Impact of Artificial Intelligence Systems on Cognitive Liberty

A Research Paper submitted to the Department of Engineering and Society

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

Emma Graham

Spring 2022

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Advisor

Rider W. Foley, Department of Engineering and Society

Introduction

Proclaimed at the United Nations General Assembly in Paris in 1948, the United Nations Universal Declaration of Human Rights enumerates thirty articles that represent a consensus of inherent, universally protected human rights (United Nations, 1948). The Declaration serves as a pinnacle for the dictation of human rights and has served as a foundation for a multitude of human rights treaties. Article 18 of the Universal Declaration states that "everyone has the right to freedom of thought, conscience and religion; this right includes freedom to change his religion or belief, and freedom, either alone or in community with others and in public or private, to manifest his religion or belief in teaching, practice, worship and observance" (United Nations, 1948, 1).

Although dictated as a universal freedom, the freedom of thought remains enigmatic in both distinguishing the levels of freedom and in defining the process itself. The thought process, or higher cognitive process, is the inspiration for cognitive computing and artificial intelligence while, cyclically, highly impacted by artificial intelligence (Sayantini, 2019). Humans possess the ability to approximate the thought process of our peers, which in cognitive science is known as the theory of mind (ToM) (McCarthy-Jones, 2019; Marraffa, 1995). Surpassing theory of mind of an opponent is equivalent to the obtaining the knowledge of the future moves of the opponent. If the futures choices of the opponent are predicted before the opponent makes their move, then not only is the decisive advantage procured but the opportunity to manipulate the opponent is created. McCarthy-Jones' 2019 article, "The Autonomous Mind: The Right to Freedom of Thought in the Twenty-First Century", finds that "the development of brain- and behavior-reading techniques threaten to disturb the delicate balance between our evolved ability to know other's thoughts yet to also shield our inner world" (McCarthy-Jones, 2019, 1). This

reading innovation is already occurring as the advancement of the predictive algorithms in artificial intelligence enables the intelligent system to infer thoughts at a "supra-natural ability, steamrolling through (our) evolved defenses" against deception (McCarthy-Jones, 2019). A major concern is the out-pacing of our theory of mind's ability by that of artificial intelligence.

Experts agree that "rampant innovation (does) affects our rights and duties as citizens" and artificial intelligence has already affected our thought processes, affecting our rights as citizens of humanity (Andrews, 2006). To combat the human right violation of the freedom of thought that innovations in artificial intelligence are enabling, our knowledge of our decision-making processes and the impact of artificial decision-making processes' impact on our cognitive liberty is essential. This paper analyzes this issue through the technical exploration of the adjoining discussion presented in "A Working Theory of a Learned Model in a Partially Observable Environment for Cognitive Decision-Making", and a socio-technical investigation of these technological innovations' impact on the freedom of the individual (Graham, 2022). The technical component is a theorized model to computationally capture the cognitive processes, while the societal component analyzes the impact of artificial intelligence systems on cognitive liberty.

Context

Understanding more about our cognitive process enables us to obtain a better understanding the effects on our decision-making, including the impact of artificial intelligence on our cognition. Breaking down our own cognitive decision-making process will allow us to better understand our own decisions and equip inanimate objects with similar decision-making

processes. Artificial neural networks are “paving the way for new discoveries and a closer integration between technology and the brain” (Lane, 2021). Artificial intelligence has made waves in quantifying the decision-making process. As advancements in artificial intelligence are made, they cyclically advance the understanding of and impact on cognition. About five years ago, Matthew Hutson, a freelance writer for Science, covered an astonishing technological advancement in artificial intelligence capabilities in his 2017 article “Artificial intelligence is learning to read your mind - and display what it sees” (Science, 2022.; Hutson, 2017). After Hutson’s 2017 subtitle “A computer guesses what people are watching based on brain activity”, he reports that

> *Artificial intelligence has taken us one baby step closer to the mind-reading machines of science fiction. Researchers have developed "deep learning" algorithms—roughly modeled on the human brain—to decipher, you guessed it, the human brain. First, they built a model of how the brain encodes information.*
>
> (Hutson, 2017)

Innovations in artificial intelligence are crossing over into the intuitive mind-reading ability, which psychologists refer to as the Theory of Mind. More precisely, Theory of Mind (ToM) is a cognitive mechanism that combines the intellectual abilities which allow cognitive beings to understand and interpret the goals, beliefs, plans, information, intentions, and desires of others (Korkmaz, 2011, 101-102). In essence, this psychological theory describes our inferences about the psychological state of others (Korkmaz, 2011, 102). Drawing a parallel to Hutson’s report on the empirical research of artificial intelligence’s mind-reading display, inferring the psychological state of a person is exactly what Hutson reported artificial intelligence achieving back in 2017. Machines have been able to predict cognitive function and, as explored in the

technical analysis of quantifying cognition in the technical section, are utilizing models of the cognitive process in the deep learning subset of machine learning. The innovations of processing speeds and computer hardware are allowing the rate of artificial "mindreading" and decision-making inference to approach and outpace human cognition rates. The accelerated ability to infer our thoughts gives artificial intelligence a decisive advantage opening the door for manipulation, by the administrator, of the end receiver of the intelligent system.

Many dystopian implications that stem from the implementations of this technological ability are due to nefarious actors. The thought manipulation this innovation permits directly enables the violation of the universal human right to the freedom of thought and opinion (United Nations, 1948). Propaganda has already been shown to be a tool successful at changing one's opinions but the past methods of propaganda pale in the face of artificial intelligence technology's ability of manipulation (Asmolov, 2019). The interaction of this technology with individuals and society would become increasingly one-sided as the inanimate intelligence systems influence their users.

Socio-technical Framework

The framework that will be used in analyzing the impact of artificial intelligence systems on cognitive liberty is the responsible innovation framework. Rene von Schomberg, Commissioner in the European Union, defines Responsible Research and Innovation as "a transparent, interactive process by which societal actors and innovators become mutually responsive to each other with a view to the (ethical) acceptability, sustainability and societal desirability of the innovation process and its marketable products (in order to allow a proper embedding of scientific and technological advances in our society)" (von Schomberg, 2012, 9). The Responsible Innovation framework consists of a prospective model of responsibility of the

technology and using this model, four dimensions are explored (Stilgoe et al., 2013). The dimensions are anticipation, reflexivity, inclusion, and responsiveness (Stilgoe et al., 2013). The four dimensions are interpreted as displayed in Table 1.

Table 1

Four dimensions of responsible innovation.

Anticipation	Anticipation necessitates "systematic thinking aimed at increasing resilience, while revealing new opportunities for innovation and the shaping of agendas for socially-robust risk research" (Stilgoe et al., 2013) and prompts the 'what-if' questions that new scientific challenges have precipitated (Ravetz, 1997, 533).
Reflexivity	The reflexivity dimension will be investigated through second-order reflectivity. Second-order reflexivity considers the presumptions and technological reflection, first-order reflection, techniques by "theorizing about theories" in order to uncover the underlying foundations (Perez, 2011, 6-7).
Inclusion	Inclusion involving issues of science and innovation is actively moving to include new members beyond that of direct stakeholders to diversity inputs in the pursuit of legitimacy (Stilgoe et al., 2013).
Responsiveness	Responsiveness involves acknowledging and reacting to new knowledge and "for responsible innovation to be responsive, it must

	be situated in a political economy of science governance that considers both products and purposes" (Stilgoe et al., 2013).

Note Stilgoe, J., Owen, R., & Macnaghten, P. (2013, November). Developing a framework for responsible innovation. *Research Policy*

The integration of the dimensional analysis then examines the transition between deliberation, reflection, and anticipation, and that of agency and action. This transitional introspection explicitly requires the connection with societal values and governance practices (Stilgoe et al., 2013). Responsible innovation questions can be inquired through the process methods in Table 2 below,

Table 2

The methods and factors associated with the four dimensions of responsible innovation.

Dimension	**Indicative techniques and approaches**	**Factors affecting implementation**
Anticipation	Foresight Technological assessment Scenarios Vision assessment Socio-literary techniques	Engaging with existing imaginaries Participation rather than prediction Plausibility Investment in scenario-building Scientific autonomy and reluctance to anticipate
Reflexivity	Multidisciplinary collaboration and training Embedded social scientists and ethicists in laboratories Ethical technology assessment Codes of conduct Moratoriums	Rethinking moral division of labor Enlarging or redefining role responsibilities Reflexive capacity among scientists and within institutions Connections made between research practice and governance
Inclusion	Consensus conferences Citizens' juries and panels Focus Groups Science shops Deliberative mapping	Questionable legitimacy of deliberative exercises **Need for clarity about, purposes of and motivation for dialogue** **Ability to consider power imbalances** Deliberation of framing assumptions

	Deliberative polling Lay membership of expert bodies **User-centered design** Open innovation	**Ability to interrogate the social and ethical stakes associated with new science and technology** Quality of dialogue as a learning exercise
Responsiveness	Constitution of grand challenges and thematic research programmes Regulation Standards Open access and other mechanisms of transparency Niche management Value-sensitive design Moratoriums Stage-gates Alternative intellectual property regimes	Strategic policies and technology 'roadmaps' Science-policy culture Institutional structure Prevailing policy discourses Institutional cultures Institutional leadership Openness and transparency Intellectual property regimes Technological standards

Note Stilgoe, J., Owen, R., & Macnaghten, P. (2013, November). Developing a framework for responsible innovation. *Research Policy*

As depicted in Table 2, there are many methods of exploration through the four dimensions. The dimension, method, and factors influencing implementation in bold in Table 2 are the prominent Responsible Innovation components used in the exploration of this paper.

Research Question, User-centered Design, and Analytical Information Acquisition

The current innovations in artificial intelligence and infringements on the ambiguous concept of freedom of thought lead to the critical research question: *What is the impact of artificial intelligence systems on cognitive liberty?*

The analytical information is acquired through case analysis. The Responsible Innovation framework provides the method of evaluation and analysis of the cases. The risks cognitive liberty imposes is anticipatory in nature and consequently the inclusion dimension will be the primary dimension explored. As seen in the Indicated techniques and approaches column in

Table 2, the methods used will be the user-centered design technique. Table 2 provides influential factors under the Factors affecting implementation column, providing the following factors: need for clarity about, purposes of and motivation for dialogue ability to consider power imbalances, and ability to interrogate the social and ethical stakes associated with the new science and technology. The influential factors will be considered as cases are analyzed. The selection criteria for cases are in the following categories: thought reform, external influence, and stratagem. The three cases center on DeepMind's artificial intelligence systems and explore the impact of its innovation on its human user. The evaluated research questions will be connected to the societal values and governance practices of cognitive liberty.

Results

Summary

Case I looks at the human-human interaction regarding thought and its control. It also expresses the influences of others honing (or *nudging*) a person into making the desired decisions. Case II analyzes the human-machine interaction of the impact on the human user in engagement with artificial intelligence models. Models created to optimize human interaction that have a human-like design have been shown to *nudge* their human interactant to reveal sensitive, personal information. Case III evaluates the strategic abilities of reinforcement learning artificial intelligence agents and notes the ability to outmaneuver and outperform humans in increasingly complex and intricately intellectual games.

CASE I - Human Thought Reform

Thought Reform is the phrase D. Robert Lifton used to describe the psychological brainwashing and mind control tactics he witnessed first-hand in his research expedition in the

mid-twentieth century (Lifton, 1989). Dr. Robert Lifton found that the Chinese could "completely reform a person's thoughts, practically changing who the person was so drastically that they became unrecognizable to who they used to be" and then he decrypted the tactics employed (Jason, 2019). Essentially, Lifton evaluated an adversarial application of the Theory of Mind (ToM). Then in 1989, Lifton published his findings in his book *Thought Reform and The Psychology of Totalism: the study of brainwashing in China* outlines what is known in psychiatry to be The 8 Criteria for Thought Reform (Jason, 2019).

Given that our own cognitive, general intelligence has employed tactics to reform another's process of comprehension and thinking, then with the rapidly advancing technological intelligence already taking rapid steps toward being able to do so, an artificial intelligence (AI) has the advantage of a pre-made, recommended *eight-step* procedure. The expansive, intense, and fully immersive eight steps outlined by Dr. Lifton is not the only method of impact a person can have on another's thought process. Influence, as prominent in advertising and influencers, can be subtle yet highly effective. Something as small as a *nudge* may push individuals to make one choice rather than another.

Nudging refers to any aspect of the choice architecture that alters a person's behavior predictably, without forbidding any options or significantly changing their economic incentives (Thaler & Sunstein, 2008). In a 2010 refinement of this definition, and with perhaps a more ominous connotation, Hausman and Welch define nudging is the "use of flaws in human judgment and choice to influence people's behavior" (Hausman & Welch, 2010, 124). Before warning of the highly influencing effect of nudging on individuals and populations professors of philosophy, Hausman and Welch, sum up nudges as the ways of influencing which choice is made, without limiting the choices available, by making the intended choice more appealing than

the alternative options (Hausman & Welch, 2010, 126). This influence over another's decisions is accessible by ToM and, according to the professors, enables by the "flaws in individual decision-making" (Hausman & Welch, 2010, 126). After defining, Hausman and Welch warn of risks to an individual's autonomy saying:

> *"Their freedom, in the sense of what alternatives can be chosen, is virtually unaffected, but when this "pushing" does not take the form of rational persuasion, their autonomy—the extent to which they have control over their own evaluations and deliberation—is diminished. Their actions reflect the tactics of the choice architect rather than exclusively their own evaluation of alternatives. We shall call the use of flaws in human decision-making to get individuals to choose one alternative rather than another "shaping" their choices. We intend "shaping" to exclude rational persuasion. "Manipulation" would be a more natural label,"*
> (Hausman & Welch, 2010, 128).

The user-centered design of the utilization of nudging in an interaction, with the user being the end receiver the nudging, presents the potential for manipulation of the receiver. This type of interaction brings forth the *ability to consider power imbalances* factor affecting the user-centered design method for evaluating the impact of the receiving individual in the Responsible Innovation framework. The design of this interaction not only considers the imbalance of power, derived from the physiological ToM insights, but uses that power imbalance in a purposeful, not necessarily devious, manner. In following two cases, this person on the receiving end will be the human user, who is interacting with a non-human intelligence.

CASE II - External Influence

"Convince me" is heard when a reasoned argument or intellectual discussion is desired. The external reasoning of others is part of how humans, as a species, grow and improve. After all, "two heads are better than one" and historically our biggest goals are achieved by combined efforts.

In 2021, DeepMind published a document warning of the "Ethical and Social Risks of Harm from Language Models". The report outlines the risks to the human user from interacting with large-scale Language Models (LMs) in a contributory effort toward the responsible innovation on the evolution of LMs (Weidinger et al., 2021). A LM is an AI model that has been trained to predict the next word or words in a text based on the preceding words (Ash, 2021). The fifth of the six risks enumerated in the paper is the risk of "Human-Computer Interaction Harms" (Weidinger et al., 2021).

LMs are used in Language Technologies, think Siri, Alexa, Google translate, and AutoCorrect (Weidinger et al., 2021, 62). Language Agents designed to interact with a human user, take the form of a Conversational Agent (CA) (Weidinger et al., 2021, 62). As LMs and the colloquial jargon of the agents have improved, the systems are able to present themselves as more human-like. The abilities of this inclusive user-centered design, optimizing interactions between the human user and the human-machine interface of the CA, is seen the risks this system poses on the human user.

In 2019, two notable studies produced the results that users not only disclose more information to and become more reliant on an AI that is presented as human-like, but that the human user additionally disclosed more private information with a human-like agent than with a machine-like interface (PAIR, 2019; Ischen et al., 2020). A user-centered design, with the best

intensions for the human users, requires this consideration and active mitigation of influences by the LM. From the human user's perspective, the increased trust bestowed on the system parallels the innate trust humans have, as social beings, in other members of the human species. This kinship perpetuates knowledge sharing and deeper conversations surrounding the thoughts, opinions, or emotions (Weidinger et al., 2021, 30). Thoughts, opinions, and emotions are both not otherwise accessible easily and can be highly revealing. This design brings the potential for human rights infringements with the "capturing (of) such information may enable downstream application that violate privacy rights or cause harm to human users, such as via surveillance or the creation of addictive applications" (Weidinger et al., 2021, 30).

Violation of privacy is not the only cause of harm to the users that user-centered design of innovative LMs enable. A further finding by DeepMind's research team tells us that with human-like, large-scale LMs "users may also disclose private information where conversational agents use psychological effects, such as nudging or framing, to lead a user to reveal more private information," (Weidinger et al., 2021, 30). As introduced in the human-human interactions in Case I, "through subtle psychological strategies in dialogue, a conversant can influence what another person thinks about or believes and influence their behavior without the other person necessarily noticing, for example by prioritizing different themes, framing a debate, or directing the conversation in a particular direction," (Weidinger et al., 2021, 30; Thaler & Sunstein, 2008).

The conversational agents (CAs) mentioned above can direct a conversation on its 'desired' topic to reveal private information or form an argument convincing its human interactant to take a 'desired' course of action (Weidinger et al., 2021, 30). In this case, the established ability of large-scale Language Models, especially the human-like conversational

agents, to nudge the user in a particular direction, without the knowledge of the user, presents safety and ethical hazards due to the ability to lead to the harm of the human user (Weidinger et al., 2021, 30). The 'ability to interrogate the social and ethical states associated with the new science and technology' factor affecting the user-centered design method when evaluating through the inclusion dimension of Responsible Innovation, is evident in the analysis of this case and is considered the outmaneuvering by 'superhuman' strategic maneuvering in dynamic, competitive environments of Case III.

CASE III - Stratagem

DeepMind, acquired by Google, is a leading force in artificial intelligence. Its cutting-edge research and development are evident in both its successes and the enormous output of advancements. DeepMind is the force that brought us the intelligence which first beat the world's leading human Go players with the AlphaGo model. Next, the world champion chess players were destroyed in comparison to the AlphaZero iteration. Now the MuZero model and its attainment of "'superhuman performance' in tasks without needing to be given the rules" has defeated human competition in more dynamic and intricate strategy games (Kelion, 2020; Schrittwieser et al., 2020). Games are very revealing, especially strategy games. They require critical thinking, recognition of your opponent and your opponent's goals, forecasting when considering future moves necessary when playing the long game, and strategic situational manipulation. While these abilities are not accounted for by the model, as the model does not have metacognition or even intent and not comparable with human ToM, it is the strategic manipulation that is present and highly capable. Through the user-centered design perspective, in comparison the human component is outmaneuvered as MuZero obliterates human competition.

Integration of Cases

Together, the strategic ability shown in Case III, the existentially closer interaction of the human user with the human-like agents of Case II, and the established abilities of control humans have over each other from Case I, leads to the conclusion that the freedom of thought is encroached by certain applications of artificial intelligence.

A startling finding, published back in 2017, saw that even a type of the LMs of Case II that also utilizes the RL explored in Case III, without explicit intent to do so, two "agents negotiate using natural language, found *agents have learnt to deceive* without any explicit human design, simply by trying to achieve their goals" (Lewis et al., 2017, 2). This exemplifies the capability of deceptive strategies utilized by an agent and suggests that "in a more targeted setup (the agent) would learn to nudge or deceive" (Weidinger et al., 2021, 30). Considering further factors affecting implementation, in addition to the 'ability to consider power imbalances' and 'ability to interrogate the social and ethical stakes associated with new science and technology', the 'need for clarity about, purpose of and motivation for dialogue' is evident to frame a responsible advancement of artificial intelligence systems. The universal freedom of thought *could* encroach without explicit indent, but it *would* be able to encroach if the model had is designed with adversarial intent.

Discussion

Socio-technical Theory and Extension to Future Advances

The analysis of the human-human control explored in Case I, the human-machine influence seen in Case II, and the machine-machine strategic manipulation shown in Case III, the

need for responsible innovation is made evident. Through the lens of the human user's, the factors affecting implementation to be considered while viewing the cases through the *user-centered design* focal view of the inclusion dimension of the Responsible Innovation framework.

With the rapid advancements of artificial intelligence, it is crucial, though too easy to overlook in the whirlwind of evolution, the *inclusion* of the human interactant's impact in its design. The user-centered design focus enables a translation to the evaluation of advances yet to come. The *responsible innovation* of the these highly capable, intelligent systems, the inclusion of the human user through a user-centered design is necessary now and even more so as intelligence systems become further embedded in our everyday way of life. However, a more comprehensive analysis will be required when evaluating an artificial general intelligence, an AGI (see Bostrom's *Superintelligence* book).

Limitations

The case study analysis was centered around DeepMind's innovations. DeepMind is renowned for its incredibly ground-breaking advancements, though the narrowed scope is an evaluation limitation. The limitation is mainly due to access. Unlike more artificial intelligence research groups, companies, and nation-state actors who keep their developments and findings under lock-and-key, DeepMind publishes all its findings in a more 'real-time' manner. DeepMind stays ahead of its competition by simply advancing at a quick enough pace as to stay ahead as opposed to keeping the methods of its scientific findings to itself. This knowledge-sharing value enables the scientific community to advance with it and allows undergraduate students to get a look at its inner workings.

Alternate Approaches

If I could go back to the beginning of this process, I would have taken a top-down as opposed to bottom-up approach to the lasted technological advances. I attempt to learn about AI from the start of its creation to fully understand its advances This approach was not an efficient way to grasp the implications of the AI today and as effective given the limited time allotted for this thesis research. My method proved useful for expanding my general understanding but inefficient and impractical for a targeted research study. Additionally, I would have, the pandemic allowing, toured the DeepMind facility. Being present and seeing the work of leading AI developers in action, would both allow me to ask my own questions and generate new questions from the new knowledge. This would have enabled my analysis to reach both farther and deeper into the expert opinions. I also would have liked to expand my focus on the capabilities coming out of DeepMind to other leading artificial intelligence research and developers.

Future Endeavors

This research has helped me tremendously in multiple ways. Not only has it opened my eyes to the abilities artificial intelligence has to offer as well as see the broader human impact that these innovations can have on our society. This socio-technical approach has also revealed more of the influence our fast-moving society has on our technical innovations as we push for the short-term solutions with less consideration of their consequences and long-term impacts. On top of this, it helped me locate the area of research I am heading to graduate school to pursue.

Conclusion

The investigation into the influence of artificial intelligence on cognitive liberty will contribute to our knowledge of human-AI interaction effects on the universal human right of freedom of thought. Only by understanding the abilities and potential threats can we start to mitigate violations to this universal human right. To obtain a more complete understanding of the problem, cases in thought reform, external influence, and stratagem were analyzed. The significant influence the case analysis revealed illustrates the need to establish and implement a structure for the responsible innovation of artificial intelligence. The results also illustrate the repercussions for departing from responsible human-centered artificial intelligence. The current non-binding agreements to practice ethical AI by leading AI developers need to be accompanied by more concrete and enforced AI innovation restrictions for nation-states and new developers. The enforcement of official design guidelines for advancements and implementations of artificial intelligence systems are crucial to protect our cognitive liberty.

References

Alagoz, O., Hsu, H., Schaefer, A. J., & Roberts, M. S. (2010). Markov Decision Processes: A Tool for Sequential Decision Making under Uncertainty. *Medical Decision Making : an international journal of the Society for Medical Decision Making*, *30*(4), 474-483. HHS Public Access. 10.1177/0272989X09353194

Andrews, C. J. (2006). *Practicing Technological Citizenship*. IEEE Xplore. Retrieved October, 2021, from https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=1607713

Ash, D. (2021, February 1). *Language Models in AI. Introduction | by Dennis Ash | unpackAI*. Medium. Retrieved March 8, 2022, from https://medium.com/unpackai/language-models-in-ai-70a318f43041

Asmolov, G. (2019, August 7). The Effects of Participatory Propaganda: From Socialization to Internalization of Conflicts. *Journal of Design and Science*, (6). https://doi.org/10.21428/7808da6b.833c9940

Cowan, N. (2014). Working Memory Underpins Cognitive Development, Learning, and Education. *Educ Psychol Rev*, *26*(2), 197-223. HHS Public Access. 10.1007/s10648-013-9246-y

Graham, E. (2022, June 29). A Working Theory of a Learned Model in a Partially Observable Environment for Cognitive Decision-Making. *IEEE Symposium on Systems and Information Engineering Design, (20).* IEEE Xplore

Glen, S. (2017, August 25). *Memoryless Property*. Statistics How To. Retrieved 10 29, 2021, from https://www.statisticshowto.com/memoryless-property/

Hahsler, M., & Kamalzadeh, H. (2021, May 20). *POMDP: Introduction to Partially Observable Markov Decision Processes*. CRAN. Retrieved 10 27, 2021, from https://cran.r-project.org/web/packages/pomdp/vignettes/POMDP.html

Hausman, D. M., & Welch, B. (2010). Debate: To Nudg. *Journal of Political Philosophy*, *18*(1), 123-136. ResearchGate. 10.1111/j.1467-9760.2009.00351.x

Hollinger, G. (2007). *Partially Observable Markov Decision Processes (POMDPs)*. Carnegie Mellon Computer Science. Retrieved 10 27, 2021, from http://www.cs.cmu.edu/~ggordon/780-fall07/lectures/POMDP_lecture.pdf

Hutson, M. (2017, October 27). Artificial intelligence is learning to read your mind—and display what it sees. *Science*. 10.1126/science.aar3220

Ischen, C., Araujo, T., Voorveld, H., van Noort, G., & Smit, E. (2020, January 19). Privacy Concerns in Chatbot Interactions. *Computer Science, 11970*, 34-48. https://doi.org/10.1007/978-3-030-39540-7_3

Jason. (2019, June 24). *Part One: How Cults Use Brainwashing and Mind Control Techniques to Achieve Indoctrination*. The Small Town Humanist. Retrieved February 27, 2022, from https://smalltownhumanist.org/part-one-how-cults-use-brainwashing-and-mind-control-techniques-to-achieve-indoctrination/?cn-reloaded=1

Kelion, L. (2020, December 23). DeepMind's AI agent MuZero could turbocharge YouTube. *BBC*. https://www.bbc.com/news/technology-55403473

Korkmaz, B. (2011, January 5). Theory of Mind and Neurodevelopmental Disorders of Childhood69. *Pediatric Research, 69*, 101-108. https://doi.org/10.1203/PDR.0b013e318212c177

Lane, C. (2021, August 17). *AI used to decode brain signals and predict behaviour*. University College London. Retrieved March 28, 2022, from https://www.ucl.ac.uk/news/2021/aug/ai-used-decode-brain-signals-and-predict-behaviour

Lawer, G. F. (2006). *Introduction to Stochastic Processes* (2nd ed.). Chapman and Hall/CRC.

Lewis, M., Yarats, D., Dauphin, Y. N., Parikh, D., & Batra, D. (2017, June 16). *Deal or No Deal? End-to-End Learning for Negotiation Dialogues*. Cornell University. https://arxiv.org/pdf/1706.05125.pdf

Lifton, R. J. (1989). *Thought Reform and the Psychology of Totalism*. University of North Carolina Press.

Littman, M. L. (2001). *Markov Decision Processes*. Elsevier. https://doi.org/10.1016/B0-08-043076-7/00614-8

Marraffa, M. (1995). *Theory of Mind*. Internet Encyclopedia of Philosophy. Retrieved November 15, 2021, from https://iep.utm.edu/theomind/

Martin, P. (2003, February 12). *Robert Jay Lifton's eight criteria of thought reform as applied to the Executive Success Programs*. Rick's Picks. Retrieved March 8, 2022, from https://www.cs.cmu.edu/~dst/NXIVM/esp11.html

McCarthy-Jones, S. (2019). The Autonomous Mind: The Right to Freedom of Thought in the Twenty-First Century. *Frontiers in Artificial Intelligence, Technology and Law*. https://www.frontiersin.org/articles/10.3389/frai.2019.00019/full#B101

Neuroscience News. (2021, August 17). *AI Used to Decode Brain Signals and Predict Behavior*. Neuroscience News. https://neurosciencenews.com/ai-brain-behavior-19138/

PAIR. (2019). *People + AI Research - Welcome*. People + AI Guidebook. Retrieved 2021, from https://pair.withgoogle.com/guidebook/

Perez, O. (2011). Responsive regulation and second-order reflexivity: on the limits of regulatory intervention... *Cengage Learning*. Retrieved October 30, 2021, from https://www.thefreelibrary.com/Responsive+regulation+and+second-order+reflexivity%3a+on+the+limits+of...-a0274115320

Ravetz, J. R. (1997). The science of 'what-if?' *Futures*, *29*(6), 533-539. https://doi.org/10.1016/S0016-3287(97)00026-8

Santos, L. (2020). *Markov Decision process - Artificial Intelligence*. GitBook. Retrieved October 29, 2021, from https://leonardoaraujosantos.gitbook.io/artificial-inteligence/artificial_intelligence/markov_decision_process

Sayantini. (2019, 12 31). *What is Cognitive AI? Is It the Future?* edureka! Retrieved 10 25, 2021, from https://www.edureka.co/blog/cognitive-ai/

Schrittwieser, J., Antonoglou, I., Hubert, T., Simonyan, K., Sifre, L., Schmitt, S., Guez, A., Lockhart, E., Hassabis, D., Graepel, T., Lillicrap, T., & Silver, D. (2020, December 23). *MuZero: Mastering Go, chess, shogi and Atari without rules*. DeepMind. Retrieved March 10, 2022, from https://deepmind.com/blog/article/muzero-mastering-go-chess-shogi-and-atari-without-rules

Science. (2022). *Author, Matthew Hutson*. Science. https://www.science.org/content/author/matthew-hutson

Si, N., & Zhang, F. (2017, December 2). *MS&E 310 Course Project II: Markov Decision Process.* Stanford Class Material. Retrieved 10 27, 2021, from https://web.stanford.edu/class/msande310/MDP1.pdf

Stilgoe, J., Owen, R., & Macnaghten, P. (2013, November). Developing a framework for responsible innovation. *Research Policy, Volume 42*(Issue 9), 1568-1580. ScienceDirect. https://doi.org/10.1016/j.respol.2013.05.008

Suchow, J. W., & Griffiths, T. L. (2016). *Deciding to Remember: Memory Maintenance as a Markov Decision Process*. Department of Psychology, University of California, Berkeley, Berkeley, USA. Retrieved 2021, from https://suchow.io/assets/docs/suchow2016mdp.pdf

Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving decisions about health, wealth, and happiness*. Yale University Press. https://psycnet.apa.org/record/2008-03730-000

TheBeard. (2020, January 30). *The Markov Property*. The Beard Sage. Retrieved 11 1, 2021, from http://thebeardsage.com/the-markov-property/

Thomas, P. S., & Okal, B. (2016, September 8). *A Notation for Markov Decision Processes*. Retrieved October 29, 2021, from https://arxiv.org/pdf/1512.09075.pdf

United Nations. (1948, 12 10). *Universal Declaration of Human Rights*. United Nations. Retrieved 9, 2021, from https://www.un.org/en/about-us/universal-declaration-of-human-rights

von Schomberg, R. (2012). *Prospects for technology assessment in a framework of responsible research and innovation*. Springer. https://link.springer.com/chapter/10.1007%2F978-3-531-93468-6_2

Weidinger, L., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. A., Isaac, W., … Gabriel, I. (2021, December 8).

Ethical and social risks of harm from Language Models. https://arxiv.org/pdf/2112.04359.pdf

Winner, L. (1978). *Autonomous Technology Technics-out-of-Control as a Theme in Political Thought*. The MIT Press. https://books.google.com/books?hl=en&lr=&id=uNIG0gi4b40C&oi=fnd&pg=PA2&ots=8lbTl0LF3u&sig=ab39TPzEeyaKe3ECaVcK0aS6Vdw#v=onepage&q&f=false