Thinking Fast And Slow In Human-Centered AI

Adam Dahlgren Lindström, Wendy E. Mackay, Virginia Dignum

Thinking Fast And Slow In Human-Centered AI

Adam Dahlgren Lindström[1]
Wendy Mackay[2]
Virginia Dignum[1]

[1]Department of Computing Science, Umeå University
[2]Inria, Paris-Saclay, Collège de France
dali@cs.umu.se, mackay@lri.fr, virginia@cs.umu.se

Abstract

Thinking Fast and Slow (Kahneman 2011) provides a simple mental model of how human intelligence builds on components with complementing responsibilities and capabilities. In computer science in general, and artificial intelligence research in particular, these ideas are used to inspire new methods and architectures. We argue that many of those methods use the concept of *Thinking Fast and Slow* as a token reference, while not living up to the definitions of dual-process systems from psychology. For instance, 'fast' is seen as synonymous with neural, and the autonomy of 'fast' seen in, e.g., fixed social interactions of social agents is lost. We further highlight that these ideas are misused in saying that humans are flawed and AI systems can fix that. Given that human bias is highly context-dependent, such simplistic applications of dual-process theory to AI are likely to fail. Thus, the narrative that AI systems will provide users with rationality is flawed. In a work in progress, we survey and categorise (mis)use of prospect theory and other dual-process theories. Building AI systems on the ideas of Tversky and Kahneman is a step in the right direction for a more human-centered artificial intelligence. With this work we want to emphasise the many things to consider in building systems that are *Thinking Fast and Slow*, and that the community is only scratching that surface.

In this expression of interest we outline ongoing work on the impact of *Thinking Fast and Slow* on AI research. We each work in different areas, that now overlap because of the rise of human-centered AI. Between human-computer interaction, machine learning and reasoning, and social and ethical AI, we take advantage of those different perspectives to challenge and inspire different kinds of human-centered AI. We aim to reach a broad audience of AI researchers that can use these guidelines for new methods and architectures.

The work of Tversky and Kahneman (T&K), *Thinking Fast and Slow*, and other dual-process theories, have influenced the discourse in AI and development of new systems and methods. For example, the field of neuro-symbolic AI is heavily influenced by the mental model of fast and slow components, such as artificial neural networks in combination with logic programming. Raedt et al. claims that "(Kahneman 2011) has put the quest for neural symbolic computation (Garcez, Broda, and Gabbay 2012; Garcez et al. 2015; Hammer and Hitzler 2007) high on the research agenda". Complementary to *Thinking Fast and Slow*, dual-process theory contain other important works (e.g. (Sloman 1996; Evans 1996; Stanovich and West 2000)). They also provide other insights, e.g. Stanovichs' tripartite Reflective, Algorithmic, and Autonomous minds (Stanovich 2009). Evans and Stanovich define fast and slow thinking as;

> *Our preferred theoretical approach is one in which rapid autonomous processes (Type 1) are assumed to yield default responses unless intervened on by distinctive higher order reasoning processes (Type 2). What defines the difference is that Type 2 processing supports hypothetical thinking and load heavily on working memory.* (Evans and Stanovich 2013)

In many cases, neuro-symbolic architectures do not reflect the systems described in many dual-process theories and would not fall under the above or other definitions. We argue that citing *Thinking Fast and Slow* is sometimes used as a token reference without clear anchoring in the theory. Simply building a dual-component system is not sufficient to realise the equivalence of cognitive dual-process systems.

We often see papers that begin with the assumption that, because T&K can show that people behave 'irrationally' in experiments that are explicitly designed to highlight specific kinds of bias, that humans are flawed and that somehow AI will solve the problem, because somehow the AI has access to a mathematically valid ground truth, and will be able to 'fix' the human biases. This is seen in, e.g., work on debiasing in interpretability (Kliegr, Štěpán Bahník, and Fürnkranz 2021). This is wrong on many levels, not the least of which is that AI does not have access to such ground truth. In reality, human behavior is neither simple nor rational, but derives from a complex mix of mental, physical, emotional, and social aspects. Instead of 'fixing' human behavior, AI systems should support reasoning in situations in which not all alternatives or consequences can be foreseen. This is incompatible with traditional approaches to rational AI, and lead to systems unable to accurately model a wide range of human behaviors. AI systems cannot suffice with only a model of fast reasoning and should support the ability to hold and deal with inconsistent beliefs and have the ability to fulfill several roles, and pursue seemingly incompatible goals concurrently (Dignum 2017). Humans clearly have biases, but

they are highly context-dependent, and simplistic applications of T&K are likely to fail. T&K's goal was not to prove that people are idiots that need to be fixed, but rather that we can use various techniques to create cognitive illusions that indicate (at least partially) how (some) people reason about the world. From a HCI perspective, there is a casual misuse of T&K to as though users can be 'fixed' with unbiased AI systems. With respect to content, T&K have designed some very creative experiments, but that they should be viewed in the same way that we view other psychology experiments on visual and auditory illustions — they help us to understand how the human cognitive, visual or auditory system works, but should not be treated as if we are saying that people are flawed. The advantage of clear, well-designed experiments is that they extract the features of interest, but they cannot be overgeneralized beyond. What the AI community can learn from the HCI community, is that the algorithm is never enough, it has to be examined in a human context.

We argue that for Human-Centered AI (HCAI), what we outline above together with the misalignment of dual-process theory and its instantiation in AI systems form a shaky foundation. For example, neuro-symbolic methods with neural networks used as input to symbolic components are still sensitive to the same type of biases that the AI community in general is tackling. Bias exists in humans' fast thinking, but is countered by slow thinking and adjusted over time accordingly. Therefore, the relationship between neural and symbolic components cannot be unidirectional. Similarly, not all cognitive theories on this topic suggests two components, but three or a plethora of components. Neuro-symbolic AI would benefit from research in this direction, acknowledging that there is existing work on this, such as Genie (Campagna et al. 2019) and other federated virtual assistants.The mental model of neural networks as neurons is powerful and has led to impressive results, but also has substantial downsides. Hence, we are cautious with respect to *Thinking Fast and Slow* for similar reasons.

An important distinction to make, is that fast is not neural. On the one hand we have the routines and heuristics that humans use to fast think (social practices fall into this category), and in AI it is about the ways fast can be implemented in computers. One is about the action or behaviour, the other about its implementation. With neural networks for image recognition, it is the same computation to recognise an image that clearly shows a cat, versus an image with occlusion and only partial clues as to what is shown. For humans, the later can engage slow thinking to reason about what is seen. For neural networks, there is no such distinction.

When building on *Thinking Fast and Slow*, it is also important to recognise the critique of dual-process theory from psychology scholars. Following is a list of critiques that are particularly interesting for neuro-symbolic AI: 1. it is a simple mental model providing a nice story, but does not hold under scrutiny (Keren and Schul 2009), 2. it is not possible to clearly distinguish between fast and slow in reality (Kruglanski and Gigerenzer 2018; Osman 2004) , and 3. only two components is not sufficient, cognitive systems build on many more subsystems (Stanovich 2009; Keren and Schul 2009). In a response to replication critique (Schimmack, Heene, and Kesavan 2017), Kahneman agrees that *What the [article] gets absolutely right is that I placed too much faith in underpowered studies.* (Kahneman 2017), referring to studies on behavioural priming. These critiques are helpful in guiding the design of AI systems.

Our main points can be summarised as:

1. T&K help us understand human cognitive, visual or auditory systems, but there is a casual misuse in that we can create unbiased AI systems and somehow 'fix' users.
2. Fast is not synonymous with neural. Social agents have fixed 'autonomous' social interactions on given ques.
3. Bias in 'fast' AI systems must be considered, even when slow components are introduced.
4. Dual-process theory is more than *Thinking Fast and Slow*, much can be learnt from other work and critiques.
5. *Thinking Fast and Slow* is a simple mental model, much like artificial neural networks are of the brain. Mental models with a life of its own outside the scope of the experiments it came from are problematic.
6. Neuro-symbolic AI rarely fulfill psychology definitions of fast/slow systems, they only borrow the concept of fast/slow.

We are working on a survey of the influence of *Thinking Fast and Slow* and other dual-process theories on AI research. The aim of this project is to categorise how they are used, and produce a set of guidelines for future research. We acknowledge works such as (Booch et al. 2021) proposing a new research direction based on *Thinking Fast and Slow*, and (Bonnefon and Rahwan 2020) which highlights the problem that computers do not 'think fast and slow' in the sense humans do, but they are built to convey that they do. This has implications for how user interact with such systems, similar to those discussed in our work.

Dual-process theory in Neuro-Symbolic AI

Sarker et al. gives a categorisation of neuro-symbolic AI, where none of the categories seem to correspond to a dual-process theory. For example, neuro→symbolic and [neuro[symbolic]] correspond to neural networks feeding into symbolic frameworks either sequentially or as subroutines, respectively. The later is closer to an analogy to dual-process, but the 'fast' system, i.e. neural network, has no autonomy. An example of neuro→ symbolic is the Neuro-Symbolic Concept Learner (NSCL) (Mao et al. 2019), in which a neural network process images and fed visual representations to a (quasi-)symbolic reasoner. Stammer, Schramowski, and Kersting extends NSCL with explanatory interactive learning, where symbolic explanations are used as reasoning to provide feedback for the system to learn, rather than purely via back-propagation. This is a step closer towards *Thinking Fast and Slow*, but the fast part still lacks autonomy, formulated by the authors as *"[..] they combine system 1 and 2 characteristics"*. A [neuro[symbolic]] system is DeepProbLog (Manhaeve et al. 2018), where neural networks are used as submodules in a logic programming language. Although neither Mao et al. nor Manhaeve et al. cite Kahneman, the surrounding literature does categorises them under the system 1/2 umbrella.

References

Bonnefon, J.-F.; and Rahwan, I. 2020. Machine Thinking, Fast and Slow. *Trends in Cognitive Sciences*, 24(12): 1019–1027.

Booch, G.; Fabiano, F.; Horesh, L.; Kate, K.; Lenchner, J.; Linck, N.; Loreggia, A.; Murgesan, K.; Mattei, N.; Rossi, F.; and Srivastava, B. 2021. Thinking Fast and Slow in AI. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17): 15042–15046.

Campagna, G.; Xu, S.; Moradshahi, M.; Socher, R.; and Lam, M. S. 2019. Genie: A generator of natural language semantic parsers for virtual assistant commands. In *Proceedings of the 40th ACM SIGPLAN Conference on Programming Language Design and Implementation*, 394–410.

Dignum, V. 2017. Social Agents: Bridging Simulation and Engineering. *Commun. ACM*, 60(11): 32–34.

Evans, J. 1996. Rationality and Reasoning. *Cognititive Psychology*.

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. PMID: 26172965.

Garcez, A. d.; Besold, T. R.; De Raedt, L.; Földiak, P.; Hitzler, P.; Icard, T.; Kühnberger, K.-U.; Lamb, L. C.; Miikkulainen, R.; and Silver, D. L. 2015. Neural-symbolic learning and reasoning: contributions and challenges. In *2015 AAAI Spring Symposium Series*.

Garcez, A. S. d.; Broda, K. B.; and Gabbay, D. M. 2012. *Neural-Symbolic Learning Systems: Foundations and Applications*. Springer Science & Business Media.

Hammer, B.; and Hitzler, P. 2007. *Perspectives of neural-symbolic integration*, volume 77. Springer.

Kahneman, D. 2011. *Thinking, fast and slow*. Macmillan.

Kahneman, D. 2017. Response to reconstruction of a train wreck: How priming research went off the rails.

Keren, G.; and Schul, Y. 2009. Two is not always better than one: A critical evaluation of two-system theories. *Perspectives on psychological science*, 4(6): 533–550.

Kliegr, T.; Štěpán Bahník; and Fürnkranz, J. 2021. A review of possible effects of cognitive biases on interpretation of rule-based machine learning models. *Artificial Intelligence*, 295: 103458.

Kruglanski, A. W.; and Gigerenzer, G. 2018. Intuitive and deliberate judgments are based on common principles. In *the Motivated Mind*, 104–128. Routledge.

Manhaeve, R.; Dumancic, S.; Kimmig, A.; Demeester, T.; and De Raedt, L. 2018. Deepproblog: Neural probabilistic logic programming. *Advances in Neural Information Processing Systems*, 31.

Mao, J.; Gan, C.; Kohli, P.; Tenenbaum, J. B.; and Wu, J. 2019. The Neuro-Symbolic Concept Learner: Interpreting Scenes, Words, and Sentences From Natural Supervision. In *International Conference on Learning Representations*.

Osman, M. 2004. An evaluation of dual-process theories of reasoning. *Psychonomic bulletin & review*, 11(6): 988–1010.

Raedt, L. d.; Dumančić, S.; Manhaeve, R.; and Marra, G. 2020. From Statistical Relational to Neuro-Symbolic Artificial Intelligence. In Bessiere, C., ed., *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, 4943–4950. International Joint Conferences on Artificial Intelligence Organization. Survey track.

Sarker, M. K.; Zhou, L.; Eberhart, A.; and Hitzler, P. 2021. Neuro-symbolic artificial intelligence: Current trends. *arXiv preprint arXiv:2105.05330*.

Schimmack, U.; Heene, M.; and Kesavan, K. 2017. Reconstruction of a train wreck: How priming research went off the rails. *Replicability Index*.

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

Stammer, W.; Schramowski, P.; and Kersting, K. 2021. Right for the right concept: Revising neuro-symbolic concepts by interacting with their explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3619–3629.

Stanovich, K. E. 2009. Distinguishing the reflective, algorithmic, and autonomous minds: Is it time for a tri-process theory?

Stanovich, K. E.; and West, R. F. 2000. Individual differences in reasoning: Implications for the rationality debate? *Behavioral and brain sciences*, 23(5): 645–665.