Supplementary Material for:

Assessing Machine Volition: An Ordinal Scale for Rating Artificial and Natural Systems

George L. Chadderdon

Department of Psychological and Brain Sciences, Indiana University, Bloomington, IN 47405

Corresponding Author:

George L. Chadderdon
Department of Psychological and Brain Sciences
Indiana University
Bloomington, IN 47405
Telephone: (812) 322-6947
Fax: (812) 855-4691
Email: gchadder@indiana.edu

S1 Examples of Measurement Using the Scale

S1.1 Example 1: A Pac-man Ghost

For our first example, we revisit the case in Section 2.3.2.2 of a ghost computer adversary in Pac-man. First, we assume it's inanimate, i.e., at Level 0.0. Then we proceed to Level 0.1 and ask: "Does the system move or act on its own, i.e., without obvious prompting by external forces?" The answer is clearly Yes since all of the ghosts can be observed moving on the screen, so a ghost is at least a schizoid automaton. So we proceed to Level 0.2 and ask: "Is the system's spontaneous behavior modified by events/conditions in the environment?" As we move Pac-man around, the ghosts follow him, so the answer is Yes. Therefore, we proceed to Level 1.0 and ask: "Does the system appear to be trying to approach or avoid any object or occurrence of an event in its environment?" The ghosts are certainly approaching Pac-man so the answer is Yes. Therefore, we move to Level 1.1 and ask: "Does the system have different sets of goals active during different environmental or bodily conditions?" If Pac-man eats a power-up pellet, the ghosts will change from approaching him to running away, so the answer is again Yes. Therefore, we go to Level 2.0 and ask: "Does the system develop new adaptive approach or avoidance patterns over time?" If we observe the ghosts long enough, we will notice that they essentially only have two behaviors and these are inflexible, predictable, and unchanging. Therefore the answer to this question is negative, which means that a Pac-man ghost is only a Level 1.1 system: a modal value-driven automaton.

S1.2 Example 2: A Chimpanzee

We now repeat the evaluation process for an entity with, arguably, a much higher degree of volition: a chimpanzee. Beginning with an (admittedly absurd) assumption of Level 0.0, we move to Level 0.1 and ask: "Does the system move or act on its own, i.e., without obvious

prompting by external forces?" Anyone who's been to a zoo could answer Yes, so we move to Level 0.2 and ask: "Is the system's spontaneous behavior modified by events/conditions in the environment?" Watching chimpanzees playing in their enclosure quickly leads to a Yes conclusion, so we move to Level 1.0 and ask: "Does the system appear to be trying to approach or avoid any object or occurrence of an event in its environment?" Chimps will approach food or other chimps and will sometimes avoid each other, so the answer is Yes. Therefore, we go to Level 1.1 and ask: "Does the system have different sets of goals active during different environmental or bodily conditions?" As chimps may forage, mate, fight, or do any number of complex behaviors that depend on their internal state and their environment, the answer is clearly Yes, so we move to Level 2.0 and ask: "Does the system develop new adaptive approach or avoidance patterns over time?" As shown by Wolfgang Köhler in the 1910s, chimps can learn to adopt such strategies as piling boxes under a banana tree and climbing on them to reach the fruit (Matsuzawa, 2002), so the answer is Yes. We now move to Level 2.1 and ask: "Can the system engage in a task that requires working memory (e.g. delayed non-match-to-sample)?" Chimps are capable of performing nearly perfectly on a visual delayed-match-to-sample task (Hashiya & Kojima, 2001) so the answer is Yes. We now move to Level 2.2 and ask: "Can the system engage in a task that requires long-term memory?" Chimps in the forests of Bossou forage for figs which means, they need to be able to remember the location of fig trees, the time of year that the fruits are ripe, the fact that red fruit is ripe whereas green fruit is unripe, and the best climbing routes to reach fruit in the highest trees (Matsuzawa, 2002). These memories would seem to require some form of long-term storage, so we answer Yes. Now we move to Level 2.3 and ask: "Can the system engage in a behavior (e.g. game-playing, navigation) that requires evaluation of multiple possibilities without action?" The behavioral pause that Köhler observed

before his chimps engaged in the box-stacking, fruit-reaching behaviors, suggests that some kind of speculative deliberation was happening. Also, it was observed in 1970 that a chimp (named Julia) was able to choose the first time the correct key for opening boxes with other keys that opened other boxes finally leading to a box being opened with a food reward (Suddendorf & Corballis, 1997). This suggests that chimps can engage in the kind of Popperian reasoning we associate with deliberation, at least in a short time-frame, so we answer Yes. Now, we can move to Level 3.0 and ask: "Does the organism send and selectively respond to social cues?" The answer is an obvious Yes, so we go to Level 3.1 and ask: "Can the system pick up and move around objects in its environment?" Under the right conditions, the answer is observably Yes, such as when chimps use sticks to forage for termites (Dennett, 1996). Now, we move to Level 3.2 and ask: "Does the system communicate using language that has syntax as well as semantics?" Although chimps such as Kanzi may be trained to use simple symbols structures, in the wild the answer generally appears to be No (Deacon, 1997), so we say that a chimp is a Level 3.1 system, i.e. a manipulative organism, though it's not clear that chimps couldn't be trained to achieve Level 3.2.

S2 Some Objections and Issues

There are a number of potential objections and issues to be addressed regarding the proposed scale and the methodology for using it. Some of the most salient of these follow.

S2.1 Volition and Consciousness

Usually, when we think of something being a volitional act, it means the act was voluntary. *That means that we consciously willed it. But many of these levels of so-called volition could be instantiated by "zombie" systems that aren't aware of anything*. Little has been said thus far about consciousness and its relationship to volition. Degree of awareness was

not built into either the definition of volition or the scale that has been proposed to measure it. One reason is that awareness is notoriously difficult to measure, except perhaps in humans who can be verbally instructed to self-report in natural language.

One of the most contentious arguments in philosophy of mind and AI has been over whether artificial systems could ever possess consciousness. Some philosophers such as John Searle are skeptical (Searle, 1980), while others (e.g. Churchland, Clark, Dennett) are more hopeful. Assuming a functionalist view of consciousness, as this paper does, however, allows one to take for granted that consciousness is a viable engineering problem.

Assuming functionalism to be true, both volition and consciousness are emergent properties of physical systems, and, in the view of this paper, both are best explained by graded scales rather than hard-threshold definitions. The lowest and (to the skeptic) least satisfying levels of volition may be possessed by creatures that are not self-aware as humans are and may even lack experience of qualia (though this is more debatable). However, both volition and consciousness increase as the information processing systems in an organism grow more complex.

Psychologist Daniel Wegner discusses the likely relation between volition and consciousness in humans in *The Illusion of Conscious Will* (Wegner, 2002). His central thesis is that "conscious will", i.e., the awareness of acting that we think causes our acts actually *follows* our acting and decision processes which can be regarded as unconscious. This view is supported by evidence from studies of neuroscience (Libet, 1999) and of cases where a sense of having willed an act are dissociated from the actual behavior of the person (e.g. "spirit possession", hypnotism, alien-hand syndrome, phantom limb movement of an amputee's absent limb). Consciousness of will may be an after-the-fact interpretation of what we actually are set to do!

However, even if this is so, the interpretation of the act is probably more than an epiphenomenon, as some would claim. Libet's own interpretation of his results (Libet, 1999) is that the awareness of the behavior that follows the EEG-measured readiness potential and precedes the actual motor act may allow the motor act to be vetoed before execution (an idea the neurologist Ramachandran has dubbed "free won't" (Dennett, 2003)). It seems likely that conscious awareness may also be used to train the organism's future acts by, for example, reinforcing behaviors according to the emotional valence evoked by their execution and its consequences ("pride" or "guilt"). One thing this suggests is that consciousness is probably an integral part of the higher-level volitional processes such as deliberation. Automatic, well-learned, well-adapted responses do not require anything more than "zombie" processing, but when things go wrong or an organism is in a novel situation, then consciousness becomes important (Minsky, 1986).

In conclusion, lower levels of volitional processing may not be accompanied by consciousness (as most people understand it), but for higher levels—particularly those above Level 2.0—phenomenal consciousness may be an integral part of the decision process. If we are able to design a deliberative organism, for example, it will probably have a level of consciousness that is commensurate with that of non-primate mammals. As things stand, this would be notable progress for AI.

S2.2 Panpsychism

Humans and apes certainly seem to have volition, and probably other mammals. It's not so clear for fish and amphibians, and it is doubtful for insects. Clearly, however, thermostats, and at the other extreme, corporations, don't have any independent will. They are not conscious. They aren't even organisms! The approach taken in this paper does force us to at

least reexamine our skeptical intuitions regarding the possibility of mind in non-animals and organizations—an earlier draft of the paper proposed collectives/societies of organisms as Level 4 entities—much in the same way the idea of extended mind forces us to consider the possible extension of mind into the environment. Acceptance of a functionalist position on materialism has the corollary that if inanimate matter or collections of individuals are organized in the right fashion, there should be consciousness, volition, and other mental attributes that accrue. This could be said to be a kind of "soft panpsychist" view of the universe, i.e., not everything in the universe necessarily possesses a mind, but (physical) systems that are sufficiently complex and organized so as to have an information processing structure analogous with that underlying animal consciousness, must be considered aware. We do not know what it's like to be a thermostat, or even (surprisingly, perhaps) a corporation or a crowd because each of our awarenesses is attached to a particular network of neurons and associated body. Even anything we know about the awareness of other human minds is inferred. Yet, it may be epistemologically useful to take the intentional stance regarding corporations and even thermostats, much as we often do with genes and memes (Dawkins, 1976). It may even be possible that, counter to our intuitions, that thermostats and societies *really are* conscious, willful, etc., in that they are at least dimly aware and qualia-experiencing beings. As we come to better understand the architecture of mind, our intuitions may (or may not) change regarding what kinds of entities have minds.

S2.3 Linear, Ordinal Scale Issues

Something seems rather arbitrary about a unidimensional linear scale for volition. It seems like it would be better to have multiple dimensions and a measure for each of these. This may be one of the most legitimate objections to the adoption of such a scale as proposed in this

paper. However, if we are willing to give up strong ontological claims about the particular analysis that has been chosen in this paper for volition, the scale may still be a valuable epistemological and methodological tool. (A Dennettian move of saying that the scale formalisms are, ontologically speaking, at least "real patterns" seems feasible (Dennett, 1991).)

There are any number of ways that the complex phenomenon of volition could be analyzed, much as there are any number of ways one might analyze the physical organization of a bacterium, e.g. a molecular vs. a cellular vs. an evolutionary or teleological analysis. In a way similar to Dennett's *Kinds of Minds* scale (Dennett, 1996), this scale attempts to roughly follow what seems to have been evolution's course in the animal kingdom from simple invertebrate organisms to symbol-using primates. At risk of anthropocentrism, the following hierarchy was assumed: insects, fish, amphibians, reptiles, non-social mammals (e.g. tigers), social, but non-manipulating mammals (e.g. horses, dogs), non-human primates, pre-literate man, and literate man. Future ethological findings may suggest reconsiderations of this assumed hierarchy, but it seems like a plausible foundation on which to build the proposed volitional hierarchy.

Even given this animal hierarchy, however, some simplifications were made for functional tidiness. Many of the volitional functional features probably evolved into existence together rather than consecutively in the ordering given. Reinforcement learning, as noted, evolved with the biological innovation of neurons. Short-term and long-term memory probably developed in parallel, as well as the attentional mechanisms that are needed to control them for deliberation. Social and manipulative scaffolding probably evolved in parallel. Many insects, of course, are both social, and organized in societies.

While all of this may suggest that a unidimensional scale for volition may not be the most accurate detailed conceptual model of the phenomenon of volition, such an approach has many

epistemological and methodological advantages for the study of volitional behavior in animals and artificial systems. It provides a comprehensive list of architectural features needed for engineering of human volition; even if we disagree on the ordering given for inclusion of the features, the list is useful. This list and its ordering suggest a plan of attack for the analysis and engineering of volitional systems. The breakdown of volition into components encourages researchers to pinpoint corresponding mechanisms in the brain and suggests to AI researchers an ordering in which to build layers of volition with simpler layers being developed and tested before the more complex layers. (For example, we might conclude that we should develop deliberative organisms before we attempt to build symbolic organisms, i.e., systems that are capable of true natural language understanding.) Finally, the scale may be a useful evaluative tool both for natural and for artificial systems. It may provide a set of empirically defined benchmarks for AI for evaluating its progress in developing intelligent systems. These benchmarks, it is hoped, connect in a more satisfying way both to formal-philosophical and to commonsense ideas of volition, than the Turing test connects to concepts of intelligence.

S2.4 Volition and Mechanistic Causality

The kind of machine (or natural) volition the scale proposes doesn't sound much like true *volition because there is a questionable underlying assumption that you can have truly volitional acts from systems governed entirely by mechanistic principles*. This is an expected argument from (some) philosophers who take an *incompatibilist* stance, i.e., who don't believe free will is compatible with determinism. The intuition is that if a behavior is deterministically caused, it can't be free, and, therefore wasn't *truly* willed. Even an indeterminist, though, might ask whether randomness will make a system's choices any freer in a sense we would care about.

For (Kane, 1996), who is an incompatibilist *libertarian*—i.e., someone who thinks determinism is false and free will exists—the critical issue is whether the agent has ultimate responsibility for at least some of its acts, and he takes the position that indeterministic causation within the choice processes of the organism would allow this, whereas a deterministic process would not. Dennett disagrees and argues that deterministic unpredictability in the decision processes would give the same results in terms of both behavior and subjective experience (Dennett, 2003).

Whether determinism or indeterminism holds, however, it makes sense to say that the acts of agents are caused by *something* and agents are not "prime movers unmoved". To say that volitional behavior can't be explained by causal mechanisms is like saying a computer's behavior can't be explained by the activities of a lot of integrated circuits that are wired together. The main difference between the arguments is that we *know* how interconnected chips give rise to computer functionality because we designed computers, but in the case of humans, natural process has engineered them and we are thrust into the position of trying to reverse-engineer. If our knowledge of computers were forgotten and it were viewed as unethical to open up computers and see how they work or analyze source-code for software, we might imagine an observer of a PC running Windows 98 attributing the behavior of the machine to any number of mysterious acausal forces!

One key argument of compatibilists is that neither acausality nor *absolute, uninfluenced* agent responsibility are necessary for volitional exercise of the will. For those unwilling to concede this point, then the proposed scale might be viewed as a functional specification for causal systems that *simulate* the behaviors of systems that have *true* volition. But if the systems

built according to these principles behave enough like real organisms, it will call into question the utility of making a distinction between simulated and real volition.

S2.5 Non-specification of Mechanisms

An abstract functional scale like this is useful to have as a benchmark maybe, but it doesn't tell us enough about specific mechanisms or algorithms. AI researchers are primarily interested in mechanisms, so indeed the scale is no detailed blueprint for a volitional organism. However, it does provide an "aerial view" of the specifications, and even suggests an order in which investigations might be made into mechanisms and algorithms. It may provide a framework, also, for interpreting the findings in ethology and neuroscience. Figuring out which areas of the human brain are implicated in the different levels of volition and then studying the mechanisms of those areas, may allow neurally-inspired AI researchers to construct more detailed specifications for the implementation of volitional systems.

References

Dawkins, R. (1976). *The Selfish Gene*.New York, NY: Oxford University Press.

Deacon, T. W. (1997). *The Symbolic Species: the co-evolution of language and the brain*.New York, NY: W. W. Norton and Company.

Dennett, D. (1991). Real patterns. *The Journal of Philosophy, 88*(1), 27-51.

Dennett, D. (1996). *Kinds of Minds*.New York, NY: Basic Books.

Dennett, D. (2003). *Freedom Evolves*.New York, NY: Viking.

Hashiya, K., & Kojima, S. (2001). Acquisition of auditory-visual intermodal matching-to-sample by a chimpanzee (Pan troglodytes): comparison with visual-visual intramodal matching. *Animal Cognition, 4*, 231-239.

Kane, R. (1996). *The Significance of Free Will*.Oxford: Oxford University Press.

Libet, B. (1999). Do we have free will? *Journal of Consciousness Studies, 6*(8-9), 47-57.

Matsuzawa, T. (2002). Chimpanzee Ai and her son Ayumu: an episode of education by master-apprenticeship. In M. Bekoff, C. Allen & G. M. Burhgardt (Eds.), *The Cognitive Animal: Empirical and Theoretical Perspectives on Animal Cognition*.Cambridge, MA: MIT Press.

Minsky, M. (1986). *The Society of Mind*.New York, NY: Simon and Schuster.

Searle, J. R. (1980). Minds, brains, and programs. In J. Haugeland (Ed.), *Mind Design II* (pp. 183-204). Cambridge, MA: MIT Press.

Suddendorf, T., & Corballis, M. C. (1997). Mental time travel and the evolution of the human mind. *Genetic, Social, and General Psychology Monographs, 123*(2), 133-167.

Wegner, D. (2002). *The Illusion of Conscious Will*.Cambridge, MA: MIT Press.