Project Specifications: growing autonomous intelligent aware* agents in a virtual environment
***aware as specified in broad terms below**

Project Initiator: Salvatore (Sam) Micheal

Initial Project Resource: Hanqi Zhuang

Preface: Several addendum were placed below with critical clarifications. Please pay attention to them. This document is essentially a feasibility proposal. Also, some considerations are proposed for 'project consequences' / implications of the project if successful. The project may be singularly unique in the sense it may be a 'watershed event' provoking similar attempts. It is hoped other project teams have similar respect for their own attempts and the responsibility implied by statements below. I cannot force this; I can only ask as part of this proposal. One final comment in that regard: how we treat our creations, specifically our *aware* creations, will *define our destiny as a race*. If we treat them *respectfully*, as we would treat our own children/ourselves, we will be fine. If we don't, we're essentially asking for mutual destruction 'down the road'. Let's not set the stage for mutual annihilation; let's *respect* our *aware* creations. Addendum: Please forgive this 'harsh tone' in caution. The source is the fact I almost invariably try my best to look at things from 'the other side' (here): the side of the agents – what would ***I*** want if I was an agent in a VE? Respect? Freedom? More on that later..

General Project Description:
Addendum: This project is small in the sense it can be accomplished by a team of five and some allocation of significant computing resources. However, the scope of this project is quite broad and I acknowledge that. In my personal experiences, individuals have been harshly dealt with by not managing this notion appropriately. I have been reprimanded myself for proposing this project to another faculty: "Sam, this project is *too big* for you." But that does *not* imply an *individual* cannot determine precise specifications for it. This document is my first attempt to do that.. Historically, AI has focused on 'intelligent' aspects of machine intelligence: information processing, expert systems, data storage and retrieval, and efficient algorithms relating to those. These are required for artificial intelligence. The concept of autonomous agent has recently gained popularity in research circles. However, there has been a notable lack of inspiration and identification of critical components relating to awareness. What is it? It's a fair question but we need to try to answer it in a non-trivial *functional* manner. We assume there are 'minimal requirements' for awareness since there is only one species on the planet capable of creating artificial awareness. This minimal set and configuration define the minimal characteristics of our agents. The environment needs to be suitable for developing those characteristics over *generations* of agents in a GP framework. In the end, it becomes a matter of detailed specifications regarding agent allocation and configuration. The computational, data storage, and configuration aspects assigned to individual agents will determine the ultimate success of the project. So we must be *extremely careful* how we assign computational resources to individual agents. In any scenario, computational resources are highly valued and we, as researchers, need to justify our requests for allocations of them. This project has the potential for being a 'game changer'. For example, we anticipate at least one team per large university allocating significant time and resources to similar projects. Inevitably, there will be competition between teams as to which can create *the first sentient agent* or set of agents. There are many ramifications: legal, moral, social,.. which result from such a historic development. We will attempt to address them, in broad terms, below.

General Specifications:
Addendum to below: this proposal is *not* asking for university allocation of a set of supercomputers; that notion is ludicrous in itself. But the team for this project needs access to significant computing

resources in order to test concepts/feasibility. Please read the following with that understanding in mind.. Typically, a virtual environment is hosted on a specialized server or set of servers. These are not 'supercomputers'. The distinction is a matter of degree; this is not a *qualitative* distinction. We assume some broad requirements that relate to supercomputers. What is conventionally available we label: *order 1*. An *order 2* set of supercomputers would be a parallel arrangement analogous to parallel CPUs in an individual computer architecture. We anticipate an order 2 set of supercomputers being a minimal requirement for this or similar project. Why? Simply because we require a virtual environment similar to our own 'real' environment which has similar salient characteristics. These are 'defined' by our two most utilized senses: vision and hearing. Vision is an extremely 'data intensive' activity. Our eyes transmit an extremely large amount of information to our visual cortex which we filter somewhat automatically depending on situational needs. Hearing is similar: we hear noises and consequences of events which register automatically in our awareness and longterm memory. Our longterm memory is similar to a 'list of events' with associated smells, feelings, perceptions, any any other individually determined characteristics of events we care to attend to. Our agents will *not* have the benefit of smell, feeling, or taste. So we must make sure our minimal representations of vision and hearing are *sufficient* for individual agent awareness. If this seems infeasible to the skeptical reader, we refer them to such achievements as flight and space travel; this can be done.

Broad Agent Specifications:
As mentioned above, minimal representations of vision and hearing are minimally required for agent awareness. Some form of 'optimal storage' of events, and therefore data representation, is required. Consider the event perceived by three nearby agents: 'bucket dropped'. They all 'hear' it. If they're looking in that direction, they all 'see' it. If we have performed our 'job' correctly, the event is added to each agent's longterm memory: *bucket dropped at time and place: (t, x)*. What do 'we' use the buckets for? Carrying water, watering plants, plants grow, plants feed us, we depend on plants for survival, and cooperative behavior increases our chances for survival. We need buckets and water to survive. We need to use them in cooperative behavior patterns. Our nutritional needs impel us to perform cooperative behavior to: use buckets effectively to carry water from a source to the place where plants are grown to feed us. We need to determine what season to plant seeds to grow plants which sustain us. We need to determine how and why to collect seeds. We need to *communicate* to: protect the clan from harmful animals, when to plant, when to water, when to reproduce, when to consume, where to defecate, and when to rest. Play is a neglected *critical* concept. We use play to: learn, recreate, reproduce, and increase bonding between individuals. All of these requirements define the minimal characteristics of agents: 'short term' memory, longterm memory, two distinct equivalent senses, a 'visualization register', some adaptive sense filtering, *a native language*, and capacity to be self-aware. Finally, a *unique connectivity* between these capacities will *define* individual awareness. Each individual agent will be *truly unique* as we are in 'real' life. The individual experiences, connectivity, and global system resources allocated to each individual agent will define the ultimate success of the project.

Environment Specifications:
As indicated above, we will utilize an *agrarian* scenario rewarding cooperative behavior with community benefits. These translate to *individual* benefits of caloric nutritional needs. Effective communication, cooperative behavior, productive play, and productive 'thoughts' all contribute to an individual's 'success' in each generation. We, as scenario controllers, must determine *exactly* what 'success' means. Otherwise, there will be no progress toward sentience. We estimate the bulk of global system resources need to be allocated to the set of *agents*. The environment is simply the *minimal frame* required for agent interaction between: agents and environment and that which allows senses to *operate*. In this perspective, senses are the *basis* for the environment and interaction.

Team Specifications:
We need two experts from computer science: a database expert and a robotics expert. The latter must have a flexible intelligence regarding sense-modeling. We also need an expert in virtual environments. Addendum: At the moment, I'm leaning toward recommending including a linguistics expert to help the team develop an *economical* agent language. Agent cooperation *requires* communication which requires we *give* agents a common language. Whether or not we gift agents with logic is an interesting concern. It might be worthwhile to investigate various scenarios.. Each expert must be 'team flexible' such that project needs take precedence. Minimally, we need a 'systems coordinator' who guides the others' efforts toward project goals.

Legal Specifications:
A. Patent Considerations:
We waive *all* legal rights to potential patents associated with this project in lieu of access to project resources and overall project goals. The university is free to patent *any* portion of the project in the interests of the university. We claim *no* intellectual property rights to *any* portion of the project. We *only* require recognition of project origin and development as appropriate and relevant. We legally *promise* ***any*** developments relating to this project specifically will *never* be used to dominate, subjugate, or exploit individuals related to or *created by* this project.
B. Responsibility Considerations:
It's inevitable we must consider legal ramifications relating to conscious entities. We, as entity creators, must be responsible and accountable for our creations. It's analogous to having children. We must be responsible parents but more-so: *we must be* ***advocates*** *for them*. We must be advocates for our children's *rights* whether or not they are our biological children or synthetic. We have a moral obligation to be good parents. Addendum: I cannot expect all team members or participants to hold this view. But I do expect at least an acceptance of this perspective and respect of *my* particular view as project initiator.

Social Implications:
Considering the scope of implications of this project, again, it's a 'game changer'. We cannot anticipate all social consequences of this project or related ones. However, some 'warnings' are evident: slavery and prostitution. We recommend avoiding social scenarios involving autonomous 'synthetic humans' where they are used as slaves or servants. This relates to 'robot rights'. A deeper concept is 'sentient rights'. The basic question is: does a 'sentient entity', regardless of 'host' (configuration of physical representation), have what we consider to be – equivalent human rights? If they are truly autonomous and sentient, it's difficult to argue otherwise. These considerations ***require*** democratic and justice attention. Addendum: it would be appropriate to ask the university's legal department to investigate this concern considering the feasibility and implications of the project.

Personal Statement of Integrity:
I, Salvatore Gerard Micheal, promise to never use any consequence of this project for personal gain. There are two main motivations for this particular project: 'can it be done?' and 'applications'. Personally, this project was motivated by what I call 'divine inspiration'. It's a spiritual consideration. Those who cannot comprehend or accept, must dismiss and 'walk away'. Those who can appreciate, may regard this as significant. It's totally up to you. Each of you must decide whether or not to accept this statement of integrity as false or true. Because I have no 'vested interest' in the project other than whether or not it's supported and pursued, I'm indifferent toward your *particular* opinion. *However*, if you feel inclined to participate, please express that directly toward our team.

Final Addendum:
There are three issues I'd like to discuss here: 1. novel forms of cooperation, 2. non-human interfacing, and 3. evaluating awareness.. We may be able to observe novel forms of cooperation that agents invent/discover in their community. So this 'experiment' would allow some interesting potential in regards to cooperative behavior; we may be able to learn from 'agent society'. The next point addresses the notion of gifting agents non-human forms of interfacing. For instance, if we created two 'mind ports' for each agent, they might be able to communicate much faster than we do. I suppose it depends on the internal architecture of implementation. If agents 'reside' as simply allocation of global resources with a unique internal connectivity, the idea of 'mind port' has no meaning. Perhaps the simplest interface between agents is the following: a display screen for each agent that allows 16 English words. Why 16? That's double our symbol register size for 'short term memory'. This is suggested to somewhat compensate for lack of sensory analogs. “Agent-22 go get water take to corn field water plants” might be an example of a 'cooperative command'. This would allow us to attend more easily 'conversations'. Each agent would permanently store conversations and local events; their memory will be perfect. If anything interesting occurs, we can detail explicitly the antecedents.. The third concern has more global / project consequences. Ideally, we will be able to rank all agents in terms of our criteria for awareness. Each generation of agents will have an explicit ranking of awareness. I suggest we adhere to the following schedule. Select the top two agents and 'mate' them in the following way. We create an entire next generation of agents based on varying combinations of aspects of each previous generation's 'Adam and Eve'. Of course, for continuity within the VE, both Adam and Eve will continue for at least one generation. Then we allow that generation to perform some cooperative behavior. Perhaps we can increase the level of difficulty as demanded by that particular generation's 'awareness capacity'. At some point, they will want to talk to us meaningfully. So we must gift them, as part of their vocabulary, notions regarding 'the outside' (of the VE). “I am Sam of those that reside outside your world” is the respectful response to “who are you?” from any particular agent. That implies we should have a 'temple' or community interface so that agents can communicate with us at will. Our responses should be truthful (regarding our world) and phrased in ways such that they can understand (limited to their vocabulary). So the project begins with the notion of respect and ends thusly. Again, we're not guaranteed all agents will respect us as their creators, but our chances are infinitely higher if we respect them *before* they're created.. Some VE specifications are in order. Clearly, to avoid confusion within the agent community, no two agents can reside in the same space. Agent destruction, murder, should not be possible. Agent harm, diminishing functionality, should not be possible. In this way, only growth will occur.

My email is micheal(at)msu.edu / smicheal(at)fau.edu