The handshake problem for human-like coordination systems.

Eugene Philalithis[1,2]

[1]School of Philosophy, Psychology & Language Sciences, University of Edinburgh
[2]School of Informatics, University of Edinburgh

Abstract

Recent work in explanatory machine learning highlights the value of symmetry between human and AI contributions in collaborative settings, in the form of a 'cognitive window' of optimal complexity that the AI collaborator must aim for. In this short paper, I argue the cognitive window is potentially mediated by processing styles, some fast and some slow, which the AI must also model to achieve the right 'handshake' with a human agent. I then consider some basic consequences of this added dimension.

Keywords

Reasoning, Comprehensibility, Explainable AI, Psycholinguistics

1. Asymmetry

One of the great natural ambitions of AI research is the creation of systems that can collaborate, take instructions from, learn from and even teach human partners in nuanced and intuitive ways that mirror how humans themselves collaborate, instruct, learn and teach. The behavioural symmetry between such (putative) human-like coordination systems, and real human action and decision-making, would allow them to replace human collaborators in many of these domains.

However, when treating coordination as a reasoning problem, an *asymmetry* exists between human and artificial agents, in the resources they each have available to reason with. All other things being equal, AI systems comfortably outpace the human capacity for storing information in working memory (cf. [1]) and for computing optimal solutions based on that information; a capacity evidenced by famous AI victories in chess [2], Go [3, 4] and Starcraft [5]. As a result, human performance is physically bounded in ways AI system performance is not. Artificial systems must adjust to this asymmetry to work with, teach, or substitute human collaborators.

Recent work in explanatory machine learning acknowledges and explores this asymmetry, making the case for a 'cognitive window' of appropriate complexity - neither too high nor too low - where an AI system positively contributes to human performance in an interactive game [6]. AI contributions either overly complex or not complex enough can harm instead of help. This cognitive window is defined by appeal to the reasoning capacity needed to fully represent a problem: e.g. the memory cost of inferring and then implementing a winning move in a game.

In this paper, I use the empirical literature to motivate the further dimension of matching collaborators' *style* of solution in tandem with its complexity. I call this the handshake problem.

HLC 2022: 3rd International Workshop on Human-Like Computing, September 28–30, 2022, Windsor, UK
E.Philalithis@ed.ac.uk (E. Philalithis)

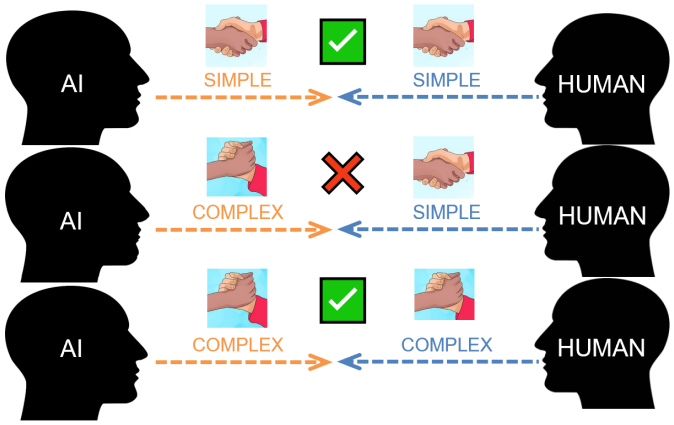

Figure 1: The handshake problem. High/low complexity both succeed when processing styles match.

2. Coordination, Fast and Slow

A basic empirical intuition behind human reasoning - well-encapsulated in the nutshell of 'thinking, fast and slow' [7] - is redundancy for the sake of efficiency. When responding to a wide array of challenges and decisions, human agents reflexively apply different styles and speeds of cognitive processing, to suit the urgency, complexity or familiarity of each application.

Where a task is routine or (importantly) where cognitive resources are strained, such as when judging the outcomes of uncertain processes, human agents rely on fast and cheap heuristics, producing a well-recorded legacy of errors and shortcuts [8, 9, 10]. Conversely, slower processes may support explicit reasoning, action monitoring, or incorporation of domain knowledge [11].

Precise theories of how 'dual-process' reasoning would be implemented in human cognition remain contentious [11, 12]; and the architectural scope of which tasks would fall under which side of such a division remains unclear and subject to debate [13, 14]. This is not a settled claim. For that reason, my present focus is not on implementing dual-process reasoning in artificial systems. Instead, this paper focuses on the methodological challenge of how computational models represent human ability for dual-process theories, relative to a single-process theory; and on the particular implications of this challenge for models of coordination and explanation.

Echoing the 'fast and slow' intuition for individual reasoning, similar contrasts range over coordination and joint problem solving. On the one hand, coordination has been successfully analysed as a 'game' of rigorous reasoning over implications [15], over an explicitly represented perceptual and knowledge domain known as common ground [16] [17]. On the other hand, parts of this representation and domain knowledge have been shown to fall off in relevance as working memory load increases [18] [19] in a trade-off with cheaper processes, like the simple repetition of past structure, which may suffice to establish coordination in many settings [20]. A familiar contrast is implied, between a slow and rigorous method, and a faster more basic one.

The thrust of this lower-level pathway for coordination is that coordination may not often be a problem of reasoning over implications, but a cheaper 'mechanistic' problem of having on-the-fly expectations of the next move or utterance by copying the last one [21]. Coordination

problems that are hard to reason through may be easier for a process based on expectation and repetition. When cheap output is often sufficient explicit reasoning becomes a fallback [22][23].

Taken on its own, the 'fast and slow' intuition already carries significant implications for the cognitive window approach to human-like AI. When different styles of solution are available, the cognitive window where a problem is tractable for human agents becomes relative to the style. An inherently simpler problem, measured e.g. by Kolmogorov complexity [24], could be slower to solve using a richer problem-solving style than a cheap one. A complex problem may turn out easier to process than a problem of slightly smaller inherent complexity, which was allocated to a cheaper style: the window moves along two dimensions. And for coordination in particular this movement can be fatal due to the need for symmetry. I turn to this issue below.

3. The Handshake Problem

I have so far considered the asymmetry in human and AI problem-solving resources, addressed by limiting AI contributions to a cognitive window; and then raised the issue of alternate processing and representation styles as a factor in human reasoning. This analysis conflicts with the definition of the cognitive window from [6], represented as a fixed measure of upper and lower bounds of computational complexity, in two main ways. The most obvious is that the real cognitive window may not be constant, even where the underlying problem remains unchanged. The same complex theory may be included in the cognitive window for a rich processing style, but be fully outside the cognitive window for some cheaper processing style. As a result, the processing style must be modelled, and condition cognitive window boundaries.

The less obvious conflict concerns any possible changes of processing style as a result of an unfavourable cognitive window. That is: a theory is initially chosen, but is replaced with another because of memory or other resource limitations. This wholesale shift is well-documented in human behaviour for collaborative tasks [17] [25] as I discuss more below. In this case the AI system aiming for a cognitive window must reject its initial style of solution to match a human.

Taking both demands together, an AI system may need to model (i) the initial challenge a human collaborator faces for processing a problem, and (ii) the style of processing chosen as a result, and its own associated upper/lower bounds of acceptable complexity. This is the *handshake problem* as illustrated in Figure 1. Like a human handshake, the lower level of the problem uses simpler information (e.g. the trajectory of a moving hand, equivalent to inherent complexity); whereas the higher level involves a demanding modelling objective (the style of handshake to try). One consequence for collaborative problem-solving is that a high complexity contribution by an AI system, using a style that does match their human collaborator, may be more appropriate and understandable than a lower complexity contribution in a different style.

For coordination problems in particular, including most forms of two-way communication, where the objective may not be just to jointly solve a problem but also to *align* [21] the symbols, signals or systems used, a missed handshake may be fatal to the outcome, not just suboptimal. A significant consequence of this constraint is that, where the richer process is not selected by a human collaborator, the best solution by the cheaper process may be the correct choice even where that solution is wrong. That is: where the cheaper process cannot solve the problem optimally, its best (wrong) solution could nonetheless be the one that an AI system must opt for.

4. Shortcuts

A formal solution or specification of the handshake problem is outside the scope of this brief paper. However, as with human reasoning itself, there are always possible shortcuts to consider.

One such shortcut is to decide the processing style in advance of the coordination problem. Where humans must necessarily rely on a priori reasoning, e.g. in virtual bargaining [26], or where past human behaviour in the same task has been extensively sampled, there may be no need for an AI system to model processing styles. If human participants only ever attempt a known task, e.g. 7x7 Noughts and Crosses, with a consistent processing style, the handshake problem disappears. This is an empirical solution to a computational problem, where domain knowledge of processing styles would be obtained before building the AI system, and 'baked in'.

Another shortcut is to lean on the cheaper, routine solution to miscoordination already used by human collaborators in noisier settings: namely rejection and replacement. In a range of coordination contexts, including the maze game [25] [27] and games from the other similar literatures (notably [17]), pairs of human players who fail to establish a 'handshake' simply reject their collaborator's contributions, until or unless a handshake is achieved with a more comprehensible alternative. Where rejected, the initial solution, representation or signal is thrown out and humans try again. An AI system could emulate this by cycling between outputs.

This second shortcut would sidestep the problem of modelling a cognitive window relative to a processing style, in favour of the more heuristic approach of presenting outputs within a range of similarity and complexity, until some output is accepted by the human collaborator.

Human responsiveness to even basic, automated clarification requests in the maze game [28] suggests that a comprehensible reject-and-replace feedback loop could be a relatively simple step to ensure for collaborative AI systems, in settings where coordination may be mission-critical. This reject-and-replace loop could be refined with a better grasp of the links between rejected contributions and their alternatives, to generate more different or more similar alternate output.

For popular coordination problems over conceptual domains - e.g. geometrical images [17], navigating a maze [25], or comparing different maps [29]) - this may be closer at hand. Recent work in explanatory AI uses domain knowledge to generate alternative examples which are conceptually 'near' or 'far' from an initial example of relational concepts [30], including spatial concept domains. A comparable approach could potentially offer intuitive alternatives when navigating spatial domains in coordination tasks, such as geometry, maze layout representations or maps, in line with observed human strategies for repairing unsuccessful coordination [27].

5. Conclusion

Throughout the above I have focused on the single and simple task of motivating the handshake problem as an extension of the cognitive window approach to human-AI coordination, then using findings from the empirical literature to challenge the existing definition and help refine it. This contribution is thus intentionally auxiliary and exploratory - intended as a basis for discussion toward the larger goal for more collaborative, comprehensible, and ultimately human-like AI.

Acknowledgments

The author is grateful to three anonymous reviewers for recommendations and a key example.

References

[1] N. Cowan, What are the differences between long-term, short-term, and working memory?, in: W. S. Sossin, J.-C. Lacaille, V. F. Castellucci, S. Belleville (Eds.), Essence of Memory, volume 169 of *Progress in Brain Research*, Elsevier, 2008, pp. 323–338.

[2] N. Tomašev, U. Paquet, D. Hassabis, V. Kramnik, Reimagining chess with AlphaZero, Communications of the ACM 65 (2022) 60–66.

[3] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al., Mastering the game of Go with deep neural networks and tree search, Nature 529 (2016) 484–489.

[4] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, et al., Mastering the game of Go without human knowledge, Nature 550 (2017) 354–359.

[5] O. Vinyals, I. Babuschkin, W. M. Czarnecki, M. Mathieu, A. Dudzik, J. Chung, D. H. Choi, R. Powell, T. Ewalds, P. Georgiev, et al., Grandmaster level in starcraft ii using multi-agent reinforcement learning, Nature 575 (2019) 350–354.

[6] L. Ai, S. H. Muggleton, C. Hocquette, M. Gromowski, U. Schmid, Beneficial and harmful explanatory machine learning, Machine Learning 110 (2021) 695–721.

[7] D. Kahneman, Thinking, fast and slow, Penguin, 2012.

[8] D. Kahneman, S. P. Slovic, P. Slovic, A. Tversky, Judgment under uncertainty: Heuristics and biases, Cambridge university press, 1982.

[9] J. S. B. Evans, D. E. Over, Rationality and reasoning, Psychology Press, 2013.

[10] N. Chater, The Mind Is Flat: The Remarkable Shallowness of the Improvising Brain, Yale University Press, 2018.

[11] J. S. B. Evans, K. E. Stanovich, Dual-process theories of higher cognition: Advancing the debate, Perspectives on psychological science 8 (2013) 223–241.

[12] M. Osman, An evaluation of dual-process theories of reasoning, Psychonomic bulletin & review 11 (2004) 988–1010.

[13] D. E. Melnikoff, J. A. Bargh, The mythical number two, Trends in cognitive sciences 22 (2018) 280–293.

[14] J. S. B. Evans, Reflections on reflection: the nature and function of type 2 processes in dual-process theories of reasoning, Thinking & Reasoning 25 (2019) 383–415.

[15] H. H. Clark, G. L. Murphy, Audience design in meaning and reference, in: Advances in psychology, volume 9, Elsevier, 1982, pp. 287–299.

[16] H. H. Clark, E. F. Schaefer, Contributing to discourse, Cognitive science 13 (1989) 259–294.

[17] H. H. Clark, D. Wilkes-Gibbs, Referring as a collaborative process, Cognition 22 (1986) 1–39.

[18] W. S. Horton, B. Keysar, When do speakers take into account common ground?, Cognition 59 (1996) 91–117.

[19] B. Keysar, D. J. Barr, J. A. Balin, J. S. Brauner, Taking perspective in conversation: The role of mutual knowledge in comprehension, Psychological Science 11 (2000) 32–38.
[20] M. J. Pickering, S. Garrod, Toward a mechanistic psychology of dialogue, Behavioral and brain sciences 27 (2004) 169–190.
[21] M. J. Pickering, S. Garrod, Alignment as the basis for successful communication, Research on Language and Computation 4 (2006) 203–228.
[22] M. J. Pickering, S. Garrod, An integrated theory of language production and comprehension, Behavioral and brain sciences 36 (2013) 329–347.
[23] M. J. Pickering, S. Garrod, Understanding dialogue: Language use and social interaction, Cambridge University Press, 2021.
[24] A. N. Kolmogorov, On tables of random numbers, Sankhyā: The Indian Journal of Statistics, Series A (1963) 369–376.
[25] S. Garrod, A. Anderson, Saying what you mean in dialogue: A study in conceptual and semantic co-ordination, Cognition 27 (1987) 181–218.
[26] J. Misyak, T. Noguchi, N. Chater, Instantaneous conventions: The emergence of flexible communicative signals, Psychological science 27 (2016) 1550–1561.
[27] P. G. Healey, G. J. Mills, A. Eshghi, C. Howes, Running repairs: Coordinating meaning in dialogue, Topics in cognitive science 10 (2018) 367–388.
[28] P. G. Healey, M. Purver, J. King, J. Ginzburg, G. J. Mills, Experimenting with clarification in dialogue, in: Proceedings of the Annual Meeting of the Cognitive Science Society, volume 25, 2003.
[29] A. H. Anderson, M. Bader, E. G. Bard, E. Boyle, G. Doherty, S. Garrod, S. Isard, J. Kowtko, J. McAllister, J. Miller, et al., The HCRC map task corpus, Language and speech 34 (1991) 351–366.
[30] J. Rabold, M. Siebers, U. Schmid, Generating contrastive explanations for inductive logic programming based on a near miss approach, Machine Learning 111 (2022) 1799–1820.