Behavioral Analysis of the Cognitive Process of Decision Making

Artificial Intelligence Systems' Impact on Cognitive Liberty

A Thesis Prospectus
In STS 4500
Presented to
The Faculty of the
School of Engineering and Applied Science
University of Virginia
In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science in Systems Engineering and Mathematics

By
Emma Graham

November 1, 2021

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.

ADVISORS

Rider Foley, Department of Engineering and Society

Leidy Klotz, Department of Engineering Systems and Environment

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). A major concern is the out-pacing of our theory of mind's ability by that of artificial intelligence. 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 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).

Andrews (2006) agrees 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. 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 explores this issue through a technical exploration of the cognitive process and a socio-technical investigation of these technological innovations' impact on the freedom of the individual. The technical component is a behavioral analysis of the decision-making through the computational capture of cognitive processes while the societal component analyzes the impact of artificial intelligence systems on cognitive liberty.

Technical Project Description

The technical component is a behavioral analysis of decision-making through the behavioral analysis of the cognitive processes. The computational-capturing, or modeling, of the cognitive process will be examined by the partially observable Markov decision process. The Markov Decision Process (MDP) is a ubiquitous mathematical framework that "models optimal decision-making process in complex dynamic systems" (Si & Zhang, 2017; Littman, 2001, 9240). In the effort to standardize the notation of the mathematical representation of the MDP, in

2015 Thomas and Okal published “A Notation for Markov Decision Processes” giving the standard notation that follows:

The MDP is expressed as a tuple, $(\mathcal{S}, \mathcal{A}, \mathcal{R}, P, R, d_o, \gamma)$, where

1. $\mathcal{S}$ is the set of possible states that the agent can be in. The state at time $t \in T$ is expressed by S_t and s is used to denote an element in $\mathcal{S}$.
2. $\mathcal{A}$ denotes the set of the courses of action available to the agent. The state at time t is expressed by A_t and a is used to denote an element in $\mathcal{A}$.
3. $\mathcal{R} \subseteq \mathbb{R}$ denotes the reward set, the set of utilities the agent can receive. The state at time t is expressed by R_t and r is used to denote an element in $\mathcal{R}$.
4. $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is the transition function and characterizes the distribution over states, denoted $\forall\ (s, a, s', t) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times T,\ P(s, a, s') := Pr(S_{t+1} = s' | S_t = s, A_t = a)$.
5. R denotes the reward function which distributes the reward over time t given S_t, A_t, S_{t+1} s.t. $\forall\ (s, a, s', t, r) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times T \times \mathcal{R},\ R(s, a, s', r) := Pr(R_t = r | S_t = s, A_t = a, S_{t+1} = s')$
6. $d_o : \mathcal{S} \to [0, 1]$ is the initial state distribution and $d_o(s) := Pr(S_o = s)\ \forall\ s \in \mathcal{S}$.
7. $\gamma \in [0, 1]$ denotes the reward discount parameter.

The Markov Decision Process and Markov models are distinguished by the Markov property. The Markov property is known as the memoryless property (TheBeard, 2020). Explicitly, the Markov property “states that to make predictions of the behavior of a system in the future, it suffices to consider only the present state of the system and not the past history” (Lawer, 2006, 9) The property is expressed mathematically as,

$\mathbb{P}\{X_n = i_n | X_o = i_o, ..., X_{n-1} = i_{n-1}\} = \mathbb{P}\{X_n = i_n | X_{n-1} = i_{n-1}\}$ (Lawer, 2006, 9).

A way of thinking of the memoryless property is the flipping of a coin. No matter how many times you flip the coin, the chance of obtaining a heads or tails result is consistently fifty percent and is independent of past flips (Glen, 2017). The memoryless property allows the

current situation to be the only perspective considered for the decision-maker, however as we make decisions, we also consider our past decisions.

But what about the memory component of cognition people and animals constantly pull from, providing information from and recall of stored information and past decisions. The memoryless property may then seem like an inconsistency in the parallel between cognition and this mathematical decision process. However, there is a solution to the modeling of working memory, the readily accessible information retained and used executing a cognitive task, and this solution is even provided by a Markov decision model. The process of maintaining the memories the mind decides to maintain in our working memory has been shown to be equivalent to a Markov decision process model. This modeling was introduced by University of Berkley's Jordan Suchow and Thomas Griffiths in their 2016 technical paper "Deciding to Remember: Memory Maintenance as a Markov Decision Process" (Cowan, 2014, 1-2) (Suchow & Griffiths, 2016).

The MDP is utilized to represent the cognitive process because its models provide puissant analytical instrument for uncertain sequential decision making and the process is a framework for producing optimal actions for decisions in a stochastic environment (Santos, 2020) (Alagoz et al., 2010). These decision processes are highly representative of our biological cognitive process which is itself a *powerful analytical tool* used to produce the best courses of action in a complex, uncertain world to achieve our goals. The current state perspective that the Markov models inherently have mimics the decision space cognitive beings have when making decisions.

Many models use the Markov property, widely utilized Markov models include Markov chain Monte Carlo, Markov decision process, hidden Markov models, and the partially

observable Markov Decision Process. The divergence between the models is represented in Table 1 below.

Table 1
Markov Models

Markov Models		**Do we have control over the state transitions?**	
		NO	**YES**
Are the states completely observable?	**YES**	**Markov Chain**	**MDP** Markov Decision Process
	NO	**HMM** Hidden Markov Model	**POMDP** Partially Observable Markov Decision Process

Note Recreation of the chart illustrated in Hollinger, G. (2007). *Partially Observable Markov Decision Processes (POMDPs)*. Carnegie Mellon Computer Science

"A partially observable Markov decision process (POMDP) is a combination of an MDP to model system dynamics with a hidden Markov model that connects unobservant system states to observations" which is "general enough to model a variety of real-world sequential decision-making problems" (Hahsler & Kamalzadeh, 2021). This *real-world sequential decision-making* allows POMDPs to be able to represent the cognition that we, as humans with our agency to make choices, *control over the state transitions*, based on our understanding, *incomplete state*

observability, of the state in which we are making the decision. The decisions we make are to, in effect, optimize a situational goal. Correspondingly, PODMPs give a policy formulating optimal actions for each state of belief (Hahsler & Kamalzadeh, 2021). These characteristics and applications make the partially observable Markov decision process a preeminent mathematical description of cognitive decision making.

Socio-Technical Topic

Understanding more about our cognitive process enables us to obtain a better understanding of external and internal effects on our decision-making, including the impact of artificial intelligence on our cognition. By breaking down our own cognitive decision-making process, we can better equip inanimate objects with understood decision-making. Artificial neural networks are "paving the way for new discoveries and a closer integration between technology and the brain" and artificial intelligence has made waves in quantifying the decision-making process (*AI Used to Decode Brain Signals and Predict Behavior*, 2021). As advancements in artificial intelligence are made, influences come full circle as artificial intelligence advances understanding 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, n.d.) (Hutson, 2017). After his subtitle "A computer guesses what people are watching based on brain activity", Hutson 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. Theory of Mind 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 the decisive advantage, opening the door for manipulation of the end receiver of the intelligent system by the administrator.

Many dystopian implications that stem from the utilization of this technological ability by 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.

Prior research indicates that responsibility in science and technology is based on both technical understanding and political and social interpretation (Winner, 1978). 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)" in the 2011 publication *Prospects for technology assessment in a framework of responsible research and innovation* (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). 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). 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 ("Responsive Regulation and Second-Order Reflexivity: On the Limits of Regulatory Intervention..." 2014). 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 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). Integrating the dimensional analysis examines the transition between deliberation, reflection, and anticipation, and that of agency and action is drawn. This transitional introspection is explicitly requiring the connection with societal values and governance practices (Stilgoe et al., 2013).

Responsible innovation questions can be inquired through the process methods in the table below,

Table 2

Four dimensions of responsible innovation.

Dimension	Indicative techniques and approaches	Factors affecting implementation
Anticipation	Foresight Technology assessment Horizon scanning 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 labour 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 Deliberative polling Lay membership of expert bodies User-centred design Open innovation	Questionable legitimacy of deliberative exercises Need for clarity about, purposes of and motivation for dialogue Deliberation on framing assumptions Ability to consider power imbalances 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[a] Value-sensitive design Moratoriums Stage-gates[b] 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.

Research Question and Methods

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?*

There are three research questions, stemming from the critical research question, that will be investigated. The three research questions are the following,

i. What are the risks to cognitive liberty associated with artificial intelligence innovations?

ii. Who is ultimately responsible to address the risks and ameliorate consequences?

The analytical information will be acquired through academic research as well as expert and stakeholder interviews. As discussed, the research exploration and interview data will then be assessed by the Responsible Innovation framework. The risks cognitive liberty imposes, question (i), is anticipatory in nature and the methods used will be the scenarios and vision assessment techniques dictated in Table 2. The inclusion and reflexivity approaches that will be taken in the exploration of question (ii) will be the inclusion techniques of consensus conferences and open innovation and the reflexivity technique of ethical technology assessment. The philosophical nature of the ultimate responsibility of artificial intelligence, question (iii), incorporations both the reflexivity and responsiveness dimensions. The embedded social scientists and ethicists in laboratories approach are most illuminating for the reflexivity dimension and the regulation technique, the telling technique for responsiveness. Wholistically,

the evaluated research questions will be connected to the societal values and governance practices of cognitive liberty.

Conclusion

The application of the partially observable Markov decision process to the mapping of cognitive decision making and the investigation into the influence of artificial intelligence on cognitive liberty will contribute to the knowledge of decision making and the universal human right of freedom of thought. Only by understanding the problem can solutions arise. The understanding attained through the technical and the socio-technological research proposed will supply the basis for innovative problem solving and the advancement of human rights.

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

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

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

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

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

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

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/

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/

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

Responsive regulation and second-order reflexivity: on the limits of regulatory intervention... (2014). Retrieved October 30, 2021, from https://www.thefreelibrary.com/Responsive+regulation+and+second-order+reflexivity%3a+on+the+limits+of...-a0274115320

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/

Science. (n.d.). *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

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

Winner, L. (1978). *Autonomous Technology Technics-out-of-Control as a Theme in Political Thought*. The MIT Press.