Edinburgh Research Explorer

Climbing the ladder

Climbing the Ladder: How Agents Reach Counterfactual Thinking

Caterina Moruzzi[a]

Department of Philosophy, Universität Konstanz, 78457, Konstanz, Germany

Keywords: Counterfactuality, Agency, Decision-making, Causal Reasoning, Robustness.

Abstract: We increasingly rely on automated decision-making systems to search for information and make everyday choices. While concerns regarding bias and fairness in machine learning algorithms have high resonance, less addressed is the equally important question of to what extent we are handing our own role of agents over to artificial information-retrieval systems. This paper aims at drawing attention to this issue by considering what agency in decision-making processes amounts to. The main argument that will be proposed is that a system needs to be capable of reasoning in counterfactual terms in order for it to be attributed agency. To reach this step, automated system necessarily need to develop a stable and modular model of their environment.

1 INTRODUCTION

Research on agency and causal efficacy has a long history in the philosophical literature, and recently there has been a resurgence of interest in this topic (List and Pettit, 2011; Müller, 2008; Nyholm, 2018; Ried et al., 2019; Sarkia, 2021). Like with many other 'suitcase words', there is no general agreement on the definition of the notion of agency. While the generally accepted definition of agent in artificial intelligence (AI) research is "anything that can be viewed as perceiving its environment through sensors and acting upon that environment through actuators" (Russell and Norvig, 2011), others are unhappy with the broadness of this definition and look for a narrower interpretation of the term, for example as an entity that can learn and be trained (Müller and Briegel, 2018).

In order to be able to act upon the environment and interact with it, it seems that an agent needs to be able to reason in causal terms, namely to understand which of the variables present in its surroundings is responsible for the change occurred (Tomasello, 2014). Still, it may be argued that causal reasoning is not enough if the agent wants to reach a level of abstraction that allows it to generalise to unknown scenarios. To reach this aim, what the agent needs to develop is the capacity of reasoning in counterfactual terms, to project itself into unknown situations it has never experienced before. Through counterfactual reasoning, the agent can strengthen its predictive abilities and understand not only what happens but also 'why' something hap-

pens. The different levels of competence possessed by a system, from correlation between data to counterfactual thinking, are described by Judea Pearl and Dana Mackenzie through the metaphor of the Ladder of Causation (Pearl and Mackenzie, 2018). They argue that data without an understanding of the causal links occurring between them are not enough: to be able of generalisation and abstraction to out-of-distribution scenarios, agents need to frame a model of the world they live in. In what follows, an agent will be understood as positioning itself on the top rung of the Ladder of Causation, that of counterfactuality.

In this paper, the relevance of organising information through frames in order to reach the third rung on this Ladder will be examined, describing the process of building robustness into a human decision-making system and considering how recent machine learning (ML) models try to implement this capacity in automated systems (Bertsimas and Thiele, 2006; Hansen and Sargent, 2011). In conclusion, it will be claimed that models that organise information through sparse and modular frames show promising results in developing better generalisation abilities.

2 AGENCY IN DECISION-MAKING

Automated decision-making systems increasingly support humans in looking for information. AI information access systems, such as recommendation systems, guide us in searching for the information

---

[a] https://orcid.org/0000-0002-9728-3873

that we absorb everyday, whether it be looking for a product on an e-commerce platform, watching our favourite series, or reading the news. They filter the information that we access and are responsible for sorting and shaping it, acting as gateways and conditioning public choices and opinions (Nielsen, 2016; Shoemaker et al., 2009).

Most of the times, we are unaware of these invisible partners that assist us in our quest for information and we have the impression of being totally in charge of the choices we make online. Other times, especially in systems with human-like interfaces, such as voice assistants, we interact with them as social agents, attributing responsibilities to them and, sometimes, placing trust on their decisions (Doyle et al., 2019; Langer et al., 2021; Pitardi and Marriott, 2021).

Given the pervasiveness of the power asserted by these information-framing systems, it is urgent to consider users' practices of attribution of agency to them and the impact that their assertion of agency over our access to information has on the trust that we place in the systems themselves (Shin, 2020). The first step we need to take in order to progress in this direction is to understand which are the essential features that a system needs to possess in order to be deemed an agent.

With this aim, in the next sections human decision-making processes are examined in the light of Pearl and Mackenzie's Ladder of Causation, in order to extract the key ingredients that allow humans to reach counterfactual reasoning abilities and, thus, be deemed proper agents.

2.1 Ladder of Causation

To understand how a system can reach the counterfactual rung, it is first necessary to describe what the metaphor of the Ladder of Causation amounts to (Pearl and Mackenzie, 2018).

Rung 1 of the Ladder is 'Correlation'. A system operating on the first rung is a mere observer of what happens in the world. The question linked to this stage is: "What is the probability that y happens, given x?" or, in symbols, $P(y|x)$.

Rung 2 is 'Intervention'. In order to climb to this level, the system needs to deliberately interact with the environment and alter it. The question is: "What is the probability that y happens if I *do* x?", $P(y|do(x))$.

Rung 3 is 'Counterfactuality'. Agents that reach this step are able to imagine counterfactual scenarios and to adapt their actions accordingly. The question the agent asks is: "What is the probability that y' would occur had x' occurred, given that I actually observed x and y?", $P(y'_{x'}|x,y)$.

While humans are good at forming causal frames of the available information, according to Pearl and Mackenzie current state of the art ML models do not progress beyond the first rung of the Ladder: that of observing the environment and finding statistical correlation between available data. Progress in AI can come only through the development of systems that are able to reason counterfactually and abstract to unknown data. State-of-the-art automated systems can produce counterfactuals, but without the capacity of selecting the relevant ones among them (de Véricourt et al., 2021). Indeed, while ML systems surpass human capacities in processing large amount of data, it is still a challenge for these systems to frame and filter relevant information and to extrapolate to unknown scenarios, a necessary ability to climb to the highest rung on the Ladder.

Analysing the decision-making process that human agents go through to climb the Ladder is helpful to understand which are the features that automated systems need to develop in order to reach agentive capacities. In addition, this description can help addressing further questions, such as: Can the counterfactuality rung be reached through the acquisition of causal reasoning skills, developed in Rung 2 of the Ladder of Causation, or are other, qualitatively different features, needed for an agent to perform counterfactual thinking? Supposing that an artificial system has computational power orders of magnitude larger than what any system can at present have and, as a consequence, it can test every possible scenario, could it reach Rung 3 of the Ladder of Causation, or are some priors necessary?

For the sake of the present analysis, mechanisms that allow to proceed from rung 1 to rung 2 on the Ladder will not be considered, while priority will be given to addressing the question of to what extent the framing of information helps agents to make the final step toward rung 3, thus reaching counterfactual reasoning skills.

3 CLIMBING THE LADDER

Suppose that an agent, Bob, wants to lose weight. In order to decide what choices he needs to make in order to achieve this aim, Bob goes through a decision-making process. The steps that he will (presumably) take follow the three rungs of the Ladder of Causation (see Figure 1).

A system operating on the first rung of the Ladder, 'Association', observes what happens in the available data, or 'Knowledge'. At the beginning of the decision-making process the agent just has data and

the way it starts to frame it is essential to the final output of the process. To achieve the aim of losing weight, Bob may start by observing with which probability the variable 'Playing basketball' is associated to the variable 'Losing weight'. The initial, approximate, frame that he creates helps Bob understand and organise the data (F_a).
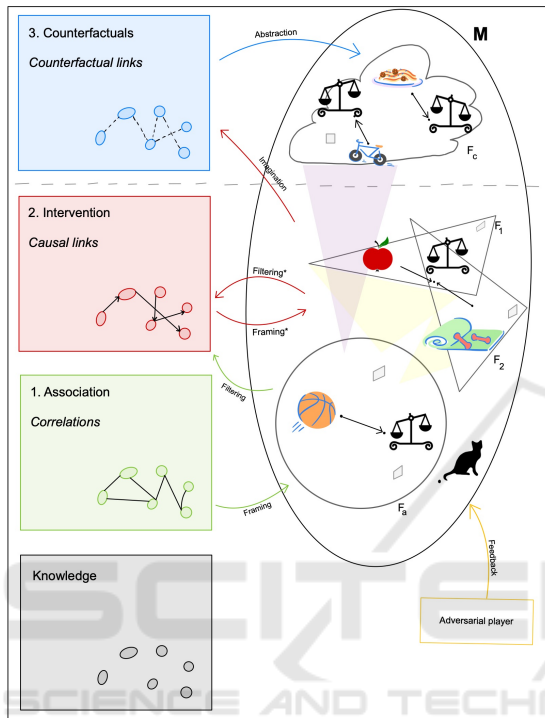

Figure 1: Decision-making process.

In order to climb to the 'Intervention' rung, the agent needs to deliberately interact with the environment and alter it. In our example, Bob may join a basketball team to confirm the connection between the variables 'Playing basketball' and 'Losing weight', observed in the previous step. His question is: "What is the probability that by playing basketball I will lose weight?" If the intervention supports the observed data, the agent can draw causal links between the variables 'Playing basketball' and 'Losing weight'. To make his frame more robust, Bob chooses among the alterations to the environment the ones that produce the desired outcome. He will use his frame to filter out variables that are not related to the outcome 'Losing weight', for example (stroking a) 'Cat'. Once Bob trusts the effectiveness of his frame, he starts interpreting data according to it. In turn, the frame influences Bob in drawing causal links between elements in future experiences (Filtering*).

The process of optimising the frame by filtering out irrelevant information is crucial to make the frames more robust and more easily adaptable to other scenarios. Another element that contributes to the robustness of the frame is the interplay between the agent and an Adversarial Player (Hansen and Sargent, 2011), which can be understood as a max-min decision rule. The decision maker maximises and assumes that the Adversarial Player chooses a probability to minimise her expected utility (Goodfellow et al., 2014). Suppose Bob cannot play because the basketball court is not available. The Adversarial Player may make Bob try different things to achieve the aim of losing weight, for example drinking lactose-free milk, taking vitamin pills, practicing yoga, and so on. Through the interplay with the Adversarial Player and the feedback received, Bob may find out that adopting a vegetarian diet works and add it to his frame (F_1). Or he may discover that exercising indoors also works for losing weight and add it to a further frame (F_2).

Frames are cognitive shortcuts used by agents to move through the uncertainty of their surroundings by identifying the variables that are responsible for change (Kahneman, 2011). Using frames to interact with the environment helps the agent react to changing conditions. By capturing the most important aspects of the world and filtering out the others, frames help agents to learn from single experiences and come up with general rules that can be applied to other situations, thus progressing toward the counterfactual rung of the Ladder of Causation.

3.1 Reaching Counterfactuality

Counterfactual thinking amounts to searching for an explanation to what happened by asking what should have happened in the past that would have changed the output (Pearl and Mackenzie, 2018). Reasoning in counterfactual terms is considered a requisite for a system to be attributed responsibility and it is a requirement that agents need to satisfy in order to provide satisfactory, interpretable explanations (Lipton, 1990; Miller, 2019; Molnar, 2020; Wachter et al., 2017).

Counterfactual reasoning occupies the third and highest rung on the Ladder of Causation. Human agents are able to imagine counterfactual scenarios ('Imagination') and to plan how their frames can be projected onto those ('Abstraction', see Figure 1). The challenge in adapting the frames the agent built in lower rungs of the Ladder to a counterfactual scenario comes from the fact that the agent cannot adjust its actions on the basis of feedback. In our example, Bob cannot go back in time and undo the action of playing basketball to know whether he would have lost weight if he did not play. He can only imagine it, drawing

inferences on how strong the links between the variables 'Playing basketball' and 'Losing weight' are by considering his previous Knowledge.

In order to be able to adapt the frame to a counterfactual scenario, Bob needs to identify fundamental causal links in available frames and understand the relation that they have with variables that he has not experienced, yet. For example, Bob may ask "What is the probability that I would lose weight if I cycle?". He could, then, draw a link between playing basketball and cycling, identify that the two activities have something in common and, through abstraction, draw a causal link between 'Cycling' and 'Losing weight'. Through counterfactual thinking Bob can also reflect on the original cause of his weight gain, for example by asking "What is the probability that I would not have gained weight, had I not eaten so much during my holiday in Italy?". If the probability is low, then he can identify 'Italian food' as the cause of his weight gain.

In order to identify the variables responsible for change, the agent can start by building a causal diagram. Figure 2 represents the cause-effect relation between the variables in our example through a diagram. This kind of causal diagram has been theorised by Pearl as a way of mapping the data available to the (alleged) agent, in order to identify cause-effect links and make better predictions (Pearl and Mackenzie, 2018). The nodes in the diagram stand for the variables and the arrows for presumed causal relations.[1] An arrow connects the variable 'Playing basketball' to the variable 'Losing weight', as the agent has concluded that playing basketball caused the weight loss. Building a causal diagram where the agent can identify the variables that are responsible for change is compelling to answer counterfactual questions of the kind "What would have happened, had I acted differently?", thus allowing the agent to be discounted from the burdensome need to experience all the possible scenarios.

Thinking about the past and about what would have changed if it acted differently allows the agent to understand which modules of the frame are responsible for change. In so doing, the agent can form an hyper-model (M) within which all the frames, real and counterfactual, can be included. Through a higher level of abstraction, Bob could identify what connects all the activities responsible for losing weight

---

[1]Causation is defined by Pearl as follows:"a variable X is a cause of Y if Y 'listens' to X and determines its value in response to what it hears." (Pearl and Mackenzie, 2018) The connection between the variables 'Gaining weight' and 'Playing basketball', 'Cycling', and 'Exercising indoors' is represented here through a dotted line as it is not a proper causal relation.
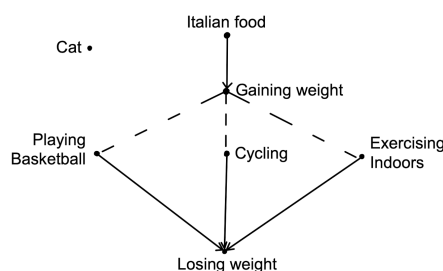

Figure 2: Causal diagram.

and make his model even more robust and capable of adapting to different contexts. The relations that link together variables within this hyper-model remain invariant and, thanks to its robustness, M can be used to understand and deal with new data and produce new estimates.

3.2 Sparse and Compositional Models

The description of how humans build robustness into their decision-making systems made above supports the idea that, in order to achieve abstraction, agents need to build frames which select and organise relevant information and to include these frames into an hyper-model which they can use to adapt their actions to both known and unknown scenarios. The exploration of this abstraction mechanism of human cognition (arguably valid also for some animals) can help practitioners understand how to program artificial systems that reach the same level of generalisation and agency.

Indeed, recent ML systems try to approximate the aim of dealing with out-of-distribution data and adapting to unknown scenarios by developing two features which are essential to the construction of a comprehensive hyper-model: compositionality and sparseness. In order to effectively deal with new situations in the world, agents conveniently build models made of smaller parts that can be recombined. This compositional ability is useful to explain new data observed in scenarios previously unknown (Kahneman, 2011), as it allows agents to identify causal connections that are valid throughout different frames. The adaptability of the model is enhanced by the ability of the agent to identify the variables that are responsible for change, providing explanations of what it observes and adapting to changes without the need of experiencing all possible scenarios.

A good model must be stable and robust to changes. The creation of a sparse and modular model has been explored as an option for building robustness in automated systems, for example through the representation of variables in sparse factor graphs (Ben-

gio, 2017; Ke et al., 2019). Another example are Recurrent Independent Mechanisms, a meta-learning approach which decomposes knowledge in the training set in modules that can be re-used across tasks (Goyal et al., 2019; Madan et al., 2021). The selection of which modules to use for different tasks is performed by an attention mechanism, while Reinforcement Learning mechanisms are responsible for the process of adaptation to new parameters.

A modular capacity is better achieved in systems that combine the data-processing capabilities of ML models with the capacity of abstraction and logical reasoning of symbolic AI methods. According to supporters of this new paradigm, referred to as the Third Wave or hybrid AI, statistical models are not enough to achieve generalisation, we need to teach systems to handle also logical and symbolic reasoning. This hybrid approach of symbolic and sub-symbolic methods allows to hold the advantages of both strategies, get rid of their respective weaknesses and, at the same time, program models that fare much better in generalisation and abstraction (Anthony et al., 2017; Bengio et al., 2019; Bonnefon and Rahwan, 2020; Booch et al., 2020; Garcez and Lamb, 2020; Hill et al., 2020; Ke et al., 2019; Mao et al., 2019; Moruzzi, 2020). The benefit of these hybrid models consists in their capacity of combining the computational power of Deep Learning with symbolic and logical reasoning to not only be able to process large amounts of data but also identify which elements within those data stay stable.

4 CONCLUSIONS

The ongoing research presented in this paper contributes to an exhaustive and accurate analysis of the notion of agency, a useful tool for the investigation of how to build reliable and flexible decision-making systems. The study of how the progression toward generalisation to unknown scenarios happens and why it is necessary to develop agency helps creating a deeper theoretical understanding of the characteristics of a robust decision-making process, contributing to address a fundamental issue within AI: whether and how systems achieve causal agency.

The analysis of the parallel between decision-making in humans and machines that has been here presented not only contributes to debates on human and artificial agency but can also provide relevant insights to research in neuromorphic engineering (Indiveri and Sandamirskaya, 2019). Indeed, one of the challenges in the development of embodied devices that interact with the environment is the design of solutions through which to generate context-dependent behaviour, adaptable to changing and unknown conditions.

This paper has identified the ability of sorting and organising information through frames as a crucial requisite for agents to build a robust model of their environment, a model which allows them to adapt and modify their choices according to the context. The analysis of the development of agency in decision-making systems is a preliminary, essential step to study whether the emulation of biological processes is a viable path for achieving power-efficient solutions with the aim to build robust and flexible artificial agents.

REFERENCES

Anthony, T., Tian, Z., and Barber, D. (2017). Thinking fast and slow with deep learning and tree search. *arXiv preprint*, arXiv:1705.08439.

Bengio, Y. (2017). The consciousness prior. *arXiv preprint*, arXiv:1709.08568.

Bengio, Y., Deleu, T., Rahaman, N., Ke, R., Lachapelle, S., Bilaniuk, O., Goyal, A., and Pal, C. (2019). A meta-transfer objective for learning to disentangle causal mechanisms. *arXiv preprint*, arXiv:1901.10912.

Bertsimas, D. and Thiele, A. (2006). Robust and data-driven optimization: Modern decision making under uncertainty. *INFORMS TutORials in Operations Research*, pages 95–122.

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., Murugesan, K., Mattei, N., Rossi, F., et al. (2020). Thinking fast and slow in AI. *arXiv preprint*, arXiv:2010.06002.

de Véricourt, F., Cukier, K., and Mayer-Schönberger, V. (2021). *Framers: Human Advantage in an Age of Technology and Turmoil*. Penguin Books Ltd, New York.

Doyle, P. R., Edwards, J., Dumbleton, O., Clark, L., and Cowan, B. R. (2019). Mapping perceptions of humanness in speech-based intelligent personal assistant interaction. *arXiv eprint*, arXiv:1907.11585.

Garcez, A. d. and Lamb, L. C. (2020). Neurosymbolic AI: The 3rd wave. *arXiv preprint*, arXiv:2012.05876.

Goodfellow, I. J., Shlens, J., and Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Goyal, A., Lamb, A., Hoffmann, J., Sodhani, S., Levine, S., Bengio, Y., and Schölkopf, B. (2019). Recurrent independent mechanisms. *arXiv preprint*, arXiv:1909.10893.

Hansen, L. P. and Sargent, T. J. (2011). *Robustness*. Princeton University Press, Princeton, NJ.

Hill, F., Tieleman, O., von Glehn, T., Wong, N., Merzic, H., and Clark, S. (2020). Grounded language learning fast and slow. *arXiv preprint*, arXiv:2009.01719.

Indiveri, G. and Sandamirskaya, Y. (2019). The importance of space and time for signal processing in neuromorphic agents: the challenge of developing low-power, autonomous agents that interact with the environment. *IEEE Signal Processing Magazine*, 36(6):16–28.

Kahneman, D. (2011). *Thinking, Fast and Slow*. Farrar, Straus and Giroux, New York.

Ke, N. R., Bilaniuk, O., Goyal, A., Bauer, S., Larochelle, H., Schölkopf, B., Mozer, M. C., Pal, C., and Bengio, Y. (2019). Learning neural causal models from unknown interventions. *arXiv preprint*, arXiv:1910.01075.

Langer, M., Hunsicker, T., Feldkamp, T., König, C. J., and Grgić-Hlača, N. (2021). "Look! It's a computer program! It's an algorithm! It's AI!": Does terminology affect human perceptions and evaluations of intelligent systems? *arXiv eprint*, arXiv:2108.11486.

Lipton, P. (1990). Contrastive explanation. *Royal Institute of Philosophy Supplements*, 27:247–266.

List, C. and Pettit, P. (2011). *Group Agency: The Possibility, Design, and Status of Corporate Agents*. Oxford University Press, Oxford.

Madan, K., Ke, N. R., Goyal, A., Schölkopf, B., and Bengio, Y. (2021). Fast and slow learning of recurrent independent mechanisms. *arXiv preprint*, arXiv:2105.08710.

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. *arXiv preprint*, arXiv:1904.12584.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38.

Müller, T. and Briegel, H. J. (2018). A stochastic process model for free agency under indeterminism. *dialectica*, 72(2):219–252.

Molnar, C. (2020). *Interpretable Machine Learning*. Lulu.com.

Moruzzi, C. (2020). Artificial creativity and general intelligence. *Journal of Science and Technology of the Arts*, 12(3):84–99.

Müller, T. (2008). Living up to one's commitments: Agency, strategies and trust. *Journal of Applied Logic*, 6(2):251–266.

Nielsen, R. K. (2016). News media, search engines and social networking sites as varieties of online gatekeepers. In Peters, C. and Broersma, M., editors, *Rethinking Journalism Again*, pages 93–108. Routledge, New York.

Nyholm, S. (2018). Attributing agency to automated systems: Reflections on human–robot collaborations and responsibility-loci. *Science and engineering ethics*, 24(4):1201–1219.

Pearl, J. and Mackenzie, D. (2018). *The Book of Why: The New Science of Cause and Effect*. Hachette UK.

Pitardi, V. and Marriott, H. R. (2021). Alexa, she's not human but... Unveiling the drivers of consumers' trust in voice-based artificial intelligence. *Psychology & Marketing*, 38(4):626–642.

Ried, K., Müller, T., and Briegel, H. J. (2019). Modelling collective motion based on the principle of agency: General framework and the case of marching locusts. *PloS one*, 14(2):e0212044.

Russell, S. J. and Norvig, P. (2011). *Artificial Intelligence: A Modern Approach. Third Edition*. Prentice Hall, Upper Saddle River, NJ.

Sarkia, M. (2021). Modeling intentional agency: a neo-gricean framework. *Synthese*, pages 1–28.

Shin, D. (2020). User perceptions of algorithmic decisions in the personalized AI system: Perceptual evaluation of fairness, accountability, transparency, and explainability. *Journal of Broadcasting & Electronic Media*, 64(4):541–565.

Shoemaker, P. J., Vos, T. P., and Reese, S. D. (2009). Journalists as gatekeepers. In Wahl-Jorgensen, K. and Hanitzsch, T., editors, *The Handbook of Journalism Studies*, pages 73–87. Routledge, New York.

Tomasello, M. (2014). *A Natural History of Human Thinking*. Harvard University Press, Cambridge, MA.

Wachter, S., Mittelstadt, B., and Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 31:842–887.