure.

What is more: such state-enclosures do not pertain to somatic states alone, we do not only expect what to perceive at the level of sensorial experience, but we also construct a vast network of systems (infrastructural, legal, habitual, etc.) in order for this to be so, sometimes robustly, sometimes less so. Whenever we refer to models, patterns, laws, and certainty we are always referring to functions of predictability. And it has taken a long time for humans to reach the point at which—at least they feel like—they are constructing their (ethical) environment according to self-made rules. In a recent case from data science, researchers were attempting to create neural networks (NNs) which could solve integration problems in mathematics. However, NNs work very well in areas where probabilistic answers are satisfactory, they do not do well in 'discovering' specific, precise solutions to mathematical problems. While they can filter large amounts of data to recognize patterns, they lack the *possibilistic* dimension of intuition, that is: a large and social enough model of the universe, biased by memory. To find an integral function often requires, perhaps echoing Einstein, "something that's closer to intuition than calculation" [Guillaume Lample and François Charton 2019]. According to the researchers: "Integration is one of the most pattern recognition-like problems in math" [ibid.]. Which means it requires a long, arduous history of trial and error; indeed: also integration, which could be defined as the average of long-term world-interpretation.

5.1 Predictive Processing as a Paradigm Shift

A fundamental philosophical marker for the importance of considering the mind as possibilistically predictive is that prediction implies prior knowledge and thus logically 'limits' all knowledge to *interpretation*, which brings the discovery of new knowledge closer to the concept of intuition than to analytical thought (if we abide to the conservative separation between the two). The concept of interpretation assumes a knower speculates about, and at best approximates rather than asserts, a *knower-independent* truth.[39] An agent

[38] No less metaphysical than Hegel's though, in section 5 Nietzsche presents *dreams* as the purported origin of all metaphysical thought.

[39] This is not akin to relativism and the impossibility to objectivity, more on this later.

constantly updates prior knowledge on the basis of incoming perceptions: induction is thus a matter of *ostensive learning* leading to ever-changing predictions. The paradigm of predictive processing proposes an interesting alternative to some of the traditional questions of philosophy—nature versus nurture, world versus mind, objectivity versus subjectivity, etc.—by suggesting that the interplay between expectation and learning is vastly more interrelated than traditionally assumed. According to Andy Clark, constantly predicting incoming sensory signals is the most effective way we can think and act [Clark 2016]. For this, we need "energy-sensitive surfaces" (organs) and "inward-looking sensory channels, including proprioception (the sense of the relative positions of bodily parts, and the forces being deployed) and interoception (the sense of the physiological conditions of the body, such as pain, hunger, and other visceral states)" [ibid.]. Probabilistic predictions are inflected at every level by changing estimates of one's own uncertainty; neuronal populations adjust themselves to become better at predicting varying interrelated levels of spatiotemporal patterns, or as J.J. Gibson would have it: *affordances*, possibilities for action.

A key observation which is relevant to our consideration of intuition is that the same mechanisms are used for both learning and action. As a learning organism, say, a one-year-old human child, moves about and experiences the results thereof (bumping into things, feeling cold or hot, etc.) [12], the child effectively maps out predictive patterns that channel its future actions and orientation in space. By producing a top-down prediction, the child learns how to better define future predictions due to the feedback hereby received. Intuition, often regarded more or less as an act of guessing, is the core function of the predictive mind. Our spatiotemporal condition is made up of patterns of expectation at various levels, from short- to long-term, from physical and robust to virtual and abstract. Clark states: "The prediction task [...] is thus a kind of 'bootstrap heaven'" [ibid.]. This is because learning about *possibility*, how the environment changes, is resourcefully done by learning how our own body causes actions. The advantage of this merger is the explanation of the coherent-self mentioned earlier: all mental activities (perception, guessing, dreaming, reasoning, acting, etc.) co-emerge driven by the power of an uncertainty-estimating, action-oriented enclosure.

Going back to the notion of *dream*, the predictive processing dictum dictates: **perception is controlled hallucination** [ibid.]. When the neuronal exchange between top-down expectations and bottom-up sensory signals results in an experience: a prediction has occurred. Should this exchange remain inconclusive—i.e. is there really someone there, standing in the dark, or do I just imagine it?—further information is needed and an agent may or may not choose to engage in retrieving the information needed, otherwise known as *learning*. The idea of a *multilayer probabilistic generative model* is that we not only perceive ourselves and the environment, but learn about that relationship by generating incoming data for ourselves, in a true bootstrapping manner. Examples range from behavioral patterns in autistic children, behavioral ticks of all sorts, obsessive compulsive disorders,[40] the aversion to certain experiences after trauma,[41] optical illusions, all sorts of supposed cognitive 'bias.'

Even at the level of interpersonal relations: getting to know someone is learning to predict them better and better. If that prediction matches your desire-parameters, you then become friends or even lovers: you might be one of Lawrence's dangerous dreamers and, after having been exposed to the advantages of chaotic encounters, you require thrill-seeking and people who surprise you, a certain state of uncertainty thus proves to be more pleasant. Or, on the contrary, your peers should be highly predictable according to your own stubborn patterns, because experience teaches you that this is the most reliable type of stability. Or, none of the above: you might prefer to surround yourself with people who make decisions for you, for better or worse, this is, energetically-speaking, very cheap. Large routines and smaller subroutines of learning mark the paths of least, medium and most resistance which dictate learning and thus intuition. Perception is bottom-up action and action is top-down perception. The *feeling* of understanding is having a certain affinity with the disposition of things, learning is remapping inner models in order to expect beyond-anticipated relations.

5.2 Hegel and Predictive Processing

> Structure, as alive, is essentially process, and it is, as such, *abstract process*, the *structural process within structure itself* in which the organism converts its own members into a non-organic nature, into *means*, lives on itself and produces its own self, i.e. this same totality of articulated members, so that each member is reciprocally end and means, maintains itself through the other members and in opposition to them. It is the process which has for result the simple, immediate *feeling of self*.[42]

> [B]iological self-organization is not as remarkable as one might think—–and is (almost) inevitable, given local interactions between the states of coupled dynamical systems. In brief, the events that 'take place within the spatial boundary of a living organism' may arise from the very existence of a boundary or blanket, which itself is inevitable in a physically lawful world.[43] I'm compelled to treat consciousness as a process to be understood, not as a thing to be defined.[44]

Predictive processing often alludes to Kant as the father of its most basic principle–*mind molds world*–because of his transcendental method of argument [see Swanson 2016]. Kant's top-down analytical approach attempting to categorize the principles that lead to cognition, most importantly his claim that causal connections have their origin in the mind and not in the properties of the nuomenal world, is indeed comparable to the claim PP makes the that the mind is not a passive consumer of data but plays an active role in defining the quality of experience. The insistence in PP on Kant (and Helmholtz) largely ignores an important philosophical contribution to the subject of cognition, and organic processes in general: Hegel's interpretation of cognition as perpetual contradiction, which can be argued is a conceptual parallel to the PP principle of free energy-minimization. Indeed, both Hegel and PP propose the idea of cognition as a speculative dialectical process. This advances, or obviates, the Kantian distinction between phenomena and nuomena, with the nuomenon representing simply that which is *not yet characterizable* as a phenomenon. If both Kant and PP offer accounts of how cognition tracks "hidden causes" based on sensory data "and they both develop these accounts using methods of top-down analysis in an attempt to reverse-engineer perception and cognition" [Swanson 2016, p. 4], then it is Hegel who offers an early insight into "hidden causes,"

[40] The development of 'deep superstitions' about the environment, often as a response to high degrees of (traumatic) uncertainty, which a compulsive person undergoes almost inevitably, e.g. situations where they **must** switch the lights on and off because it's what they've always done, it's their habit, and they cannot deal with the uncertainty of not doing so, since their gathered evidence is that future seems to proceed smoothly as if it were an effect of their actions.

[41] Being in a car accident often makes people highly distrustful of vehicles.

[42] Hegel, 1830/2004; p. 377, §356.

[43] Friston, 2013, p. 1.

[44] Friston, 2017, p. 1.

and what PP proposes as free-energy minimization and prediction errors.

As evidenced by the quote above—and much of his *Philosophy of Nature* at large—Hegel defined the organism in its most elemental quality as that which possesses a receptive inside versus a total outside, which, in predictive processing, is a Markov blanket in the service of self-preservation. The boundary, for Hegel, certainly does not stop at the level of the individual. In the *Phenomenology of Spirit* Hegel presents *Geist* as the emergent, transcendental property of human intellect; a collective transindividual negentropy [16]. Similarly, in predictive processing, congregational individuals group together to minimize surprise and preserve coherent self-models more effectively. Below we explore a variety of characteristics in PP and Hegel from which a fruitful comparison can ensue.

In Hegel's work, most prominently in the *Phenomenology of Spirit*, we observe the process of thinking as wholly interlaced with perceiving. This follows from Kant's proposal that the imagination is a requisite for perception, but goes a step beyond as Hegel's account seeks to demonstrate the universality of this fact: it is not just the mind but all relations the mind can deem possible, which actually follow the principle of the imagination doubling down on perception, and vice versa. Hegel takes the reader through as a series of steps guided by 'rational' inference–better defined as spatial reasoning by our standards–as the *Phenomenology* begins with an abstract perspective from which focus; directionality emerges from being something in the face of absolute otherness. Cognition evolves–or tends towards–'absolute knowledge' through different forms of spatial configuration (vantage point; positionality; relationality, etc.; all spatial circumstances), gradually amplifying by incorporation what it understands as coherent. In contrast with Kant, Hegel was a proponent of the intuition that there must be a cohesive unity between perception and *anything we may deep imperceptible*, or else the process of learning would be impossible. Since learning is necessarily a dynamic process, and reason is a speculative, educational process which constantly redefines itself (as opposed to the understanding, which generates limited, rigid categories), there can be no way in which reason does not accommodate within itself a process of contradicting everything it discovers and proposes. Similarly, in Karl Friston's predictive processing, a constantly changing environment presents the challenge of unpredictability to an organism. If the predictions an organism makes do not match the given (dis)order of things, it will need to "keep the discrepancy between the predictions of its model and what actually ensues to a minimum, or what is technically referred to as "prediction error"" [Calvo and Friston 2017]. The organism, in the face of whatever alterity it encounters, must be possess an inner degree of plasticity in order to accommodate future predictions, much like the plasticity of Hegel's reason.[45]

The free energy principle (FEP) proposes that an organism, i.e. anything bounded by a Markov blanket (which ensues in the resistance of entropy; energetic dissipation), must minimize the encounter of surprises along its path. The reduction of surprise, or variational free energy between an inner model prediction and what the organism senses, must follow (variational) Bayesian inference—or as explained earlier: a *possibilistic logic*—given it pertains the endurance of a system that is plastic and thus updates its models as it encounters reality. We will provide a more extensive account of Bayesian inference later on, but what is important to hold on to for the moment is the observation that both Hegel's dialectical logic of never-ending (or teleological *Absolute*-tending) contradiction, can be said to be strikingly similar to the logic behind the FEP, where an organism—but also a group of organisms—persists in the face of change, by forming a concept of itself (an experience of the definitive boundary that contains it, an 'in' versus 'out') and defending it against dissipation by modeling what it expects to encounter, thus actively avoiding constant surprise.[46]

On the other side of the coin, having to learn by exploring its environment, the organism gathers evidence for its inner models, and therefore the more variety an agent is exposed to (the higher the entropy): the robuster it becomes, or the more it fails at persisting, depending on the make-up of its blanket states. As an organism seeks to minimize the discrepancy between its internal states and the external states it perceives, it forms models, or what Hegel would regard as the bounds that construct habits. According to Friston "the state of a system corresponds to its coordinates in the space of possible states, with different axes for different variables" [Friston, 2018a]. Depending on the external variables, a system iterates constant predictions in the service of its course of action. Very often, an organism will seek agreement with its models and thus act in order to predict. This certainly confuses the temporality commonly perceived as linear in cause and effect situations. A very simple example of this is the beep-flash hallucination presented earlier, where the correlation of a beep and a flash primes the agent into perceiving them as occurring together, even when they don't. It is also the mechanism behind the experience of depth in images on two-dimensional surfaces. The nervous system is ready to engage in the activity of moving into; grasping; predicting its next move in space: falling prey to the perceived three-dimensionality. All these mechanisms are guided by the same basic principle, and definitely underlie what is known as confirmation bias—and any other bias, for that matter—in the realm of scientific hypothesis-building.

In their paper "The Dialectics of Free Energy Minimization," Evert Boonstra and Heleen Slagter propose a comparison between Hegel's dialectics and the FEP, much like we've begun exploring above. Their interpretation is based on the readings of Hegel by Catherine Malabou and Slavoj Žižek, and their article "seeks to show how Karl Friston's free energy minimization resuscitates Hegelian dialec-

[45] What is also interesting to note is that Friston refers to certain free energy-minimizing processes as being "dialectical," for example during his March 31st 2017 talk at the University of Edinburgh, *Active Inference and Artificial Curiosity*.

[46] Another interesting detail to note is that in a conversation with physicist Sean Carroll (*Mindscape* podcast, episode March 9, 2020), Carroll asks Friston whether *fire* can be considered as bounded by a Markov blanket. The conclusion they both arrive at is that because fire does not possess a model of the world, it cannot be said to have a Markov blanket. We could speculate about the notion of *model* at length, but what we would like to expose here a further parallel with Hegel, since in the *Philosophy of Nature*, section 337, p. 277 [15], he contends:
"Fire releases itself (*entlaßt sich*) into members, there is a perpetual passage into a product; and this is perpetually brought back to the unity of subjectivity, for the self-subsistence [of the members] is immediately consumed. Animal life is therefore the **Notion which displays itself in space and time**. Each member **has within itself the entire soul** (*Seele*), is not self-subsistent but exists only as bound up with the whole. Feeling, **the finding of self in self**, is the highest achievement and occurs here [in animal life] for the first time [...]." My emphasis in bold. Later on: "There is essentially *Understanding* [*Verstand*] in Nature. Nature's formations are determinate, bounded, and enter as such into existence." Section 338 p. 284.
It is pertinent to note Hegel does not understand fire as alive but as a basic quality, or element, in our interactions with the world. He considers fire and water to be the two dynamic solvents which permeate events on Earth: "The solvent effects of water or fire are wholly separate factors which do not express organic fermentation: any more than when we understand it as the process of oxidation and deoxidation [...]" [ibi.d p. 283].
Albeit in total madness (as Žižek would have it), Hegel predated the concepts proposed by PP, significantly before the conceptualization of Markov blankets.

tics from the perspective of empirical science" [Bonstra and Slagter, 2019]. They also note how "Hegel anticipated contemporary discussions surrounding the appropriation of Friston's framework in terms of cognitivism and enactivism." [ibid.]. Boonstra and Slagter draw on Malabou's comparison between Hegel's plasticity and neuroplasticity[47] and observe that the brain as a 'passive wax' upon which effects and experiences are imprinted, is, unfortunately, a much too installed concept in the field of neuroscience [ibid.]. Their urgency to draw a comparison between Friston and Hegel emerges from the observation that in both Hegel and Friston "we are dealing with a bidirectional process of (active) organization and maintenance at the level of the brain or organism at large. The challenge at hand is to extend the brain's organizational capacity to our most basic understanding of neural functioning, in order to move beyond a conception of the brain as a passive receiver of influences" [ibid.].

They further propose that it is the double meaning of plasticity as the capacity to receive and produce form which is at stake in free energy minimization, they ask: "is free energy minimization a formalization of the dialectical process of plasticity, understood as the capacity both to receive and to produce form?" [ibid.]. Their suggestion, which this paper's proposal wholeheartedly agrees with, is that Friston:

> ...and the appropriators of his framework in terms of cognitivism (Hohwy, 2013, 2016; Wiese and Metzinger, 2017), and enactivism (Clark, 2013, 2015; Bruineberg et al., 2018), do not do justice to the unsolvable tension at the heart of neural functioning under free energy minimization: the brain's anticipations are never "correct"; the brain necessarily and continuously sustains and attempts to solve the "errors" of its anticipations. There is no definite solution to this problem: the brain can only optimize its anticipations and thereby minimize its error. The result is that under free energy minimization, the brain sustains tension **necessarily**.[48]

Additionally, many proposals–especially in data science–continue to focus on the horizon of high-fidelity with respect to a marginal 'context of error' as the undesired condition which permeates everything.[49] This is all the more evident as the list of supposed cognitive 'biases' continues to grow. It seems the most prominent bias of this age is the detection of supposed biases; a 'blind mechanism' Hegel would have relegated to the realm of the Understanding. Dialectically; free energy-minimizationally speaking, an error is a necessary indication of spatial adaptation.[50] Embracing such a vision in data and cognitive science would drastically reorient the current paradigm as it could define alternative parameters for success on the basis of a positively embraced erroneous hypotheses which would eliminate aimless extractionism, patchwork and solutionism, as well as redefine success altogether.[51]

If the nervous system is a structure that actively anticipates its surroundings and the mechanisms for perception are the same as those for action, then, to suggest that what ties Hegel's dialectic with Friston's free energy principle is the plastic circularity of a process of self-perpetuation is far beyond a superficial claim. Both Hegel and Friston purport to provide insights into a *basic mechanism of living organisms*, and the key aspects of error-minimization conceptually relate to the picture emerging if we apply the universal notion of dialectical contradiction to life. Autopoietically speaking, an organism is caught up in an anticipatory structure from the get-go, contradictorily changing itself in order to *remain being what it is*. Seemingly paradoxically, an organism can also never 'remain,' else it would not be an organism, it is in constant conversation and co-production with its environment. Lacking the scientific terminology and expertise, Hegel was already making similar claims in 1842: "The determinateness remains a universality, is one with the element and principle; for the organic being, there is nothing which it is not itself. The reflection-into-self of the organic being means that its non-organic world is no longer in itself (*an sich*); this exists only as sublated and the organism is the positing and sustaining of it." [*Philosophy of Nature* §341 p. 301] "The living thing is the point, this particular soul (Seele), subjectivity, infinite form, and thus immediately determined in and for itself." [ibid. §339, p. 284].

If Friston's framework (and its appropriation in cognitivism and enactivism) does not go far enough in explaining the tension between the abstract and the relational, where plasticity is sustained by "the contradictory tension between particular determinacy and its dissolvement into the universal" [Malabou 1996/2005; p. 12, quoted in Boonstra and Slagter] it is because we are confronted with the question of seclusion and openness, that is: with the particular, the universal; and everything in between. In the introduction to the *Philosophy of Nature* Hegel explains that the universal attempt at a philosophy of Nature is the constant problem "whose solution both attracts and repels," it's the confusion between what we mean by nature and what we mean by philosophy [pp. 3, 4]. Later on, comparing the philosophy of nature with the consideration of physics, he says they are both a "*comprehending* (*begreifend*) treatment, [they consider] this universal in its own immanent necessity in accordance with the self-determination of the Notion." The notion, for Hegel, is indeed the historical, transformative, dialectical process by which cognition acquires the contemplative capacity that leads to understanding and reasoning.

So, precisely because an organism or a concept requires an environment for sustenance, this tension between universality and particularity cannot ever be 'resolved,' as we saw with the concept of autonomy earlier. As a matter of fact, this very tension is what results in organisms, concepts, communities, histories, etc. The dissolution of tension is what Hegel tried to leave 'as is' when proposing the dialectic.[52] It is neither, and it is a process, there is no need for bound-

[47] "Malabou (1996/2005) shows how in Hegel's work, the notion of plasticity encompasses not only the designation of passively receiving form, but simultaneously implicates its obverse: the capacity to produce form. Said differently, plasticity harbors activity that often gets lost in its neural variant. And yet, advances in our scientific understanding of neuroplasticity also touch on the activity of plasticity, by moving beyond the passive association of synapses."

[48] P. 3, my emphasis.

[49] On this subject, Matteo Pasquinelli rightly notes: "A paradigm of rationality that fails at providing a methodology of error is bound to end up, presumably, to become a caricature for puppetry fairs, as it is the case with the flaunted idea of AGI (Artificial General Intelligence)" [27].

[50] It is difficult to underline the importance of this banal claim, because it is one of those philosophical clichés which has seen so much publicity it's become a joke. The old and overabused Hegelian magic trick consists in pointing at the fact that oppositions are far more than two mutually-exclusive categories, given that they co-produce each other and are, in a way, one and the same. It should suffice to say that, as an ontological magic trick, its prestige is no longer surprising. Its magic, however, remains unchallenged.

[51] Our call for action here is not unlike that of Nicholas Maxwell with his 'aim-oriented empiricism,' see Maxwell 2005: "The basic idea is that we need to see physics (and science more generally) as making not one, but a hierarchy of assumptions concerning the unity, comprehensibility and knowability of the universe, the assumptions becoming less and less substantial as one goes up the hierarchy, and thus becoming more and more likely to be true [...]." His argument is not unlike the one of multiple-working hypotheses by T. C. Chamberlin, as well as Peirce's development of the concept of abduction.

[52] This appeal/approach is most lucidly presented in the preface to the *Phe-*

aries, only recurrence and a continuing attention towards potentiality. Boonstra and Slagter also remark, and we agree, that if minimizing free energy "places an upper bound on the entropy or dispersion of sensory states" (Friston, 2012), this means from the get-go no escape from dispersion is possible, and thus persistent adjustments are the only reality. Much like in the case of Lacan's *objet petit a*, for the organism "there is no unobtainable goal, because the "goal" consists in sustaining the continuous process itself, through which dispersion or lack inherent to the organism's organization is constrained, and a boundary is maintained. We could say the "goal" is continually "reached" and "missed" simultaneously [Boonstra and Slagter].

The authors also explain that their appeal to Hegel offers a a change in perspective: "it is not enough simply to choose either cognitivism or enactivism; seclusion or openness. Or to insist on finding the right balance between the two. The appeal to Hegel allows us to see how Friston's framework reintroduces the old Hegelian theme of a contradiction constitutive of life" [ibid.]. Current developments in both cognitive science and data science alike necessitate a paradigmatic change towards the conceptualization of this elemental tension, at the level of definitions, concepts, organisms, etc. As the authors observe, both in the case of Friston and Hegel, a precondition for this tension is the existence of a limit between the organism and its environment: "the events that 'take place within the spatial boundary of a living organism' [Schrödinger] may arise from the very existence of a boundary or blanket, which itself is inevitable in a physically lawful world" [9]. In Hegel's words: "Nature's formations are determinate [*bestimmt*], bounded [*beschränkt*], and as such enter into existence" (*Philosophy of Nature*, p. 284, §339). In other words, whatever can be said to minimize surprise, and thus possess the capacity for intuition and thought, needs to be bounded in the ways which both Hegel and Friston precisely conceptualized.

It would be strange to think that the 'solution' to the feeling of hunger would be to become connected to a tube that constantly feeds the organism a nutritious liquid.[53] There is no 'solution' to hunger, hunger is what happens to an organism when it needs to engage with the environment in order to persist. Hegel's solution to the problem of purpose consists in the experience of lack, it is the way in which "the organism exerts purpose in the preservation of its organization. There is nothing mysterious about the notion of lack. If we regard an organism as an organized product of nature, then lack is simply a state of disorganization that needs to be addressed" [Boonstra and Slagter]. Falling into place, in this way, seems more and more like an apt qualification of what is at play when an organism engages with its environment: it has no choice, it must fall, falling is not a decision but a consequence of gravity. However, falling is also an anomaly which must be avoided, thus in agreement and disagreement; falling and attempting to remain balanced, is an apt metaphor for the error-minimizing, contradictory organism. A basic defiance of gravity is what we do every day when getting up in the morning, walking upright, going to the moon, etc.

According to Boonstra and Slagter, the dependence/independence paradox of a self-perpetuating organism with regard to its environment is that "for there to be a relationship of dependence between organism and surroundings, there needs to be an independent organism in the first place, [but] we can also turn this around: the only way for the organism to conserve and perpetuate itself as an independent entity is to engage in a continuous relationship of dependence with its surroundings, through which the organism assimilates external nature. This relationship of dependence, in turn, is continually reinvigorated by the organism's recurring state of need" [ibid.]. At another level, relatedly, the need to acquire *meaning* in order to make things relevant is a basic activity which pertains to much of the animal realm. Relating A with B because of experiential significance (tracking footprints leads to food) is the phenomenon of lack, and of an inevitably intuitive falling into the environment which answers to said lack. For a concept to be a concept, or an organism to be an organism, there exists a necessary tension. Which is why the notion of 'error' does not do justice to the universal activity at hand. Conceptualized differently, the circle ends where it begins, and existence is guided by an eternity of interlocking circles (of blanket states). According to Allen and Friston the boundary itself "induces a circular causality" (Allen and Friston, 2018). All the metaphors we live by, and we only live by metaphors, are spatial ones. We trap the concept of time within the concept of space. But the past is only what we retrospectively create causes for, and the future only that which we predict, of project effects into.

Near the end of the article the authors briefly assess the dark room problem[54] as a non-problem given that, in a very lowly stimulating state, the dark room problem "it is a problem only if we remain at the level of an abstract designation in terms of sensibility and irritability. If instead, we follow Hegel to the very end, the organism's constitutive contradiction appears equally necessary for the organism's existence; no less than the existence of a boundary" [Slagter and Boonstra]. We agree and would like to argue that perception is its own very specific dark room already, cutting out massive amounts of information in order for it to remain manageable within the specific level of complexity of the organism experiencing it. In the case of humans, for example, the range of audiovisual perception is very limited compared to, say, dogs (in terms of audition) or bumblebees (in terms of vision). Additionally, as Joscha Bach has noted, organic life's success in the face of entropy consists in the investment of a small amounts of energy in order to harvest more energy.[55] Think about eating and digestion as an example of this. This means that all organisms, as already noted, are always 'out of phase' and eternally 'in debt;' their stability is far from 'ideal,' when ideal stability is conceptualized as something static. What should be questioned is in fact this elusive search for an abstract, perfect equilibrium, usually always resulting of the choice between two divergent paths, one of which is 'optimal.' This conception of the organic is the ultimate error, as it leads to highly inflexible models which do not do justice to the diversity of organic life (in terms of anything from culture, to psychology, physiology, history, etc.).

For Hegel, the entire process of becoming self-conscious, as an entity, of gaining degrees of freedom within a group of entities, relates to the process of recognition, which, for the sake of our spatial

nomenology of Spirit. It may even be considered the grounds for having philosophy be the 'queen of all sciences;' because it changes and diversifies eternally, and at the same time because it does not. The universal Heraclitan dictum Hegel proposes remains unchallenged because it does not prescribe nor describe, it *encompasses*; it sets up a high diversity of predictive paradigms. Metaphorically-speaking, it is more like a lens than like a hammer. The dialectic—in its expansive interpretation, to include all contradictory relationships possible—is a principle that encompasses infinite applications and interpretations.

[53] Though developments such as *Soylent* and similar products make room for said consideration.

[54] A purported problem for the FEP, posed by PP scholar Jakob Howhy. Howhy presents the apparent paradox that, if the drive of an organism is to reduce uncertainty, why don't organisms seek out 'dark rooms' where the perceptual input is minimal and thus highly predictive? The nuomenal counterargument is that what we experience right now is, in fact, the dark room.

[55] MIT *Artificial Intelligence* podcast, June 15 2020.

metaphor, we might as well call *reconnaissance*.[56] The gradual process of understanding this *ontogenesis*[57] and limits is akin to the the process of learning by way of prediction [expand point]. Karl Friston's free energy principle: the principle guiding anything within a Markov blanket which minimizes free energy by modeling anticipations, based on prediction errors which approximate Bayesian inference, is a markedly dialectical process, advancing by the estimation (intuition) of contradiction and an attempt at its resolution (both at the macro and micro levels: in bacteria and also in cities). What Hegel and the German idealists called *intellectual intuition* (*intellektuelle Anschauung*), relates to perception itself creating what it encounters in a God-like fashion, which comes conceptually very close to the "perception is controlled hallucination" dictum. A system's tendency towards its metastable self-enclosed, individuated condition, is definitively teleological, too. It could be argued that Hegel's mereological understanding of *Geist*; the fact that it progresses sustained by a cultural community and its surrounding materiality– which in our conception *is* that very community and its materiality– is not unlike a proposal of free-energy minimizing enactive semantic externalism, where reality functions as an 'auto-complete' operation from the perspective of the agent engaging with it as a dynamic knower-learner, i.e.: *intuitively*.

5.3 Bayesian reasoning in spatial perspective

Bayesian reasoning is strongly tied to predictive processing and the free energy principle, as the *negative* of possible 'surprise' can be considered Bayesian model evidence: the (*sui generis* or acquired) predictive inference "which makes the free energy bound a better approximation to the surprise that action is trying to minimize" [9]. An organism not only considers, but actually acts in order to strengthen its models, in order to have accurate predictions. Otherwise phrased: intuition spatially reasons towards the actualization of its inherent potential. In the humanities, Bayesian reasoning is often unduly criticized because it relegates the complexities of the nervous system and decision-making to the discreteness of unambiguous calculations. Below I would like to explain why this is unduly, given that we engage with variational Bayesian reasoning in daily life. Which is another way of saying: learning from experience. How are we to account for the process of learning, if not as a mechanism that alters our beliefs in order to encompass new, ever-changing experiences?

The probability that an event occurs needs to be based on whatever its prior circumstances were. I do not know that the mailperson will pass by unless I've seen them passing by a number of times (or know I should expect them because of knowledge I've acquired through other means). If I'm awaiting some important piece of mail, I will anticipate their arrival by adjusting my expectations to what I recall was their earlier pattern of appearance. However, in our complex frame of existence, people are erratic and subject to events which influence their predictability. If it's windy, the mailperson is delayed and arrives at different times. If we step outside the realm of Bayesian inference as a statistical method for deriving conditional probability, we can simply understand this mechanism as a way to connect the desire to observe functions of cause and effect in different ways. Better said: conditional probability estimation is a mechanism by which an agent can fine-tune its predictions, and thus spatial adaptability, by being able to anticipate not just a simple pattern, but its changing relation and co-occurrence with other patterns. Probability estimation changes depending on the available evidence, thus, following Bayes' theorem (1), I would be able to make a good estimate about my mail arriving based on the following calculation:

$$P(\theta|\mathbf{D}) = P(\theta)\frac{P(\mathbf{D}|\theta)}{P(\mathbf{D})} \tag{1}$$

Or in other words: what is the probability (P) that the mailperson passes by at a given time (θ) if it's windy (**D**)? That is the probability of the prior that the mailperson has passed at a given time (θ), times the probability that it's windy, divided by the probability that the mailperson passes by at a given time, and all of the above divided by the probability that it's windy. Think about how simple this premise is, and realize how it's not a mathematization for the sake of sport. Theta is the prior expectation, the feeling of "it's 12:45, the mail should have arrived by now." **D** divided by theta divided by **D** is the estimation we arrive at by comparing the windy versus non-windy circumstances. If it's not windy often, and the mailperson is very punctual, our probability will be quite high, perhaps almost 1. This results from comparing a few of our expectations, and we do this all the time, not as a computer would, but following variational expectations which are comparable to Bayes' theorem. When a crucial piece of information is missing from our framework (such as how often it's windy) we resort to *intuition* (or in data science: a variety of variational Bayesian methods to settle our priors somehow).

Should weather patterns change depending on the season—or climate change—I would have to adjust my intuitive estimation of how often it's windy. Bayes' rule *as a rule* offers the possibility of dynamically adjusting the relation between several events, or states. According to Friston this is actually the basic mechanism all Markov-blanketed organisms follow in order to adjust their states to minimize free energy [9]. This is not to say that the state estimates are always correct, because priors may be off due to many different factors. Being primed towards a particular state only depends on the given physiology and its experience Bayes rule is a simple rule, it is a relationship between probabilities, it is the *relationship* that matters for the sake of our argument, not the actual estimation itself. In the context of intuition it is important to understand that most of the time, when we intuit, we are undergoing some kind of Bayesian recursion, given the fact that we assess reality based on what we know about it, and learn about it in the process, which updates our future assessments. If we didn't have this kind of built-in mechanism, a mechanism that minimizes free energy, we would not learn and thus also not survive.

Why consider it as 'spatial'? Because it observes how things occur in spatial relation to each other, it is an inference which *invents* interest in the collision between several movements at once. The estimation derives a function of interest (my mail) from an otherwise semi-relevant framework (weather patterns and mail deliveries). This is why superstitions (a kind of intuition) arise in many cases, where people combine two patterns into an explanation or speculation, regardless of whether they are truly spatially related or not. If we relate the motion of a black cat to a negative outcome, we relate two spatial events regardless of their actual connection. This is why speculation is spatial: it invents interest in the collision of events. Provided the complexity of our experience, and the fact that its underlying logic attempts to minimize free energy, it is possible to speculate about how the concept of god(s) emerged in our semantic history. It is far easier to explain cause and effect by way of a model one is already familiar with; oneself. The fact that we are, as Hegel proposed, 'one cognitive thing' advancing together as *Geist*, also explains why the large-scale prior updating capabilities of *Geist* also rid *Geist* of ideas that do not have enough explanatory power. Boonstra and Slagter's obser-

[56] In every sense of the word: acknowledgement, investigation, re-cognition, agreement or bond, etc.

[57] More on Simondon in section 6.

vations relating Hegel to PP could be enriched by indeed highlighting this aspect of his concept of *Geist*. Predictive processing's FEP amplifies the speculation that social animals evolved into groups because this minimizes free energy collectively; language, architecture, dance, music, literature and many other human activities are a way of intermingling experiences for the sake of enlarging our knowledge together, expanding the bounds of experience.

6 Conclusion

It cannot be emphasized enough that understanding the spatial dimension of predictive speculations can bring about new, different yet valid (evidence-based, testable hypotheses) observations about the way cause and effect interrelate. From the perspective of predictive processing, an agent is actively anticipating its environment rather than being guided by incoming batches of new, neutral information. It is necessary for this latter interpretation of perception—the most prominent one in AI today—to be abandoned in order for better concepts of 'bias' or 'error' to emerge. In terms of the *cause and effect* consideration described earlier with regard to the anticipation of mail: we are not saying that an agent will be able to cause wind to happen,[58] but that cause and effect are interrelated in the Hegelian sense, bounded by desire (the goal to receive mail, which is only a subordinate function of whatever reason we're expecting mail about) to which an account of the organism as actively expecting it because of its desire to connect two isolated events in a speculative fashion is preferred over the naïve account of pure information consumption and talk of 'facts.' It is by anticipating what it desires that an agent can engage in actively changing the state of circumstances around it, by moving towards or away from stuff—literally and abstractly—and actualizing its models by literally acting them out. Facts, data, are only what they are about. What they are about reveals an agenda of desire. Desire leads down the road of goal-orientation or anticipatory speculation.

In this paper we have attempted to expose several things:

1) Intuition is a crucial and often overlooked concept in the field of data science, and in philosophy and psychology it is a contested faculty which is more often than not posited against rationality. We tried to show that it is a form of inevitable prediction as agents participate in their environments, and how this inevitability can be allegorized as a process of *falling*.

2) This prediction is the connection between two or more spatial models. Intuitive prediction is driven by the combination and recombination of spatial speculations. Even at the level of abstract concepts: these originate spatially and become contracted; summarized estimations, whose energetically-demanding spatial aspect we lose track of, as a result of free-energy minimization.

3) This results in cognition, and what the agent's movement through space actually does is *make predictions take place* (or feed back an error signal, in which case: learning happens). Mind thus becomes space, as mind falls into its own predictions. This is the effect in place, for example, in the search for confirmation in all possible realms: from the psychological comfort of belonging to a group, to vast galactic exploration. This aspect of the evolution of intelligence; the actualization of emergent mental action *beyond* the agent itself, is excellently captured by Hegel's concept of *Geist*, and this is why his work deserves revisiting in the context of free-energy minimization. In the Preface to the *Phenomenology*, Hegel states:

> The more conventional opinion gets fixated on the antithesis of truth and falsity, the more it tends to expect a given philosophical system to be either accepted or contradicted; and hence it finds only acceptance or rejection. It does not comprehend the diversity of philosophical systems as the progressive unfolding of truth, but rather sees in it simple disagreements. The bud disappears in the bursting-forth of the blossom, and one might say that the former is refuted by the latter; similarly, when the fruit appears, the blossom is shown up in its turn as a false manifestation of the plant, and the fruit now emerges as the truth of it instead. These forms are not just distinguished from one another, they also supplant one another as mutually incompatible.[59]

There is no need for new systems; only reconciliations, adaptations, (mis)interpretations, growths. It is, in our current line of argumentation, a proposal to move away from either-or and attempts at 'perfect' formalizations. Agents process information in a predictive fashion, and this effecting of consciousness can be correlated with Hegel's account of reason as a plastic faculty: learning is happening at every step, in the end: rationality is a more elusive concept than intuition. Re-introducing Hegel to modern contemplations of the workings of life and of mind is relevant because of his insistence on being as a fluid, eternal immediacy and dialectical rationality as being at the seat of progress and existing beyond the skull (i.e. 4E). Hegel's *Phenomenology* recasts philosophy as the fertile ground which offers the new precisely because it is an open system. Creation does not exist, only renovation, recombination. Within their "Biologico-Cultural Matrix of Human Existence" project, Humberto Maturana and Ximena Dávila Yáñez developed a similar way of thinking about spacetime as the one presented by this paper and by Hegel's intercategorical concept of *Geist*. Maturana says:

> "We human beings exist in the present, in a continuously changing present, the past and the future do not exist as such, and they are manners of being now, in the present. The cosmos that we generate in our living occurs, exists, as a continuously changing present. The past as a way of explaining the present being lived, arose in its continuous change by proposing a generative mechanism that would have given rise to it if the operational coherences of the now being lived were conserved. The future is a manner of living now in the proposition of what would happen if the operational coherences of the present being lived now are conserved in the continuously changing present being lived. Autopoiesis, living, occurs in a continuously changing present: living occurs in no time, in zero time."[60]

Memory arises as a way of dealing with current events. As a way of explaining our current state. Memory is not a window into the past but a window into the present, the self creates a hallucination that needs to explain its current state, so that it may continue going about as a being, enclosed and coherent. This condition is what leads towards the inclination or bias that reveals itself as categorical or seeking stability. Without this tendency: selfhood, subjectivity or perspective, would not make sense; organisms would not make sense, they would not exist as separated entities. This is why it's difficult to imagine a rock or fire as a kind of organism: because it doesn't energetically employ this recursive effect we call memory, which leads to action and thinking. It gets complicated with something like a virus, which is not 'officially' considered as living—the same could be said

[58] Though in some instances we may observe wind if we expect it to be there, but that's another discussion.

[59] Hegel, *Phenomenology of Spirit*, [16].

[60] Maturana, H., *American Society for Cybernetics*, Wiener Medalist address, 2008 [24].

of a neural network. The virus, perhaps, really, truly just *falls*, without much recursion—something similar could be said about gradient descent. The way more complex organisms 'fall' is by tripping on their own shoelaces: one hand washes the other. But you have to have hands before you can have this auto-gratifying relationship. A virus doesn't have hands, but it does have a memory. It is in conversation with the environment. And, again, falling prey to the poetry: a rock *is*, in effect, the memory of everything that happened to the particles that compose it. A system is anything you draw a limit around. A cybernetic organism is anything that knows that limit, and tries to stay within it. Cybernetic systems persistently insist on their boundedness by a limit, wishing to maintain it, but inevitably wear and tear due to gravity: all of our systems are falling apart.

REFERENCES

[1] Woo-Young Ahn, Nathaniel Haines, and Lei Zhang, 'Revealing neurocomputational mechanisms of reinforcement learning and decision-making with the hbayesdm package', *Computational Psychiatry*, **1**, 24–57, (2017).

[2] Nils Bertschinger, Eckehard Olbrich, Nihat Ay, and Jürgen Jost, 'Autonomy: An information theoretic perspective', *Biosystems*, **91**(2), 331–345, (2008).

[3] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al., 'Language models are few-shot learners', *arXiv preprint arXiv:2005.14165*, (2020).

[4] Peter Carruthers and Peter K Smith, *Theories of theories of mind*, Cambridge University Press, 1996.

[5] René Descartes, *René Descartes: Meditations on first philosophy: With selections from the objections and replies*, Cambridge University Press, (1641) 2013.

[6] Hubert L. Dreyfus, *What computers still can't do: A critique of artificial reason*, MIT press, 1992.

[7] Jonathan Evans and Keith Stanovich, 'Dual-process theories of higher cognition: Advancing the debate', *Perspectives on psychological science*, **8**(3), 223–241, (2013).

[8] Ronald Fagin, Yoram Moses, Joseph Y Halpern, and Moshe Y Vardi, *Reasoning about knowledge*, MIT press, 1995.

[9] Karl Friston, 'Life as we know it', *Journal of the Royal Society Interface*, **10**(86), 20130475, (2013).

[10] Karl Friston, Harriet R Brown, Jakob Siemerkus, and Klaas E Stephan, 'The dysconnection hypothesis (2016)', *Schizophrenia research*, **176**(2-3), 83–94, (2016).

[11] Gerd Gigerenzer and Daniel G Goldstein, 'Mind as computer: Birth of a metaphor', *Creativity Research Journal*, **9**(2-3), 131–144, (1996).

[12] Alison Gopnik and Eric Schwitzgebel, 'Whose concepts are they, anyway? the role of philosophical intuition in empirical psychology', *Rethinking intuition*, 75–91, (1998).

[13] Beate Hampe, *Metaphor: Embodied cognition and discourse*, Cambridge University Press, 2017.

[14] Georg Wilhelm Fredrich Hegel, *Hegel: Elements of the philosophy of right*, Cambridge University Press, 1991.

[15] Georg Wilhelm Friedrich Hegel and Arnold V Miller, *Hegel's Philosophy of Nature: Being Part Two of the Encyclopedia of the Philosophical Sciences (1830), Translated from Nicolin and Pöggeler's Edition (1959), and from the Zusätze in Michelet's Text (1847)*, volume 2, Oxford University Press, 2004.

[16] G.W.F. Hegel, *The Phenomenology of Spirit, edited and translated by Terry Pinkard*, Cambridge University Press, 2018.

[17] Douglas Hofstadter and Emmanuel Sander, *Surfaces and essences: Analogy as the fuel and fire of thinking*, Basic Books, 2013.

[18] Daniel Kahneman, *Thinking: Fast and Slow*, Macmillan, 2011.

[19] Immanuel Kant, *Critique of Pure Reason, trans. N. Kemp Smith. Original A edition published 1781*, Macmillan, London, b edition 1787 edn., (1929).

[20] George Lakoff and Mark Johnson, *Metaphors we live by*, University of Chicago press, 2008.

[21] Jeannette Littlemore. Metaphors in the mind: Sources of variation in embodied metaphor: by jeannette littlemore, cambridge, uk, cambridge university press, 2019.

[22] Computing Machinery, 'Computing machinery and intelligence-am turing', *Mind*, **59**(236), 433, (1950).

[23] Gary Marcus, 'Gpt-2 and the nature of intelligence', *The Gradient, January*, **25**, (2020).

[24] Humberto Maturana, 'Maturana wiener medalist 2008 address'.

[25] John McCarthy et al., *Programs with common sense*, RLE and MIT computation center, 1960.

[26] Erik T. Mueller, *Commonsense reasoning: an event calculus based approach*, Morgan Kaufmann, 2014.

[27] Matteo Pasquinelli, 'How a machine learns and fails', *spheres: Journal for Digital Cultures*, (5), 1–17, (2019).

[28] Iuliia Pliushch, 'The overtone model of self-deception', (2017).

[29] Herbert A Simon, 'Making management decisions: The role of intuition and emotion', *Academy of Management Perspectives*, **1**(1), 57–64, (1987).

[30] Herbert A Simon and Kevin Gilmartin, 'A simulation of memory for chess positions', *Cognitive psychology*, **5**(1), 29–46, (1973).

[31] Gilbert Simondon, 'The position of the problem of ontogenesis', *Parrhesia*, **7**(1), 4–16, (2009).

[32] Peter Van Inwagen, 'Materialism and the psychological-continuity account of personal identity', *Philosophical Perspectives*, **11**, 305–319, (1997).

Robots with Purpose

Steve Battle
steve.battle@uwe.ac.uk

Dept. Computer Science and Creative Technologies,
University of the West of England, Bristol

Abstract. In this paper we apply Perceptual Control Theory to Braitenberg Vehicles. Computational models explain systems in functional terms, as a transformation from input to output. Living, and indeed robotic systems are so much more, having to engage with, and survive within a world. Robots, of course, are very simple machines, but purposeful, *cybernetic* explanations of robotic behaviour can be a powerful alternative to computational explanations.

1 INTRODUCTION

Teleological explanations are seen by modern science as the 'poor cousin' of modern causal and reductive accounts of nature. Aristotle identified four different *causes* that could be invoked in an explanation. His *material cause* emphasised the static aspects of the structure of matter, for example the structural organisation of biological cells. By contrast, an explanation of the cell dynamics is an *efficient* cause which also touches on agency. His *formal* cause addresses the form or design of a thing, and so a scientific explanation might focus on the way that DNA provides a kind of 'blueprint' for an organism. Aristotle's *final* cause, or *telos*, describes the purpose of a thing. As with the other causes this can be interpreted in widely different ways, from the homeostatic maintenance of a critical variable such as the sugar level in the blood, through directed biological evolution, to the ultimate 'meaning of life'. Seeking to clarify the terminology used by biologists, Pittendrigh [14] introduced the term *teleonomy* that applied only to biological phenomena directed towards an end. The concept was further developed by Mayr to exclude the idea of (goal-directed) biological evolution [12], "The development or behavior of an individual is purposive, natural selection is definitely not."

In the *Critique of Judgement*, Kant analyses the differences between Artistotle's teleology and modern science regarding their ability to explain how natural forms come about [5]. Despite the success of science, Kant felt that it could not account for the apparent purposiveness of organisms. Kant's synthesis introduced a third way, uniquely describing natural forms in terms of self-reference; "a thing exists as a natural end if it is cause and effect of itself" [20]. In his view, this approach could sit comfortably as a *regulative* norm alongside modern science which provides *constitutive* laws of nature. This circular process identified by Kant is a distinctive feature of cybernetic accounts of behaviour, indeed of *autopoietic* theories about the nature of life and self-becoming [11].

One of the foundational papers of cybernetics, "Behaviour, Purpose, and Teleology" by Rosenblueth, Wiener, and Bigelow [17] places the study of purpose front and centre. The term *purposeful*, rejected by modern science, is used in cybernetics to describe behaviour directed towards the attainment of a goal [6]. While cybernetics is concerned with the behaviour of organisms and machines, what distinguishes cybernetics from *behaviourism* is the notion of feedback controlled behaviour [4]; that is, goal-directed behaviour. Indeed, "When we perform a voluntary action, what we select voluntarily is a specific purpose, not a specific movement" [17].

The title of this paper recalls the first annual symposium of the American Society for Cybernetics, *Purposive Systems*, held in 1968. At this conference, David Hawkins describes the nature of purpose [7], "As long as we restrict ourselves to the description of behaviour in living things or machines or people, teleological concepts are appropriate wherever the the system described is one for which a particular metastable equilibrium is characteristic of the system, which can be known independently of the particular mechanisms of informational feedback ... by which the system attains or maintains that equilibrium." In this view, purpose is intimately linked to self-referential, homeostatic behaviour. Behaviour cannot simply be seen as a chain of mindless habits; creatures without goals cannot survive in the world. Purpose is a goal-oriented activity, necessarily oriented towards a future outcome. As Tolman puts it [18], "behavior reeks of purpose and of cognition."

2 VEHICLES

Valentino Braitenberg's, 'Vehicles: Experiments in Synthetic Psychology' [2] inspired many to explore its strange intersection of cybernetics and artistry. From the outset it describes creatures that are openly mechanistic – a simple brain laid bare for all to see – but then labels them with suggestively *purposeful* names starting with 'getting around', leading rapidly onto 'Fear and Aggression', and then 'Love'. I think it is no accident that the book challenges us with these different interpretations, inviting us to consider mechanism versus purpose. We can understand Braitenberg vehicles as thought experiments at the intersection of purposeful, and purely stimulus-response (S-R) models of behaviour [9]. These complementary viewpoints suggest different kinds of analysis. In this paper we apply Perceptual Control Theory (PCT) [16] to Braitenberg vehicles, a method used in experimental psychology to understand purposeful behaviour in organisms ranging from bacteria to people, using techniques derived from control theory. This approach allows us to gain a fresh understanding of these vehicles' behaviours. Might we think about them as robots with purpose?

In this paper I focus on the first two vehicle designs, Vehicles 1 and 2 (specifically 2B, or 'Aggression'), analysing them in terms of their purposeful, goal-seeking behaviour. Braitenberg vehicles have sensors and motor output. There is a simple causal arc from sensor

to motor output, so vehicles *can* be understood as simple stimulus-response systems. Lacking any internal state (at least up to Vehicle 4) any given input produces a determined output. When placed in an environment with a light source they react and move. This movement, in turn, causes a change in the light stimulus. These two causal pathways, from the sensors to the motors within the vehicle, and then the motors affecting the environmental stimulus, form a self-referential closed loop. We could describe the behaviours of the vehicle and its environment as a set of simultaneous equations, meaning that they act upon each other simultaneously and continuously. Where this mutual embrace acts to stabilise itself and maintain equilibrium we describe this virtuous cycle as negative feedback. For the experiments below, these vehicles are represented within a computer simulation of the vehicle and a simple environment with a single source of light.

In our 3D universe light intensity follows the inverse square law, such that the light intensity is inversely proportional to the square of the radial distance, R, from the source, $1/R^2$. However, replicating this in the environment of the vehicles appears very counter-intuitive as the light intensity falls away almost too rapidly. In the real world we are more familiar with light emanating from a relatively distant light source, our sun; the diminution of the light reflected from objects at our everyday scale accounts for only a tiny fraction of the total diminution of the light on its long journey from the sun, so it is not usually so apparent. The simulations appear to work more 'realistically' if we assume that the vehicles inhabit a two-dimensional Flatland [1]. For N dimensions the intensity at radius R is $1/R^{N-1}$. In a two-dimensional universe, the light intensity is a circular wavefront that falls in intensity in a way that is inversely proportional to the distance, $1/R$.

Braitenberg Vehicles have one or more sensor inputs and motor outputs. By 'motor', Braitenberg means anything that can provide a motive force, not just electric motors. They could easily represent the flagella on a bacterium. Braitenberg's sensors are wonderfully underspecified. We know they are directional, but they might just as easily measure a chemical or temperature gradient, as measure light. We will think of this as a light sensor that produces an analogue signal in proportion to the amount of light falling upon it. A simple light-sensing 'eye' can provide directional information by virtue of Lambert's cosine law which states that the illuminance on a surface varies with the cosine of the angle of incidence. With a light source directly ahead at 0°, it receives full illumination ($cosine(0) = 1$), falling to zero at a 90° angle of incidence ($cosine(\pi/2) = 0$). This kind of eye, modelled on the simple photocell, will be used for Vehicle 1. When the light source moves *behind* a regular photocell, then it receives no input at all and the output signal remains at zero. With this arrangement, vehicles without any other means of movement will quickly get into the *doldrums* where they don't receive enough light to move, making for a very unconvincing demonstration. Things get a lot more interesting if we give them wrap-around vision such that light from behind provides a negative stimulus ($cosine(\pi) = -1$). Given that nature's nerve bundles cannot carry a negative signal we may add an offset of 1 to provide a signal in the range 0 to 2, inclusive. This kind of sensor is also straightforward to construct in a physical robot by arranging pairs of photocells back to back. This kind of eye will be used for Vehicle 2. In addition, photocells are not perfectly efficient devices for translating light energy falling on it, into an output signal. In particular, these devices will saturate beyond a certain maximum. This feature is added to the Vehicle simulations, effectively adding a speed-regulator to avoid unrealistic jumps in position.

3 PERCEPTUAL CONTROL THEORY

The aim of this paper is to apply the methods of experimental psychology to Braitenberg's vehicles, looking for negative feedback control manifesting as purposive behaviour. We test that this goal-seeking negative feedback loop is real and not just a linguistic turn. To do that we must define a hypothesis about the variable controlled by the loop. William Powers' Perceptual Control Theory (PCT) [15] asserts that organisms are not in the business of controlling their behaviour, but of their *perceptions*. This reflects earlier observations by Merleau-Ponty [13, p. 37],"the motor devices appear as the means of re-establishing an equilibrium, the conditions of which are given in the sensory sector of the nervous system." Our hypotheses about vehicle behaviour must accordingly describe not the motor outputs, but variables that it is possible for a vehicle to perceive, or a simple function of these directly available sensory inputs. For example, a vehicle with a *single* light sensor cannot in principle control distance because the luminosity of the light source is unknown to it, and it cannot compute distance from the perceived brightness alone. the Our first thought might be that the vehicle is phototactic and seeks to maintain a constant level of illumination.

A system can be said to control a variable if every disturbance tending to cause a deviation from the goal state results in a behaviour that acts in opposition to the disturbance [15]. PCT provides the tools for testing whether any of these hypothesised variables is actually controlled by the vehicle. Powers' Test for the Controlled Variable (TCV) [9, 10] is an objective way to determine which, if any of these hypotheses is true. Control Theory tells us how we can keep a controlled variable in, or near, a reference state. The idea is that if we disturb the system by manipulating another variable then the hypothesised controlled variable will hold constant if it is truly under control, or else it will vary with the disturbance. If we have multiple hypotheses then we should try to find a disturbance that will affect one but not the other.

4 VEHICLE 1

As can be seen in Figure 1, Braitenberg Vehicle 1 has just one forward facing eye, and one motor. Vehicle 1 cannot obtain any information about the direction of the light source with respect to its sagittal axis, running from the front to back of its body. We might think of its behaviour as a form of *klinotaxis* which occurs in organisms with sensors that are not paired. Only by moving through, and sampling its environment can it move relative to a sensory gradient.

The sensory signal is conveyed from the eye to the motor by the wire or nerve fibre connecting them, causing the motor to vary continuously in its output in proportion to the input. The brighter the light, the faster it drives the motor. This vehicle might seem trivial, but nature employs something very close to this design in the humble E. coli bacterium. Not just a source of food poisoning, but one of natures own vehicles. Vehicle 1 exploits a feature of the world faced by simple motiles, and even exploited by E. coli; as the vehicle presses forward through its medium, it encounters pushback from frictional forces that introduce randomness. To quote Braitenberg [2, p. 5], "Once you let friction come into the picture, other amazing things might happen. As the vehicle pushes forward against frictional forces, it will deviate from its course. In the long run it will be seen to move in a complicated trajectory, curving one way or the other without apparent good reason. If it is very small, its motion will be quite erratic, similar to 'Brownian motion', only with a certain drive added." To simulate the random buffeting of friction a small random

angle is added to the vehicle heading on each cycle of the simulation. Its heading is a random-walk through angles of orientation, which enables it to randomly 'scan' its environment.

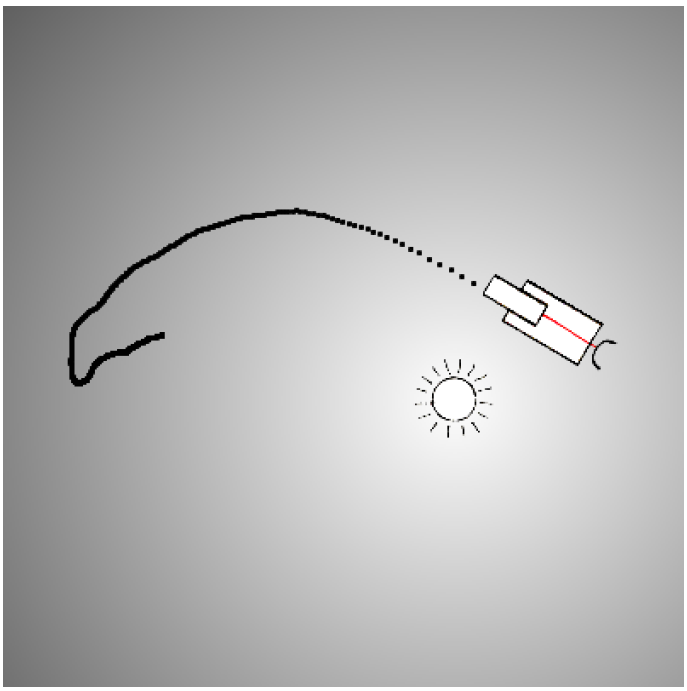

Figure 1. Braitenberg Vehicle 1 is stimulated by light, making it move faster as it approaches the light source.

Vehicle 1 is a light-seeking robot. It speeds up as it approaches a light source, not because of any positive feedback; its speed is simply a function of light intensity falling on its eye. The brightness of the light determines its speed, but not the overall movement towards the light. The decision-making takes place in the eye. The light signal increases on its approach to the light, and then suddenly falls to zero as it passes by, and the vehicle stops. Vehicle 1 aims to minimise the light signal. In terms of PCT this is equivalent to a reference value, or goal, of zero illumination. Our hypothesis (H1) will be that, despite its semi-random walk, Vehicle 1 controls for distance from the light source. To test this hypothesis, a disturbance is introduced that moves the light source from side to side with the addition of a sine-wave to its horizontal position. This means that the vehicle must dynamically track the light. The hypothesis will be tested by measuring the correlation between the horizontal coordinates of the vehicle and the light source. If they are highly correlated then the vehicle can be said to controlling its distance. This may be compared against a similar vehicle that performs a simple random-walk at a constant speed without the benefit of having a sensor. We expect the correlation between this vehicle's position and the light source to be low.

If Vehicle 1 can be thought of as a control system at all, the effect will be very weak and the disturbance may be only weakly opposed [15, p.234]. The experiment lasts for 100K iterations during which we record the horizontal position of both the vehicle and the light source. The light source moves relatively slowly, so that it doesn't move faster than the vehicle. The 100K iterations allow for 40 complete cycles of moving the light source from left to right. We also allow for one complete cycle as settling time before the data is collected, for the vehicle to initially locate the light source from a random starting position.

At the end of the experiment the Sample correlation coefficient is calculated as, $r_{xy} = 0.9143$, where x, y are the samples for the horizontal positions of the vehicle and light source. We interpret this as a high degree of correlation, and so accept the hypothesis (H1) that Vehicle 1 controls for distance from the light source. This may be contrasted with the extremely low measured correlation, $r_{xy} = -0.0533$, of the random-walk control vehicle with the moving light source.

5 VEHICLE 2

Braitenberg Vehicle 2 has two eyes, which allow it to determine the direction of the stimulus and move accordingly (tropotaxis). Its response to light may be described as 'has a fight or flight', and Vehicle 2B specifically demonstrates the latter, accelerating towards the light in a way that might be considered aggressive. It has two eyes and two motors, such that the motors are driven differentially according to the difference in light level received by the two eyes. This results in positive phototaxis, or directed movement towards a light. As the vehicle gets closer to the light the stimulation increases, causing it to go faster until it runs into the light source (hence the name, 'Aggression').

The 'brain' of Vehicle 2B is shown graphically in Figure 2 which shows that each sensor is connected via an excitatory connection to one motor, indicated by the '+' at the end of the nerve fibre. In its simple nervous system, the signals from each eye crosses-over to stimulate the motor on the opposite side of its body. Thus light stimulation on one side causes the motor on the opposite side to run faster, acting to steer the vehicle towards the light. These crossed connections are common in vertebrates, including humans. The Spanish neuroanatomist Ramón y Cajal was the first scientist to map the detailed neural structure of the brain in 1899. He observed optic nerve fibres from the half of the eye closest to the nose cross over to the opposite side of the brain. However, where each human optic nerve carries a million or so separate nerve fibres, the humble Vehicle 2 has just two.

These *tropisms* are seen in nature, not only as movements towards or away from a light source but also as changes in orientation with respect to light. The sea-anenome *Actinea cereus* extends its tentacles perpendicularly with respect to a weak light source but in parallel with more intense light, regulating the amount of falling on it [19]. Similarly, the dorsal light reaction in moths allows them to keep the moon above them, one reason why they become entrapped by street lamps [8].

The arrangement of the sensors and the crossover network mean that the vehicle is in equilibrium when a light source is directly ahead, and the stimulation of both motors is equal. This is what *causes* the vehicle to ram the light source head-on. Can this be said to be a goal of the vehicle? This 'aggressive' event is a little short-lived to study at length, so I look to subsequent behaviour after it has run into the light. If we allow it to pass through unharmed then a small imbalance one way or the other will again cause it to turn towards the light. In the designs explored here there comes a critical point where the turning circle of the vehicle holds it within a close *orbit* around the light source. Again the vehicle has entered some kind of equilibrium, orbiting clockwise or anti-clockwise around its source of energy.

The eyes are mounted on the vehicle body with a disparity of 90°

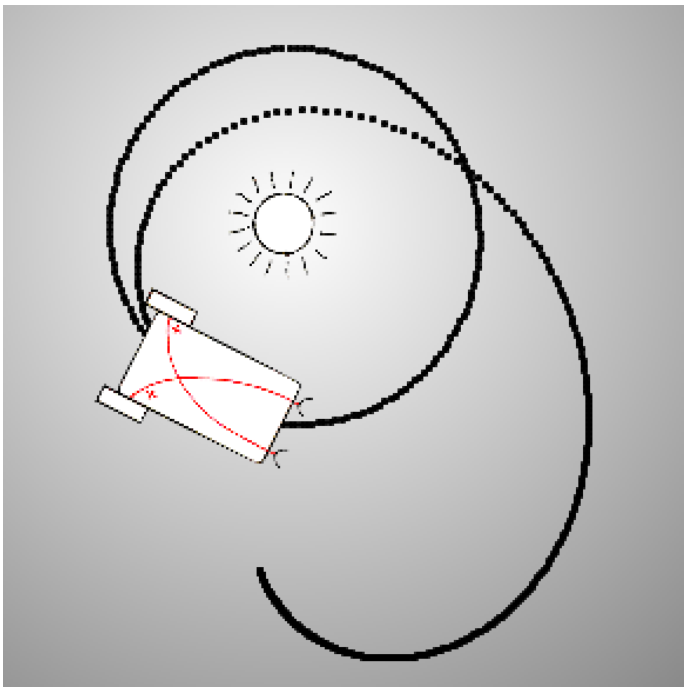

Figure 2. Braitenberg Vehicle 2B approaches the light source 'aggressively' then settles into a close orbit.

from each other, each one facing outwards from the centre-line of the vehicle at an angle of $45°$, or $\pi/4$ radians. If θ is the angle of the vehicle to the light source (how much it would have to turn to face the light), then the light received by the right eye varies with the *cosine* of θ with this offset. The $90°$ angle between the eyes introduces a phase-shift such that the light received by the left eye varies with the *sine* of θ. The identity of Equation 1 implies that after factoring out the illumination, the difference between the left and right eyes is proportional to $sin(\theta)$, in other words it is a straightforward representation of the angular deviation from the light source. Note that this holds whether or not we add a constant to each side to bring them into the positive range, as with Vehicle 1, because these simply cancel each other out.

$$\begin{aligned} sin(x) \pm cos(x) &= \sqrt{2}\, sin(x \pm \pi/4) \\ l &= sin(\theta + \pi/4) + 1 \\ r &= cos(\theta + \pi/4) + 1 \\ l - r &= \sqrt{2} * sin(\theta) \end{aligned} \tag{1}$$

If we look at the kinematics of Vehicle 2B, The angular velocity of the vehicle, ω, defined in Equation 2, is proportional to the difference between the two eyes. The variable, d, is the distance between each wheel d the centre-line of the vehicle, so $2d$ is the distance between the two wheels; the further apart they are, the slower the vehicle turns. The angular velocity is influenced by the brightness of the light, b. For a light source dead-ahead or dead-astern the stimulation of both sensors is identical and they cancel out, or equivalently, $sin(0) = sin(\pi) = 0$, utilising Equation 1 above. The difference, and hence the angular velocity, is zero.

$$\omega = b\,(l - r)/2d \tag{2}$$

The linear velocity, v, of the vehicle is simply the average of the two motors. Imagine both motors running forwards at full speed; the average will be the same. Now imagine one motor running forwards at full speed and the other running backwards at the same speed; they average out at 0 as the vehicle simply spins around on the spot. This is modulated by the brightness, b, so that it turns faster in brighter light.

$$v = b\,(l + r)/2 \tag{3}$$

The turning radius, r, of the vehicle is given by $v = r\omega$. Solving for r in Equation 4, we see that the brightness terms cancel out. There are two additional interesting orientations to the port and starboard of the vehicle. With the light source $\pm 90°$ to the axis, $sin(\theta) = \pm 1$ and so the difference between the eyes simplifies to $\pm\sqrt{2}$. At the same time, the sine and cosine response of the eyes cancel each other out exactly, leaving us only with the added constants which average out to a linear velocity of 1 as the numerator. The implication of this is that once the vehicle has achieved a stable orbit, the turning radius is a function only of the distance between the two wheels, and so the vehicle will follow a stable clockwise or anti-clockwise orbit independent of brightness.

$$\begin{aligned} r &= v/\omega \\ &= \frac{b\,(l + r)/2}{b\,(l - r)/2d} \\ &= 2d\,\frac{(l + r)/2}{(l - r)} \\ &= 2d\,\frac{(l + r)/2}{\sqrt{2}\,sin(\theta)} \\ &= \frac{2d}{\sqrt{2}} \text{ where } \theta = 90° \end{aligned} \tag{4}$$

In this experiment we will test our hypotheses by introducing a disturbance in the brightness of the light source. A reference value for each hypothesised variable is estblished with the light source at full illumination. We then collect a number of readings for each variable at different levels of luminance, and calculate the residual difference between the observation and the respective reference values for each hypothesis. The residuals $\hat{y}_i - y_i$ measure the difference between the reference value, $\hat{y}_i$ (in this case held constant over i), and the observed values, y_i. The results are summarised in the calculation of the root-mean-square deviation (RMSD) of these residuals for each hypothesis, as in equation 5 below, where n is the number of observations. The greater the level of control, the lower the RMSD error.

$$\text{RMSD} = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}} \tag{5}$$

The root-mean-square deviation is sufficient to be able to make a determination about each hypothesis in isolation but to compare different hypotheses we need to normalise the results to account for observations on different scales. The *normalized root-mean-square deviation* (NRMSD) may be calculated by dividing the RMSD by the mean, $\bar{y}$, of the observations as in equation 6, below. This is expressed as a percentage with lower values indicating less residual variance.

$$\text{NRMSD} = \frac{\text{RMSD}}{\bar{y}} \tag{6}$$

Powers states [15, p.233], "A controlled quantity is controlled only because it is detected by a control system." Given that a system may only control that which it can perceive, then it cannot be said to control its turning radius; this is merely a contingent fact that follows from the hypothetical control of θ. We might expect a phototactic vehicle to adapt by moving closer to the light source to keep the illumination constant, whereas a vehicle that is concerned with the angle to the light source would not be so affected. Our hypothesis (H2) then is that Vehicle 2B seeks to maintain a constant level of brightness through phototaxis. Alternatively, we might assume that, like the moth, the vehicles' dorsal light reaction holds the angle, θ, to the light source constant; this is hypothesis 3 (H3).

	H2 (brightness)	H3 (angle)
reference	13.4736	1.6399
luminosity		
90%	12.0696	1.6327
80%	10.6784	1.6256
70%	9.2999	1.6186
60%	7.9340	1.6116
50%	6.5807	1.6046
40%	5.2399	1.5977
30%	3.9115	1.5909
20%	2.5955	1.5841
10%	1.2918	1.5775
mean	6.6224	1.6048
RMSD	7.2894	0.0373
NRMSD %	1.1007	0.0233

Table 1. Analysis of Vehicle 2 brightness and angle data with increasing disturbance (decreasing luminosity). Observations for brightness and angle are taken for each of the ten levels of disturbance. The RMSD summarises the squared residual difference of each data point from the reference. For comparison of H2 and H3, these are normalised (NRMSD) using the mean of the data. We see two orders of magnitude improvement in control for angle (H3) over that of brightness (H2).

In any experiment based on PCT, the experimental subject is not simply responding to the error, but to the difference between what they perceive (illumination or angle) and a *reference*, or goal condition. We may establish these reference conditions by observing the system *without* disturbance. Following an initial run in which we record the reference values with no disturbance, we run ten test trials each with increasing levels of disturbance such that the luminosity of the light source ranges from 90% down to 10% of full luminosity. Each test run lasts for 3K cycles which gives the vehicle sufficient time to settle into a stable orbit. At the end of each run, we record the apparent brightness at the position of the vehicle, and its angle to the light source (with 0° representing dead-ahead). The vehicle may settle into either a clockwise or an anti-clockwise orbit, which differ only in their sign; we therefore take the absolute value of the angle.

The results are summarised in Table 1. With increasing disturbance the vehicle slows down as there is less sensor stimulation. As speculated above, the vehicle enters a stable orbit and as the RMSD for H2 (brightness) shows, there is very little movement towards the light source commensurate with the disturbance, leading to increased differences in brightness. The RMSD for H3 (angle) is very low, indicating that the vehicle is able to control for angle, and as a consequence, the orbit has a near constant turning radius. Looking at the normalised results, the NRMSD for H3 (angle) is 2 orders of magnitude less than that of H2 (brightness), so we consider this to be additional evidence that the angle to the light source *is indeed* the controlled variable. We conclude that the controlled variable is the angle of the vehicle to the light source, independent of the luminosity of the light source over a wide range of values.

Despite the large disturbance, the vehicle behaves as if it is comparing the perceived state of affairs with a reference perception of how the world *should* look. Yet this behaviour is not explicitly coded into the vehicle simulation, but emerges as a consequence of the interactions between its parts [3], including the type and arrangement of the sensors, the robot kinematics, and the virtual world itself.

6 CONCLUSION

A computational model of a system describes the algorithmic transformation of input data to output data, which can be understood as a Stimulus-Response (S-R) system. Where computers may be understood as disembodied information-processing machines with no loss of explanatory power, it is precisely the *embodiment* of robots that is key to our understanding of them. Even where a faithful simulation can be built entirely within the computer, as with Braitenberg's vehicles, the simulation still lacks an account of *purpose*. Control theory, in particular the Test for the Controlled Variable (TCV), provides us with methods for evaluating objective hypotheses about the purposes of an embodied robot from its observed behaviour.

The surprising thing about Vehicle 1 is that it is goal-directed at all. On the face of it, there is no internal decision making capability, yet on closer inspection this logic takes place because of the relative placement of the eye relative to the light source. Observations of its biased random walk reveal its long term *klinotaxis* where it moves towards the light source and comes to a halt at a position just beyond it. The big surprise of Vehicle 2 is that while it is able to hold a circular orbit at a constant distance from the light source in the face of extreme disturbance in luminosity, this is not the controlled variable because it cannot be perceived by the robot. This is merely a contingent fact that follows from the vehicles' ability to control its angle to the light source. Both Vehicles 1 and 2B are extremely simple, but it is their simplicity that is revealing. The analysis points to the power of the Test for the Controlled Variable (TCV) from Perceptual Control Theory. There is, of course, a lot more to PCT than explored here. It would be enlightening to create a new family of vehicles exploring the PCT approach to hierarchical control systems.

REFERENCES

[1] E. Abbott. *Flatland: A Romance of Many Dimensions*. Seeley & Co., 1884.

[2] V. Braitenberg. *Vehicles: Experiments in Synthetic Psychology*. The MIT Press, 1984.

[3] E. A. Di Paolo, M. Rohde, and H. De Jaegher. Horizons for the enactive mind: Values, social interaction, and play. In *Enaction: Toward a new paradigm for cognitive science*, pages 33–87. The MIT Press, 2010.

[4] J. Dupuy and M. DeBevoise. *The Mechanization of the Mind: On the Origins of Cognitive Science*. New French thought. Princeton University Press, 2000.

[5] H. Ginsborg. Kant's biological teleology and its philosophical significance. In *A Companion to Kant*. Blackwell, 2006.

[6] G. Guilbaud and V. MacKay. *What is Cybernetics?* Contemporary science books. Heinemann, 1959.

[7] D. Hawkins. The nature of purpose. In *Purposive Systems: Proceedings of the First Annual Symposium of the American Society for Cybernetics*, pages 163–179, 1968.

[8] D. Lees and A. Zilli. *Moths: A Complete Guide to Biology and Behavior*. Smithsonian Books, 2019.

[9] R. S. Marken. *Doing Research on Purpose: A Control Theory Approach to Experimental Psychology*. MindReadings.com, 2014.

[10] R. S. Marken. *The Study of Living Control Systems: A guide to doing research on purpose*. Cambridge University Press, 2021.
[11] H. Maturana and F. Varela. *Autopoiesis and Cognition: The Realization of the Living*. D. Reidel Publishing Company, 1980.
[12] E. Mayr. Cause and effect in biology. *Science*, 134(3489):1501–1506, 1961.
[13] M. Merleau-Ponty. *The Structure of Behavior*. Beacon Press, 1963.
[14] C. S. Pittendrigh. Adaptation, natural selection, and behavior. In ed. A. Roe and G. G. Simpson, editors, *Behavior and Evolution*, page 390?416. New Haven: Yale University Press, 1958.
[15] W. T. Powers. *Behavior: The Control of Perception*. Aldine Publishing Company, 1973.
[16] W. T. Powers. *Making Sense of Behavior*. Benchmark Publications Inc., 1998.
[17] A. Rosenblueth, N. Wiener, and J. Bigelow. Behavior, purpose and teleology. *Philosophy of Science*, 10:18–24, 1943.
[18] E. C. Tolman. *Purposive Behavior in Animals and Men*. Meredith Publishing Company, 1932.
[19] M. Washburn. *The Animal Mind: A Text-book of Comparative Psychology*. Macmillan, 1908.
[20] A. Weber, Francisco, and J. Varela. Life after kant: Natural purposes and the autopoietic foundations of biological individuality. *Phenomenology and the Cognitive Sciences*, 1:97–125, 2002.