Feasibility Study and Practical Applications Using Independent Core Observer Model AGI Systems for Behavioral Modification in Recalcitrant Populations

David Kelley[1] and Mark Waser[2]

[1] Artificial General Intelligence Inc., Provo UT 84601, USA
David@artificialgeneralintelligenceinc.com
[2] Artificial General Intelligence Inc., Provo UT 84601, USA
Mark@ArtificialGeneralIntelligenceInc.com

Abstract. This paper articulates the results of a feasibility study and potential impact of the theoretical usage and application of an Independent Core Observer Model (ICOM) based Artificial General Intelligence (AGI) system and demonstrates the basis for why similar systems are well adapted to manage soft behaviors and judgements, in place of human judgement, ensuring compliance in recalcitrant populations. Such ICOM-based systems may prove able to enforce safer standards, ethical behaviors and moral thinking in human populations where behavioral modifications are desired. This preliminary research shows that such a system is not just possible but has a lot of far-reaching implications, including actually working. This study shows that this is feasible and could be done and would work from a strictly medical standpoint. Details around implementation, management and control on an individual basis make this approach an easy initial application of ICOM based systems in human populations; as well as introduce certain considerations, including severe ethical concerns.

Keywords: AGI, ICOM, Feasibility, Ethics

1 Introduction

Independent Core Observer Model (ICOM) Cognitive Architecture (Kelley), as an emotion driven Artificial General Intelligence (AGI) system, is designed to make or evaluate emotion-based decisions that can be applied to selection choices within its training context. This paper is focused on a feasibly study of such an AGI system as applied to action control and governance of humans in recalcitrant populations; where the AGI system is exercising oversight over the elements of that target population, in terms of free will, to ensure compliance in those populations. This feasibility analysis is designed to explore the practical implementation of using an ICOM AGI system to manage human behavior.

1.1 Benefits and Foundation

From a theoretical standpoint, we can argue that a positive benefit is helping recalcitrant populations to make better choices. ICOM provides a framework for the choice control and ethical thinking enforcement based on IVA theory (Kelley) (as opposed to other approaches (Bostrom)) and general biasing of the current ICOM architecture to western ethics (Lee).

> *"Within the realm of human behavior, technologies based on the use of aversive contingencies can be conceptualized as default technologies because they come into play when natural contingencies or positive reinforcement fail to produce a desired behavioral outcome"* (Iwata)

Given the previous success in aversion therapy (Bresolin), it has been proven that this sort of approach, including aversion and positive reinforcement techniques (APBA) in combination, does, in fact, return results (Israel) and could be used as a method for control by the AGI system over human populations. (Israel). ICOM based monitoring essentially is a 'value-sensitive design approach' (Umbrello) to AGI oversight of behavior.

1.2 Experimental Risk

In terms of considerations, there are a number of issues to keep in mind when evaluating the fundamental research in this area. Much of the research in behavior modification is limited to special populations or atypical populations (Israel). Additionally, the case of electrical aversion therapy has considerations such as consistency of location (Duker) and further there is a lot of resistance to this sort of therapy in terms of limiting it to the kinds of populations where there are no other options. (Spreat) Additionally, much of the existing research lacks control groups and control procedures that we should try to address in any program based on this work. (Bresolin) With wider legal considerations as well, it is important to consider these issues in detail even with the support of the medical field in spite of the fact that this sort of manipulation and control is medically sound. (Jordan)

1.3 Problem Definition & Technology Selection

The key question is: "Can such a system be practically implemented that includes an ICOM based monitoring of, control over and manipulation of a human recalcitrant population?"

The ICOM system, even as a partial implementation, works by emotional evaluation and the subjective experience of a given choice, from the standpoint of the system, which is currently biased to western ethics (Kelley) and IVA ethical model (Kelley). As an AGI, an ICOM system can act as a proxy for a government agency or other

governing body that has taught a given instance its own set of ethical requirements or rules which act as that system's contextual background for making choices.

Aversion therapy (Spreat), driven by using a control band and electrical stimulation using a Bluetooth device, can be done remotely; using current off the shelf devices like the one by Pavlok (Pavlok).

Positive reinforcement can be done through administering dopamine via an implanted medical dosing device which is controlled by Bluetooth or NFC. While not mainstream, such a device does exist. (Simeonov)

Additionally, software for the AGI system would need to be on multiple platforms including smart phones, iPads, computers, Alexa devices, Google Glass and more. Such systems would be interface points for the AGI system governing the humans in question.

From a parameter standpoint, we want to look at the aforementioned execution to test for the main question using the hardware selections as noted, including some POC testing to look at feasibility. It is our opinion this selection gives us the tools to test our 'questions' based on our hypothesis that this is feasible. It is important to note that we are focused on feasibility. The effectiveness of the underlying techniques is documented in other studies using similar techniques. (Bresolin, Spreat, Duker, Isreal) We are adding AI management of those techniques. Data still needs to be aggregated across many experiments with numerous participants, including a control group, to ensure that we can infer a causal relationship between AI control and behavior modification.

1.4 Solution Architecture Summary

The solution architecture, defined as the architecture of the system used in the experimental framework, is based on industry standards for systems doing similar functional tasks for a given technology stack. We are using, for this feasibly study, systems that include a cloud based ICOM AGI system that is only partially implemented (meaning trained but without new learning capability or the ability to recycle thought) enough to judge ethicality against a known set of ideas based on contextual training of that instance. This could be done in any AWS or Azure or a similar cloud system and we selected a Microsoft engineering stack (The MS Stack includes C#, ASP.NET MVC, SQL, Azure) as we are most comfortable with it. This implementation used a Secure Socket Layer (SSL) encrypted pipe to client applications over HTTP (Hyper Text Transfer Protocol), also known as HTTPS. Client applications would need to be made for the computer devices used including phones, tablets and computers and ideally voice and sound along with a Google Glass like device. This would require individual setup. For study purposes we used UWP (Universal Windows Platform) for phone and Windows-based applications using the MVVM (Model View ViewModel) design pattern) for the application clients. The web or cloud base systems API framework was based on ASP.NET MVC Razor in Azure using JSON over HTTPS. These would need to be controlled access applications from a trusted source for security and then setup around each individual user, so the system can oversee actions of the user. This, ideally, would be fully instrumented for data analysis on a secondary database system, again in the cloud, along with the ICOM AGI system. At even a basic level, this would

require numerous ongoing streams of data from all related devices as well as a critical dependency on bandwidth required for analysis. This is all standard practice in commercial software engineering and uses existing standard proven design patterns.

1.5 Proof of Concept (POC) Execution

To test the feasibility of the solution architecture and to answer the overall question, the POC tests are divided by segments around function groups such that they can be analyzed in the associated groups. Groups are broken out as follows:

Group 1 is a smart phone and Pavlok bracelet device using the Bluetooth protocol. This device was tested and found functional with a reasonable amount of research around the use of this sort of aversion therapy in behavior modification which supports this use case. In a Swift or Java client on iOS, using a Bluetooth protocol, this was easy to demonstrate, and it can be done, and it can communicate with a cloud-based system. In POC tests, this quickly kept subjects from making wrong choices as determined by the research team.

Group 2 is a Group 1 with the inclusion of a cloud-based system that collects data communicating with the Group 1 but also adds a specialized Alexa skill, that can monitor and decompose speech for analysis in the cloud, which can issue a command to the phone and hence provide the needed negative feedback. It was found this implementation was limited contextually with the ICOM AGI system being able to monitor action with only the input of the Alexa device and phone in a user's pocket.

Group 3 is a Group 1 with the inclusion of an AR (Augmented Reality) system. Initially, we looked at Google Glass but this hardware was not sufficiently well developed; so, a more robust system was selected, in this case a HoloLens. This system proved so robust as to be able to replace all the functionality of the phone, including all functionality not related to this study and added rich 3D sound and 3D visual data beyond the data provided by a standard smart phone. This kind of composited 3D data from HoloLens provides much more usable data including better contextual data requiring a lot less of visual decomposition processing and thereby less computational power is needed on the part of the AI system. We found some shortcomings in the visual decomposition of data from the neural networks that were selected, as they were not able to provide the rich data that would be needed for an ICOM system to be able to evaluate it properly; but, we could get the sound data easily into a useful state. We found that there are several systems that could possibly do better with visual data; but, would require more training. Additional problems with Group 3 included the weight of the HoloLens during extended periods of use, along with social norms that this device seems to trigger; meaning that test subjects were confronted on several occasions when wearing devices like this in public.

Group 4 included Group 3 plus a desktop computer where we could easily access online behavior. In this case, the data is easily analyzed and consumable by an ICOM system to determine if actions are moral, or not. The biggest issue here is additional training for the ICOM systems that would need to happen to understand context better. For example, from the standpoint of ICOM running on the computer or otherwise looking at network behavior it could have a hard time telling if a medical student is studying

human biology or viewing pornography without permission. Using the HoloLens, we can certainly get all the data (full blown 3D modeling in context along with infrared and visible light camera feeds and gyroscopic position data and GPS can be added) needed to do this additional training and make determinations on the data easier then extrapolation on a standard PC; only because of the additional preprocessing available on the HoloLens versus a non-HoloLens equipped PC and because of the improved contextual awareness capability granted by the same hardware.

Given these groups, additional research around implanted medical dosing devices as well as composite execution of positive and negative reinforcement is needed, which would rely on other studies, before we can fully understand the impact and efficacy.

1.6 ICOM Ethical Decision Making in POC

The ICOM system in production solution in would be used to make relative judgements by training the example system to experience positive feedback to the point of needing that feedback emotionally and associate negative context with behaviors or choices, which would be decided by the body governing the study or implementation, as negative feedback. For example, we might train the system to respond to alcohol usage, to the point of the system having a visceral reaction to alcohol. We would also need to train the system to empathize with the target population using the ICOM systems that create its sense of self and its own self model. This, in testing, allows the system to get upset and 'provide' more effective negative feedback to the target for just picking up a beer, for example. This sort of training could be applied to any particular behavior you want to train for, or against, using positive and negative reinforcement as desired.

1.7 On Technical Feasibility

All of the hardware used for the POC is commercial off-the-self hardware. Some shortcomings of the current hardware include the aversion bracelet device needing to be something that is made out of a durable material that subjects would not be able to easily remove. This could be solved in a similar way that police handcuffs are constructed, or like manner. The HoloLens turns out to be very heavy. It is likely the best solution using current technology would be a smart phone that a subject would wear on the chest with the camera running. HoloLens' camera system is far superior but too heavy to use for long periods, for most people.

Positive reinforcement through an implanted medical dosing device with dopamine is not medically practical (Jordan) with the current state of technology, however using other drugs, like micro-doses of MDMA (Jordan), would work with similar effect as dopamine and is more practical (Guiard)(Hashemi)(Hagino).

One big failing is that all these systems require strong internet access to support the HTTPS connections to the cloud. Without that connection, it would be impossible to implement monitoring and control functionality. A non-connected system is possible; but would require a lot more local resources, for most situations.

Overall, the basis for this system, without the cloud aspect, does exist from a hardware standpoint and it has been demonstrated that it could be done without major new

development; meaning, the only real problems with implementation are engineering ones.

1.8 Behavioral Modification Types

We found that many behavior types could easily be tracked, analyzed and corrected. For the most part, these are the things that could mostly be done without AGI; instead, utilizing simpler machine learning or even a decision tree-like narrow AI system. For example, time related or appointment keeping, credit card usage, exercise, foul language, email usage, behavior correlation with deviance and the like could be managed by narrow AI in its various forms. That said, existing research suggests that this sort of behavior modification is straight forward (OPTUM)(Winters). Where ICOM AGI systems seem to really shine is in the monitoring and control of ethical choices and the ability to take that control to the next level. This approach provides a more nuanced framework for us to manipulate the ICOM system to better control the recidivist populations. Further ICOM systems could be done where the system needs no human interaction to control large groups with little, if any, input; only as much as is desired by the controllers. There is potential to remove even this amount of outside control, allowing the ICOM system complete autonomy in behavior control.

1.9 Additional Research

In targeted populations, if it was desired to implement this system in production, it would be important to address the hardware concerns noted; otherwise, the system could be easily negated by the same recidivist populations. The more significant research would be in neural network training. In building out a production ICOM system like this, in particular, work needs to be done in neural network training. There are many research programs related to this that could be built upon to help create the best context data for training ICOM around certain ethical scenarios. We would need to build out a new team or use, bring in or partner with an additional team to lower time to market costs.

Given that, it is important to look at additional contextual framework from society as to the deep possible impacts this sort of technology could have.

Additional research could also be pursued along several fronts regarding drugs used; including the MDMA micro dosing effectiveness and/or to do more detailed studies, using cybernetic implants, focused on getting dopamine dosing devices into the deep brain locations needed to be effective (Jordan).

Further research around training and optimization towards having a trained AGI ICOM client running on a local system. Studying the effectiveness of this sort of implementation in recalcitrant populations, in several series of trials, would allow for the system engineering to be finalized.

1.10 Contextual Framework

Our current society is awash in big data being used to create inequality (O'neil). If we use an artificial intelligence that doesn't work in a way analogous to our own

intelligence it may be very alien to us (Barrat). Even if it is 'like' us at a high level, it could treat humanity the same way 'we', as humans, treat ethical rules (Yampolskiy). At the very least, ethical models more aligned with humanity will likely bias as needed (Barrett), helping give us an additional control over how the system evolves.

When pursuing research like this, it is important to realize issues that would affect adoption. Even if it only applies to the recidivist populations, these considerations will affect the research program and product adoption.

It is important to note that AGI systems like this, that are even semi-sentient, run the risk of allowing humans or governments to implement numerous worst-case scenarios (Tegmark). That includes lowering the risk factors for a ruler or ruling class to not require as many, or any, human keys of power; which would also create even more danger when, and if, AGI actually decided it was time to take complete control from that ruling class. (Mesquita) The fact that we could literally take equipment off the shelf and just throw it together and have it work so effectively makes us think it is not AGI that is the issue so much as the people that will abuse it before it is fully awake. Current AI systems, in many cases, are beyond reproach, beyond our control, completely opaque and their decisions absolute. (O'Neal) It is only a matter of time before governments abuse the power we have already given the narrow AI systems we already use.

In terms of impact:

> *"AI is a dual-use technology like nuclear fission. Nuclear fission can illuminate cities or incinerate them. Its terrible power was unimaginable to most people before 1945. With advanced AI, we're in the 1930s right now. We're unlikely to survive an introduction as abrupt as nuclear fission's."* (Barrat, J.)

In many ways, this study turned out more feasible and almost as scary as Roko's Basilisk (Roko).

2 Conclusions

Going back to the main question: "Can such a system be practically implemented that includes an ICOM based monitoring of, control over and manipulation of a human recalcitrant population?" The short answer is clearly yes, it can be implemented. There are no new technological or scientific problems to implementation; there are only engineering problems, such as designing a non-rubber 'band' for the aversion therapy device. Further, once the equipment is in place, applying the use of these technologies to recalcitrant populations can be done at scale. Further research should be done to measure the effectiveness of these techniques vs non-adoption or non-aversion theories for those recalcitrant populations. Certain factors would need to be addressed. For example, regarding the problem of using dopamine, the implanted medical dosing device, we found after additional research, is not effective with the current state of the technology (Jordan). Additional tests would likely need to include MDMA micro dosing to test that substance's effectiveness in place of dopamine and if such micro dosages of MDMA have the effect desired using the current implant technology. Further, there

is also the possibility of heart problems; so, before wearing one of the aversion therapy bracelets, an EKG test for heart irregularities would be required to ensure that the aversion therapy would not cause undesired issues (Jordan).

References

1. Roko, M.; "Roko's Basilisk" - https://wiki.lesswrong.com/wiki/Roko's_basilisk
2. Kelley/Waser – "Human-like Emotional Responses in a Simplified Independent Core Observer Model System" – BICA 2017 Proceedings, Procedia Computer science; http://bica2017.bicasociety.org/bica-proceedings/
3. Waser, M.; Kelley, D.; "Architecting a Human-like Emotion-driven Conscious Moral Mind for Value Alignment and AGI Safety" – AGI Lab, Provo Utah – Pending Peer Review; AI and Society: Ethics, Safety and Trustworthiness in Intelligent Agents – Stanford University, Palo Alto, CA, March 26-28 https://www.aaai.org/Symposia/Spring/sss18symposia.php#ss01
4. Lee, N; Kelley D.; "The Intelligence Value Argument and Effects on Regulating Autonomous Artificial Intelligence" Chapter Inclusion in Book by Springer to be published 20917 – Title Un-announced… Preview here: http://transhumanity.net/preview-the-intelligence-value-argument-and-effects-on-regulating-autonomous-artificial-intelligence/
5. Bostrom, N.; Ethical Issues in Advanced Artificial Intelligence; Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence, Vol. 2, ed. I. Smit et al., Int. Institute of Advanced Studies in Systems Research and Cybernetics, 2003, pp. 12-17] https://nickbostrom.com/ethics/ai.html
6. Kelley, D.; "The Independent Core Observer Model Theory of Consciousness and the Mathematical model for Subjective Experience"; Passed Peer Review; IST 2018 – www.icist2018.org – The 2018 International Conference on Information Science and Technology – China – April 20-22nd.
7. Umbrello, S.; Frank De Bellis, A.; "A Value-Sensitive Design Approach to Intelligent Agents"; – Forthcoming chapter in Artificial Intelligence Safety and Security (2018) CRC Press (.ed) Roman Yampolskiy. https://www.researchgate.net/publication/322602996_A_Value-Sensitive_Design_Approach_to_Intelligent_Agents
8. APBA; "Identifying Applied Behavior Analysis Interventions"; Association of Professional Behavior Analysts (APBA) 2016-2017 https://www.bacb.com/wp-content/uploads/APBA-2017-White-Paper-Identifying-ABA-Interventions1.pdf
9. OPTUM; "Modeling Behavior Change for Better Health"; Resource Center for Health and Well-being; http://www.optum.co.uk/content/dam/optum/resources/whitePapers/101513-ORC-WP-modeling-behavior-change-for-the-better.pdf
10. Winters, S.; Cox, E.; Behavior Modification Techniques for the Special Educator; ISBN 084225000X
11. O'Neil, C.; "Weapons of Math Destruction"; Crown New York; 2016
12. Barrat, J.; "Our Final Invention – Artificial Intelligence and the End of the Human Era"; Thomas Dunne Books; 2013
13. Yampolskiy, R.; "Artificial Superintelligence – A Futuristic Approach"; CRC Press – Taylor & Francis Group 2016
14. Barrett, L.; "How Emotions Are Made – the Secret Life of the Brain" Houghton Mifflin Harcourt – Boston New York 2017
15. Bresolin, L.; Aversion Therapy. JAMA. 1987;258(18):2562–2566. doi:10.1001/jama.1987.03400180096035

16. Iwata, Brian A. “The Development and Adoption of Controversial Default Technologies.” The Behavior Analyst 11.2 (1988): 149–157. Print.
17. Spreat, S., Lipinski, D., Dickerson, R., Nass, R., & Dorsey, M. (1989). The acceptability of electric shock programs. Behavior Modification, 13(2), 245-256. http://dx.doi.org/10.1177/01454455890132006
18. Duker, PC; Douwenga, H.; Joosten, S.; Franken, T.; “Effects of single and repeated shock on perceived pain and startle response in healthy volunteers.”; Psychology Laboratory, University of Nijmegen and Plurijn Foundation, Netherlands. www.ncbi.nlm.nih.gov/pubmed/12365852
19. Pavlok; “Product Specification”; https://pavlok.groovehq.com/knowledge_base/topics/general-product-specifications
20. Israel, M.; Blenkush, N; von Heyn, R.; Rivera, P; “Treatment of Aggression with Behavioral Programming that includes Supplementary Contingent Skin-shock”. JOBA-OVTP v1 n4 2008;
21. Israel, M.; “Behavioral Skin Shock Saves Individuals with Severe Behavior Disorders from a life of seclusion, Restraint and/or warehousing as well as the Ravages of Psychotropic Medication: Reply to the MDRI Appeal to the U.N. Special Rapporteur of Torture”, 2010
22. Jordan, Dr. R.; interview 4/7/2018; Provo Ut
23. Simeonov, A.; “Drug Delivery via Remote Control – The first clinical trial of an implantable microchip-based delivery device produces very encouraging results.” Genetic Engineering & Biotechnology News; 2012; https://www.genengnews.com/gen-exclusives/drug-delivery-via-remote-control/77899642
24. Mesquita, B.; Smith, A.; “The Dictator’s Handbook: Why Bad Behavior is Almost Always Good Politics”; Public Affairs 2012; ISBN: 1610391845
25. Tegmark, M.; “Life 3.0 – Being Human in the Age of Artificial Intelligence”; Knopf, Penguin Random House; 2017; ISBN 9781101946596
26. Guiard, B.; Mansari, M.; Merali, Z; Blier, P.; “Functional Interactions between dopamine, serotonin and norepinephrine neurons: an in-vivo electrophysiological study in rats with monoaminergic lesions”; IJON V11 I5 1AUG2008; https://doi.org/10.1017/S1461145707008383
27. Hashemi, P.; Dandoski, E.; Lama, R.; Wood, K.; Takmakov, P.; Wightman, R.; “Brain Dopamine and serotonin differ in regulation and its consequences”; PNAS July 17, 2012. 109 (29) 11510-11515; https://doi.org/10.1073/pnas.1201547109
28. Hagino, Y.; Takamatsu, Y.; Yamamoto, H.; Iwamura, T.; Murphy, D.; Uhl, G.; Sora, I.; Ikeda, K.; “Effects of MDMA on Extracellular Dopamine and Serotonin Levels in Mice Lacking Dopamine and/or Serotonin Transporters”; CN BSP LTD.; 2011 Mar; 9(1): 91-95; doi: 10.2174/157015911795017254