Fast & Slow Thinking Machines

A Proposal for Dual-Process Reasoning in Machines

Fast and Slow Thinking Machines

A Proposal for Dual-Process Reasoning in Machines

Anaïs Siebers
Cognitive Science, Ruhr-University, Bochum, Germany,
anais.siebers@rub.de

12th October 2022

Abstract

A machine which rationally reasons like humans sounds like science fiction and is far from becoming part of everyday life. But the demand and necessity for explainable artificial intelligence which does not "decide" / "act" like a black box, but can explain its processes and reason about its decisions and judgements, strongly increases. In cognitive science, reasoning is also a currently strongly researched topic. The most dominant model at the moment is the dual-process theory, which attempts to combine the previously widespread symbolic, heuristic and probabilistic approaches. Processes of type 1 are fast, but biased, unconscious and strongly contextualised, whereas processes of type 2 are slow, conscious, controlled and unbiased. This paper proposes how machine reasoning on the basis of this dual-process theory could look like. The greatest difference to current attempts in machine reasoning is that the probabilistic black-box-like algorithms are accepted as processes of type 1. Future research is required concerning the transition and interplay between the processes and, if existent, the monitoring process controlling both processes and their interaction, to build a foundation for a successful implementation for machine reasoning.
Keywords: dual process theory; machine reasoning; logic reasoning; probabilistic reasoning; recognition; cognition

Contents

1 Introduction

Recently, the attempts to artificial intelligence (AI) and general AI (GAI) have focussed more and more on reasoning because current implementations of AI systems lack explainability, interpretability and accountability (Cyras et al., 2020, 4; Garcez et al., 2019, 1; Lin et al., 2019, 1; Zellers et al., 2019, 6727). The research of *explainable AI* (XAI) aims to make AI systems more understandable and acceptable to other users – humans as well as other machines. In the first wave of AI research, the focus was on rule-based systems, also called good old fashion AI (GOFAI) (Duan et al., 2020, 1). The second wave focussed more on expert systems and statistical learning. Both major approaches failed to be understandable, traceable and at the same time efficient. The decision processes are often referred to as black boxes (Cyras et al., 2020, 3). Thus, a machine who could reason and explain its decisions is desired. Therefore, research of the human phenomenon of reasoning is a good basis to advance *machine reasoning* (MR) – as *machine learning* (ML) was inspired by human learning as well. Both fields aim to "computationally mimic abstract thinking" (Cyras et al., 2020, 3) but have almost been advanced separately (Zhou, 2019, 1).

There is an interesting parallel between the history of machine reasoning and cognitive-psychological / -philosophical research of reasoning: in the beginning of reasoning (or Judgment and Decision Making (JDM)) research, the focus was on *logical approaches* like first order logic reasoning and then turned towards *probabilistic approaches* and *heuristics* (Baron, 2014, 138f.). No approach could yet on its own explain human reasoning behaviour. A very recent kind of theories are called *Dual-Process Theories* (DPT), which combine two different types of processes (Thompson, 2009, 171): a "fast and unconscious system 1 (S1)" and a "slow but thorough and conscious system 2 (S2)". Although there is a strong debate about the exact distinction between the two process types[1], there is a consensus that the first system is rather fast and requires few cognitive resources whereas the second system is slower but allows assessing and estimate decisions / actions (Pietzko, 2020).

1.1 Relevance

As outlined before, one reason to occupy oneself with machine reasoning is that current AI systems lack explainability. But another aspect is acceptance and the ease of use, as well as the endeavour to create socially cooperating multi-agent systems or autonomous agents (Cyras et al., 2020, 4). Reasoning would empower machines to make assumptions about social interaction (Zellers et al., 2019, 6727), comprehension, understanding physics (Zellers et al., 2019, 6727), situation prediction, commonsense (Duan et al., 2020, 2; Lin et al., 2019, 9) and improve human-robot interaction (Lin et al., 2019, 1). Furthermore, reasoning allows humans to go beyond perception and

[1] I agree with some researchers that the usage of the term systems suggests two distinct entities, which is misleading. In the following, I will only use the terms process of type 1 (P1) and process of type 2 (P2) because the transition between them seems to be flowing. Furthermore, DPTs are rather descriptive than normative and thus I accept the uncertainty this descriptive nature entails.

contributes to the "effectiveness and reliability of communication" (Mercier and Sperber, 2011, 58, 71f.). It is "often seen as a crucial driver for the real-world deployment of trustworthy modern AI systems" (Cyras et al., 2020, 4), but also as the "bottleneck of GAI" (Lin et al., 2019, 1).

In social settings, reasoning can also accelerate work because it divides cognitive load – it works by mutual verification: instead of checking one's own arguments and hypothesis all the time, other persons find flaws during communication (Mercier and Sperber, 2011, 73). This process improves knowledge and leads to better decision making (Mercier and Sperber, 2011, 57) – a task which will also be relevant for machines in future. Current systems are able to recognize and classify, but they still lack the ability to understand (Zellers et al., 2019, 6720).

The seamless integration of ML and MR would allow for advanced intelligent technologies to emerge – ML as a data-driven process with a focus on recognition and perception and MR as a knowledge-driven process, focussing on cognition (Zhou, 2019, 1). Reasoning and learning are essentially associated with intelligence (Bottou, 2014, 133). To complete these technologies, a last step concerning meta-cognitive processes like the influence of knowledge on learning and recognition has to be taken. It is a key competence which manifests itself in a reflective system (Ricco and Overton, 2011, 122).

1.2 Argumentative and Epistemic Goal

The central question and epistemic goal is to outline what we take to be the current state of machine reasoning and reasoning research to develop suggestions for *dual-process machine reasoning* (DPMR). The dual-process approach is chosen as basis, because it is very descriptive and thus comes closest to human reasoning, although this automatically implies challenges for the transfer to computationally realisable models and algorithms. Especially, since the exact distinction and change between processes is unclear. Furthermore, dual-process models are the currently most acknowledged approach.

To develop suggestions for DPMR, first, current theories of reasoning and decision making will be studied, focussing on DPT and the distinction and transition between the two process types. Then, current computational approaches to MR are outlined, taking a closer look at the approaches. This, subsequently, allows developing a list of suggestions for machine reasoning. Nevertheless, these are only a theoretical descriptive proposals and there still is a long way to go considering the current level of research. But they offer a starting point for discussion of how to realise the connection of type 1 and 2 processes of reasoning in machines.

2 Theories of Reasoning and Decision Making

As outlined in the introduction, there are basically three main systems in JDM and reasoning research (Bottou, 2014, 138f.): *first order logic reasoning* – also called *symbolic reasoning* (SR), *probabilistic reasoning* (PR), *heuristic reasoning* (HR), *dual-process theory*. There are more explanations of how reasoning works like causal reasoning, Newtonian mechanics, spatial reasoning or non-falsifiable reasoning systems, etc., but they are less hotly debated in recent history or concern only a specific application of reasoning.

The oldest and prevalent theory stemming from philosophy is logical / symbol reasoning. Since Aristotle, rationality has been considered to distinguish the homo sapiens from animals (Knauff and Spohn, 2021, 1). Although e.g. first order logic reasoning is very expressive and can be translated to natural language (thus it is understandable), it is insufficient because not every detail from natural language can be translated to logical statements (Bottou, 2014, 138). Given the assumption that rational is, if humans behave according to the reasoning system underlying human reasoning, SR fails since humans tend to systematically violate logical rules which would be very irrational behaviour (Oaksford and Chater, 2001, 349). One strategy to avoid this dilemma is to follow Chomsky, that rationality only is a *competence* and the *performance* of it might be defective. Another strategy is to reason that symbol reasoning is the wrong normative standard: people seem to reason with sound logical reasoning systems, but are restricted by cognitive limitations, which makes a probabilistic approach interesting (Oaksford and Chater, 2001, 349).

Probabilistic reasoning can be described as a "space of models formed by all the conditional probability distributions associated with a predefined collection of random variables" (Bottou, 2014, 138f.). The probabilities are also called a "measure of degree of belief" and represent the people's judgements that a proposition is in alignment with e.g. reality or coherence (Baron, 2014, 8). Compared to SR, PR has a continuous nature, a good performance and allows for uncertainty. Nevertheless, this has the drawback that it is even less expressive as first order logic (Bottou, 2014, 138f.).

Following experiments which displayed the inadequacy of the mathematical approaches of Bayesian PR (Baron, 2014, 3), another form of reasoning was proposed: heuristic reasoning. In HR, the judgement / reasoning process is biased and not all available information is considered, which allows for robust and efficient reasoning in uncertain or labour-intensive situations (Gigerenzer and Brighton, 2009, 107). Heuristics can be imagined as "rules of thumb" (Baron, 2014, 3). A lot of heuristics such as "Do no harm" (avoid and prevent harm) or "Status quo" (stick to the default) were discovered (Baron, 2014, 14). Two major arguments for heuristics are worth mentioning (Gigerenzer and Brighton, 2009, 110):

- **Accuracy-effort trade-off** Information and computation cost time and effort; therefore, minds rely on simple heuristics that are less accurate than strategies that use more information and computation.

- **Less-is-more effects** More information or computation can decrease accuracy; therefore, minds rely on simple heuristics in order to be more accurate than strategies that use more information and time.

But, humans do not solely rely on heuristics. Another theory of reasoning worth naming, although not mentioned earlier, is the *prospect theory* by Kahneman and Tversky (1979, 288). It includes heuristics where the probabilities of decisions are not given, which introduces a personal bias (Kahneman and Tversky, 1979, 289). A kind of prospect theory is *utility theory* which states that people always choose to increase their utility (Baron, 2014, 8). But compared to utility theory, prospect theory allows for choices that do not increase the utility but include other considerations.

All previously presented different aspects of reasoning. For centuries, distinctions between reasoning systems and approaches have been made: mainly the distinction between rule-based and probabilistic reasoning systems, which each fulfil different functions and explain specific behaviour (Sloman, 1996, 3). But to pinpoint one reasoning theory, which is more mighty and potentially leads to computational algorithms, has not yet been achieved (Bottou, 2014, 139). Recent publications encourage including different approaches and combine them to a single approach. Dual-process theory (DPT) explains reasoning and JDM by a combination of two processes which share properties with PR / HR (type 1 processes) and SR (type 2 processes) (Thompson, 2009, 171).

2.1 Dual-Process Theory

As shortly touched upon before, the dual-process theory describes two interacting processes / systems with distinct cognitive set-ups and usually also different functions (Ricco and Overton, 2011, 120). Processes of type 1 (P1) are fast, associative, unconscious, biased, autonomous, driven by heuristics. They are emotionally affected and give a default response if they are not interrupted by a higher cognitive reasoning process (process of type 2). Whereas processes of type 2 (P2) are, on the other hand, rather reflective, the opposite of P1, hypothetical, resource-intensive, slow and conscious, ... (Baron, 2014, 18; Evans and Stanovich, 2013, 223).

To continue, it is important to define how the dual-process theory is understood and used in this paper. There are many different interpretations and definitions of dual systems (Evans and Stanovich, 2013, 224): dual processes, dual types, dual systems, modes of processing, autonomous set of systems, ... In this paper, a combination of dual processes and dual types is used. *Dual processes* often refer to and are equated with dual types. It assumes that there are two forms of processing for cognitive tasks. The terminology *dual types* "implies that the dual processes are qualitatively distinct. Type 1 processes are (broadly) intuitive and Type 2 processes reflective" (Evans and Stanovich, 2013, 225).

2.1.1 Fast Thinking

Fast thinking is mostly done by P1. This process has a high capacity and works independently of neither working memory nor cognitive abilities (Evans, 2011, 87). Hence, it also does not require a lot of attention and works unconsciously. P1 is highly associative and works in episodic and semantic or procedural memory (Ricco and Overton, 2011, 120). Therefore, it is embedded in a highly problem-centered context There are many biases which are probably caused by heuristics. These can be avoided by learning or training and intervention of P2 (Baron, 2014, 19).

2.1.2 Slow Thinking

On the other hand, slow thinking is mostly executed by P2. P2 is described to perform analytic, deductive, inductive, abductive tasks (Zhou, 2019, 1) and works on a more abstract and context-unspecific, domain-general level and corresponding representations (Ricco and Overton, 2011, 120). The process is resource-intensive, controlled, conscious and limits cognitive abilities. Due to the dependence on cognitive capabilities (Evans, 2011, 87), P2 is highly individual and correlated with intelligence (Ricco and Overton, 2011, 120). It is supposed that P2 is the part of the mind, distinguishing human behaviour from mere animal behaviour – enabling humans to simulate and meta-represent content and override intuitions (Evans and Stanovich, 2013, 236). Essential to P2 seems to be the ability to abstract and decouple from primary contextualised representations and keep secondary representations (Evans and Stanovich, 2013, 237).

2.1.3 The Monitoring Process

Having two process types dealing with cognitive work necessarily raises the question of how these are interdependent, work together and which processes or events elicit a change of the process type (Sloman, 1996, 3; Thompson, 2009, 171). Psychologically, recognition and reasoning in humans is rather entangled and not disjoin (Zhou, 2019, 3). It is generally accepted that a – as Freud calls it – "primary process thought" is succeeded by a secondary purposive thought (Sloman, 1996, 17). The *default-interventionist* view claims that the default rapidly prompted P1 is intervened by a reflective P2 (Evans and Stanovich, 2013, 237). Most reasoning is assumed to take place as P1, but the crucial question is about the balance between P1 and P2, since P2 usually is slower than P1 and intervening requires a temporal correspondence (Oaksford and Chater, 2001, 356)

It seems that next to P1 and P2 a *meta-cognitive judgement* or *second-order judgement* is needed to decide when to use P1 and when to switch to P2 (Thompson, 2009, 171, 190). This meta-cognitive judgement will further be referred to as *monitoring process* (MP). It could observe and handle cognitive capacity and performance, controlling and monitoring the processes and resources (Thompson, 2009, 191).

Thompson (2009, 175f.) describes different possible factors, which are judgements

contributing to the MP such as:

- **Feeling of Rightness (FOR)** A feeling that provides "a means to assess the output of one's cognitive processes and determine whether further action should be taken. Under this view, the explanation for the compellingness of many cognitive illusions is that the heuristic response is generated with a strong intuition that the answer is correct."
- **Feeling of Familiarity (FOF)** The feeling one has if one e.g. recollects a correct answer from memory. This phenomenon feels familiar.
- **Feeling of Knowing (FOK)** The feeling that one knows something without actually retrieving the knowledge.
- **Judgment of Learning (JOL)** The judgement that one has correctly learned something and will be able to accurately recall it later.
- **Fluency of Processing (FOP)** There are various influencing factors such as the aesthetic pleasure or the assessment of truth and influences the perception of difficulty.

The author suggests that the individuals differ regarding their MP-skills (Thompson, 2009, 180), which implies that the individual differences supposed in P2-capacities are not necessarily given, but could be a consequence of the differences in their MP-abilities. Another consequence is that a strong FOR correlates with the acceptance of P1's heuristic results whereas a low FOR starts the P2 process (Thompson, 2009, 179).

2.2 Challenges

One challenge the dual-process theory has to face is that the processes merely describe process styles and are not clearly defined. Furthermore, they are yet unmapped to actual neural processes in the brain and the clustering of labels for the processes take place while there might be flowing transitions (Evans and Stanovich, 2013, 226f.). Another problem is that the processes are competing: since P1 is faster than P2, it is questionable how P2 can interrupt P1. An MP would lessen this problem, but it would nevertheless require parallel processing and a temporal connection (Evans and Stanovich, 2013, 237). Another general problem is, that a lot of reasoning theories are normative but therefore inadequate, whereas a computationally feasible model requires a descriptive but applicable model (Elqayam and Evans, 2011, 233, 248; Gigerenzer and Sturm, 2012, 244).

3 Machine Reasoning

After taking a look at the theoretical dual-process framework for reasoning, the current state of research concerning MR and ML is outlined. The dual-process the-

ory is a descriptive theory aiming to explain how people reason, judge and decide. It is formulated in cognitive psychology and composed of heuristics, strategies and mathematical models (Baron, 2014, 5). Current attempts at machine reasoning try to combine "the precision and productive power of symbolic rules with the learning, automatic generalization, and constraint satisfaction power of connectionist associations" (Sloman, 1996, 19) just like the dual-process theory tries to join P1 and P2.

3.1 Approaches and Methods

Since there is a lot of research done in this area, a lot of different approaches are evaluated and studied. There are various attempts to implement machine reasoning (Duan et al., 2020, 1). They differ not only in their implementation, but also in their definition of reasoning and the reasoning questions they aim to answer. Bottou (2014, 133), for example, defines reasoning as: "algebraically manipulating [of] previously acquired knowledge in order to answer a new question". Three types of explanations given by AI systems can be differentiated (Cyras et al., 2020, 46): *Attributive explanations* which explain why an AI system returns a specific output given a specific input concerning its association and attribution of the system with the output. *Contrastive explanations* explain why an AI system choses one output in comparison to another ouput, reasoning for and against the alternative outputs. *Actionable explanations* are explanations about what the user of the system can do differently to get another output. Thus, a direct comparison is difficult, but in the following, different approaches will be listed. In general, machine reasoning aims to build an explainable and interpretable AI system (Duan et al., 2020, 1). The goal is to connect lower-level information processing (which is strongly data-driven and comparable to recognition) and high-level abstract representations (which are knowledge-driven and comparable to cognition) (Garcez et al., 2019, 1).

Various different approaches have been proposed and published[2] (Duan et al., 2020, 1):

- Corresponding to symbolic reasoning theories, there are **symbolic reasoning methods** which use symbolic logic to represent and argue with knowledge. Used algorithms are e.g. the truth-table approach, inference rules, resolution, non-monotonic logical reasoning, forward and backward chaining, probabilistic logic programming (PLP) and statistical relational learning (SRL) (Zhou, 2019, 1). They cannot well handle uncertainty, although PLP and SRL are attempts to bridge between symbolic and probabilistic approaches.

- The counterpart to probabilistic reasoning theories are the **probabilistic reasoning methods** which mainly use probabilities, but partly also a bit of symbolic logic. They can deal with uncertainty and can be easily interpreted, but in large spaces these methods lead to combinatorial explosion and are inflexible due to the discrete and finite symbolic representation. Used algorithms are Bayesian Networks or the Markov Logic Network.

[2]this list does not claim to be complete

- Another form of machine reasoning are **neural-symbolic reasoning methods**. Here, knowledge symbols are represented as mathematical representations (like vectors and tensors) to permit large and effective learning because the components are differentiable. A drawback is that it is not easily interpretable. Nevertheless, the system is more readable than black-box neural networks, because it usually realises inferences as a chain of modules which allows seeing the time-line of different functions which were called.

- Compared to all previously mentioned methods, **neural-evidence reasoning methods** allow for communication to the outside environment and thus learn and search for evidence for reasoning outside the scope of the model. They are least developed.

Another aspect – not much considered in the methods above – is commonsense reasoning. As outlined in the introduction of the paper, humans evaluate the reasons and actions of others as rational or irrational with the purpose of distinguish correctness and incorrectness of reasoning processes. But most of the time, this happens unconsciously in form of commonsense knowledge and evaluation (Chater and Oaksford, 2000, 93f.) and seems trivial to humans (Sap et al., 2019, 3027). Moreover, human-level performance will only be reached by basic knowledge of the world (Davis and Marcus, 2015, 92). If machines reason, they also have to comply to these "unwritten rules" of common sense humans share. A recent proposal for commonsense reasoning was published by Sap et al. (2019, 3027). They aim for "simple and explainable commonsense reasoning" which exceeds task-specific correlations (Sap et al., 2019, 3027).

Moreover, also relevant for MR is grounding. Previously mentioned approaches were all based on existing knowledge bases – often engineered by hand by experts or large communities (Davis and Marcus, 2015, 92) – and a properly defined problem. But in real-world applications of reasoning, the pipeline of machine reasoning contains more than just reasoning about the knowledge. For understanding visual scenes or spoken text, the content first has to be grounded – extracting the needed information and translating them to the correct representation so that the machine is able to reason at all (Zellers et al., 2019, 6721). Reasoning can also help to improve the performance of grounding by checking for coherence and generalise over different domains (Sridharan and Mota, 2022, 1, 32).

3.2 Limitations

There already are a lot of approaches and methods to realise MR, but they are all still limited. Although there are few forms of (commonsense) reasoning or domain-specific reasoning, progress has been very slow. The techniques used are not yet sufficient (Davis and Marcus, 2015, 92). As indicated before, each method is relatively domain-specific and focuses on fixed knowledge (Cyras et al., 2020, 46). Most efforts focused on a knowledge of "what" and less on e.g. procedural knowledge

of "how" (Sap et al., 2019, 3033). Additionally, there is a dilemma between interpretability and performance, similar to the trade-off compensated by heuristics in P1. Black-box systems are much more performant compared to abstract, generalised and interpretable systems (Duan et al., 2020, 1f.).

Apart from the problem that representation and inference rules first have to be realised efficiently, not every knowledge can be formulated at the appropriate level of abstraction and much knowledge is implicit (Davis and Marcus, 2015, 98). For example, a robot housekeeper should realise that the pet of the child (a rat) is not a vermin, but a pet loved by the child, but that the rat running around in the kitchen is not the same and is a vermin. Since commonsense reasoning is very natural to humans and happens unconsciously and automatically, it is very difficult to get conscious access to it (Chater and Oaksford, 2000, 94; Davis and Marcus, 2015, 97). A well-known problem, the *frame problem* describes the difficulty to break down the knowledge to an adequate level (Chater and Oaksford, 2000, 94). The integration of recognition and cognition requires complex inferences, which state-of-the-art-systems are still struggling with (Zellers et al., 2019, 6721).

Lastly, the monitoring process imposes a big challenge for machine reasoning. Just as for the cognitive-psychological reasoning models, the approaches to machine reasoning have not yet much discussed and researched how this "middle layer, already a form of reasoning, but not yet formal or logical" could be realised and when it intervenes (Bottou, 2014, 134).

4 Proposals to Advance Machine Reasoning

After summarising the current state of reasoning research and machine reasoning, this section will propose some approaches to advance machine reasoning. They are inspired by the DPT of reasoning, as well as they incorporate / integrate state-of-the-art algorithms and approaches to ML and MR. As indicated in the beginning, the dual-process approach as basis for the proposals is chosen, because until now it captures human reasoning best by its descriptive nature. To better understand the suggestions, the process of face recognition is chosen exemplary to explain the interworking of the processes.

In accordance to the DPT, there should be two different kinds of processes which differ in their main features, although there are no strict borders between the process types. Thus, specific algorithms or models are not designated to one process type.

The first process type is like P1 and contains heuristic and probability methods. Additionally, where applicable, it could also contain some form of short-term memory or – computationally spoken – cache which adds a strong bias to the most recent results or ideas. In humans, very recently acquired knowledge also dominates. Hence, the system is strongly biased and very fast as P1. Given the example of face recognition, neural networks could retrieve the name of a face. Similarly to human face recognition, there is no long reasoning process required.

Following this approach, this implies that P1 are realised by algorithms which act as a black-box and do not allow for traceability, understandability and explainability. Furthermore, the result can be very irrational like recognising a dead person in the face of a living person. Depending on the algorithm used, the answers can not be retraced and even more, not even a decision pattern can be detected, leaving a very untrustworthy random impression.

So what about these aspects which are currently the strongest motivation to increase machine reasoning abilities? As for humans, we could engage more thorough decision making and reasoning in machines if confronted with a question. If a person identifies a person wrongly or does not recognise the person, others will ask about the person's features like e.g. "Did you not see his red curly hair?" or "Why did you think it was Sam?". These are obvious situations to engage P2. The same could be valid for machines. Their P2 would entail commonsense knowledge, be deductive or inductive and work on traceable and understandable symbolic reasoning processes. Just like reasoning with P2 in humans requires a lot of resources, P2 is resource-intensive and slow in machines. One important feature has to be, that these processes have to be able to explain P1. But this does not require the P1 to be understandable. In the given example of face recognition, an answer to not recognising a person could be "[name] looks very different than I remember" or "I saw [name of other person] very often and mistook [name of other person] for [name of seen person]". In terms of human-machine interaction, these arguments would represent a sufficient explanation. Furthermore, in regard to future development, the gained knowledge and the finding of these P2 could influence the training and learning of P1 in a meta-cognitive and reflective way.

Does this mean that every cognitive process is, by default, a P1 and P2 is only activated upon question? No, there still is and has to be a monitoring processes, switching between P1 and P2. This is the area which is still the least researched. Therefore, these proposals are only starting points, hoping that they will enrich and advance the debate and research in the transition between P1 and P2.

Thompson (2009) described various feelings like the feeling of familiarity or knowing, which could determine the likeliness and the certainty with which humans answer with P1 or whether they engage P2. Many algorithms which would fall under P1 and act like black-boxes are probability based. Therefore, the result is a probability which can be used as a measure of certainty, and, if too low, lead to an engagement of P2. Since humans expect "rational behaviour" which folk-psychologically can be described as coherent and consistent behaviour, commonsense knowledge can be used to check P1's decisions / results. Commonsense knowledge and reasoning can also trigger P2, if reality and expectation mismatch. An example would be if P1 during face recognition comes to the result that the face belongs to a dead person. The commonsense knowledge, that the person is dead, would highlight an incongruence between expectation and reality as well as be inconstant with the knowledge, that dead people cannot walk around. As a consequence, this would start P2. Another reason for P2 activation is if P1 cannot deal with the cognitive input. But the timing problem of P1, P2 and the monitoring problem remains.

To summarise, these are first ideas in the hope that they can build a foundation for further research and discussions how machine reasoning can be improved and should develop in future in accordance with and based upon the dual-process theory of human reasoning. They feature two main types of processes which are monitored and controlled by another process and parameters like coherence or certainty.

5 Conclusion, Implications and Outlook

Integrating reasoning in attempts to produce advanced intelligent systems is relevant for current and future research. It finds application in many real-world scenarios like expert systems and diagnosis (medical, mechanical, ...), knowledge base completion, human-machine interaction, fact checking, search, ... (Duan et al., 2020, 3) Interdisciplinary work is key to understand the cognitive processes and techniques and to be able to implement them without necessarily replicating them (Davis and Marcus, 2015, 103). Currently, the dual-process theory best reflects human reasoning, although it mainly on a descriptive level, lacking information about the monitoring process and transition between the two process types. If reasoning should be implemented in machines to achieve advanced intelligent technologies, using the dual-process theory as basis is obvious. Until now most approaches in machine reasoning resemble the development of research in cognitive science with current attempts to bridge between symbolic and probabilistic methods. Nevertheless, the approaches aim at explaining all processes. A new approach, which this paper offers, is to accept unexplainable, untraceable, biased processes. The only drawback in implementing biased machines is the human attitude in the infallibility of computers. Otherwise, humans are used to biased and heuristical behaviour in other humans: It seems that usually cognitive processes happen unconsciously and without logical inferences (Bottou, 2014, 140), which indicates that machines do not need to exclusively have explainable processes.

Most problematic and difficult is the design of the monitoring process. This paper raises first ideas such as coherence or certainty as indicators that P2 engagement is required, but with continuing research in cognitive science as well as artificial intelligence, the design and description of the monitoring process will probably become more certain. Another open aspect is that not much is known about social reasoning and rationality. In future scenarios where machines might also work together with humans in social settings, rationality and reasoning in social settings becomes very important. These cognitive processes seem to differ from individual reasoning (Knauff and Spohn, 2021, 51). Further research in meta-cognitive and reflective processes could also enhance the influence of P2 on P1.

To summarise, the current development in the area of machine reasoning indicates that human-like reasoning is still far away. Further research in cognitive science as well as artificial intelligence, machine reasoning and machine learning is required to address the dichotomy of the processes and their interplay. Very domain-specific reasoning is getting better, but general (commonsense) reasoning still has a long way

to go. Main advances need to be made concerning the difference and interplay of both processes. This could then build a good foundation for computational feasibility and an implementation of reasoning in machines.

References

Baron, J. (2014). Heuristics and biases. In Zamir, E. and Teichman, D., editors, *The Oxford handbook of behavioral economics and the law*, pages 3–27. Oxford University Press.

Bottou, L. (2014). From machine learning to machine reasoning: An essay. *Machine Learning*, 94(2):133–149.

Chater, N. and Oaksford, M. (2000). The Rational Analysis Of Mind And Behavior. *Synthese*, 122(1/2):93–131.

Cyras, K., Badrinath, R., Mohalik, S. K., Mujumdar, A., Nikou, A., Previti, A., Sundararajan, V., and Feljan, A. V. (2020). Machine Reasoning Explainability. arXiv:2009.00418 [cs].

Davis, E. and Marcus, G. (2015). Commonsense reasoning and commonsense knowledge in artificial intelligence. *Communications of the ACM*, 58(9):92–103.

Duan, N., Tang, D., and Zhou, M. (2020). Machine Reasoning: Technology, Dilemma and Future. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, pages 1–6, Online. Association for Computational Linguistics.

Elqayam, S. and Evans, J. S. B. T. (2011). Subtracting "ought" from "is": Descriptivism versus normativism in the study of human thinking. *Behavioral and Brain Sciences*, 34(5):233–248.

Evans, J. S. (2011). Dual-process theories of reasoning: Contemporary issues and developmental applications. *Developmental Review*, 31(2-3):86–102.

Evans, J. S. B. T. and Stanovich, K. E. (2013). Dual-Process Theories of Higher Cognition: Advancing the Debate. *Perspectives on Psychological Science*, 8(3):223–241.

Garcez, A. d., Gori, M., Lamb, L. C., Serafini, L., Spranger, M., and Tran, S. N. (2019). Neural-Symbolic Computing: An Effective Methodology for Principled Integration of Machine Learning and Reasoning. arXiv:1905.06088 [cs].

Gigerenzer, G. and Brighton, H. (2009). Homo Heuristicus: Why Biased Minds Make Better Inferences. *Topics in Cognitive Science*, 1(1):107–143.

Gigerenzer, G. and Sturm, T. (2012). How (far) can rationality be naturalized? *Synthese*, 187(1):243–268.

Kahneman, D. and Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, 47(2):263.

Knauff, M. and Spohn, W., editors (2021). *The handbook of rationality*. The MIT Press, London.

Lin, B. Y., Chen, X., Chen, J., and Ren, X. (2019). KagNet: Knowledge-Aware Graph Networks for Commonsense Reasoning. arXiv:1909.02151 [cs].

Mercier, H. and Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences*, 34(2):57–74.

Oaksford, M. and Chater, N. (2001). The probabilistic approach to human reasoning. *Trends in Cognitive Sciences*, 5(8):349–357.

Pietzko, J. (2020). Der Kobra-Effekt. *Philosophie Magazin*, (online).

Ricco, R. B. and Overton, W. F. (2011). Dual systems Competence ←-→ Procedural processing: A relational developmental systems approach to reasoning. *Developmental Review*, 31(2-3):119–150.

Sap, M., LeBras, R., Allaway, E., Bhagavatula, C., Lourie, N., Rashkin, H., Roof, B., Smith, N. A., and Choi, Y. (2019). ATOMIC: An Atlas of Machine Commonsense for If-Then Reasoning. arXiv:1811.00146 [cs].

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, 119(1):3–22.

Sridharan, M. and Mota, T. (2022). Combining Commonsense Reasoning and Knowledge Acquisition to Guide Deep Learning in Robotics. arXiv:2201.10266 [cs].

Thompson, V. A. (2009). Dual-process theories: A metacognitive perspective. In *In two minds: Dual processes and beyond*, pages 171 –195. Oxford University Press.

Zellers, R., Bisk, Y., Farhadi, A., and Choi, Y. (2019). From Recognition to Cognition: Visual Commonsense Reasoning. arXiv:1811.10830 [cs].

Zhou, Z.-H. (2019). Abductive learning: towards bridging machine learning and logical reasoning. *Science China Information Sciences*, 62(7):76101.

Competing Interests

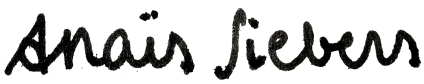
This paper was written in the context of the course "Reasoning, Normativity and Cognitive Biases" at the Ruhr-University Bochum. The paper is rewarded with credit points and evaluated with a grade.

Statement of Authorship

I hereby certify under oath that the paper I am submitting is entirely my own original work except where otherwise indicated. I have not used any auxiliary means other than those listed in the bibliography or identified in the text and any use of the works of any other author, in any form, is properly acknowledged at their pint of use with indication of the source.

Anaïs Siebers
(firstname, surname)

12.10.2022, Siegburg
(date and place)



(signature)