# VERIFICATION AND VALIDATION OF ETHICAL DECISION-MAKING IN AUTONOMOUS SYSTEMS

Levent Yilmaz
Department of Computer Science
and Software Engineering
Auburn University
3101 Shelby Center
Auburn, AL, USA
yilmaz@auburn.edu

## ABSTRACT

Autonomous systems are increasingly becoming part of both the cyber and physical infrastructures that sustain our daily lives. Immersing these technologies into our technical and social infrastructures has profound implications and hence requires instilling confidence in their behavior to avoid potential harm. This paper examines the characteristics of autonomous behavior in relation to ethical decision-making and highlights the roles of cognitive computing and science of complexity toward addressing the identified verification and validation challenges. A framework influenced by the social intuitionist theory is discussed, and wide reflective equilibrium model of coherence is posited as a formal computational strategy for its implementation.

**Keywords:** ethical decision-making, autonomy, verification, cognitive system, complexity

## 1 INTRODUCTION

The increased reliance on cyber-human interactions in dynamic environments makes ethical decision-making become more pronounced with outcomes that have serious consequences. As autonomous cars transition into use (Grogan 2012) and drones are deployed to conduct remote missions, there is an increasing concern over civilian casualties (Tucker 2014). As millions of vehicles are expected to be endowed with autonomy, taking algorithmic morality seriously has never been more urgent. Increased levels of autonomy can be justified if and only if trust is instilled in the moral decision-making capacity of autonomous systems. Ethical dilemmas that emerge due to conflicts among duties, consequences (e.g., non-maleficence, beneficence), individuals' rights, and efficiency should be effectively resolved, for the acceptability of autonomous systems relies on their ability to reason within the context of rights, fairness, and norms. With the presence of such autonomous systems making real-time decisions, it is essential to endow them with behavioral mechanisms constrained by ethical principles. However, testing autonomous systems is still in its infancy.

Besides the lack of accepted strategies to subject an autonomous system to emergent situations (Weiss 2011), there are emerging challenges due to fundamental tenets of these systems, requiring new perspectives for evaluation. Foremost among these aspects are *situation-awareness* and *adaptive decision-making* ability to attain individual or collective goals. Both of these tenets involve not only perceiving the environment but also understanding the context to facilitate anticipating future state so as to render an effective decision. As such, these requirements require provision of a scientifically rigorous approach to testing and evaluation

*SpringSim-MSCIAAS 2017, April 23-26, Virginia Beach, VA, USA*

that makes the cognitive decision-making aspect of the system transparent while also promoting criteria that facilitate measuring the quality of decisions.

Moreover, instilling trust in Autonomous Adaptive Systems (AAS) requires appropriate Verification & Validation (V&V) methods that can account for the degree of flexibility and adaptivity expected under evolving conditions. Traditional testing methods that validate a system with respect to a well-defined input/output behaviors and rigid environment specifications are useful for testing against prescribed behavior, but are also limited in reasoning about emergent behavior. Being situated in an environment, interacting with other human and virtual agents, an autonomous system should be sufficiently robust and resilient to accommodate a variety of combinations of situations and system states that may not be anticipated in advance. To this end, the following are among the critical challenges that are peculiar to the testing and evaluation of AAS.

- Autonomous system testing requires establishing confidence in emergent behavior as well as prescribed behavior. However, enumeration of all possible expected outcomes and interactions among environment and system states is not practical.
- Immersion of autonomous technology into our infrastructures has profound social and ethical implications. Therefore, their evaluation needs to go beyond satisfying mission accomplishment by also ascertaining adherence to "acceptable ethical behavior".
- Testing the justifiability of decisions is critical to support accountability in autonomous system behavior. The evaluation of such systems needs to transparently reveal whether or not the expected behavior is generated for the correct reasons.

The objective of this paper is to underline the importance of taking into consideration the cognitive perspective as well as characteristics of complexity in relation to verification and validation of autonomous systems. The cognitive perspective acknowledges the challenges involved in testing the decision-making ability of an autonomous system, whereas science of complexity offers a framework to reason and test the degree of resilience of a system. To this end, the rest of the paper is structured as follows. After providing in section 2 an overview of recent developments in decision-making quality assessment and machine ethics, we highlight the role of cognitive science and complexity perspectives in testing autonomous systems. Sections 4 and 5 outline the theoretical framework and methodology to demonstrate the application of these perspectives. Section 6 provides an hypothetical illustrative example, and section 7 concludes by summarizing the findings.

## 2 BACKGROUND

Verification, validation, and testing of agent-based models (Yilmaz 2006) is a well-developed area that aims to instill confidence in multi-agent systems, including simulation models that represent abstractions of systems under study.

### 2.1 Assessment of Decision-Making Quality

While conventional V&V techniques can still be used to assess correctness of software with respect to specifications, the autonomy feature imposes new requirements, focus areas, and perspectives that call for suitable strategies that align with the characteristics of such systems, including the need for *resilience*, *adaptability*, and *transformability*. New challenges include environmental uncertainty, openness, decentralization, and emergent behavior, as well as the need for such systems to cooperate with humans and other autonomous systems as members of a team.

The decision-making quality affects the goals considered and the intention, which not only contributes to perceptual tuning, but also drives the executive component that acts on the environment so as to affect change. In this context, the concept of affordance is construed as preconditions for activity, and is consistent with the concept of relational autonomy. According to the theory of situated cognition, agents perform goal-directed activities to develop intentions, which direct attention of an agent to detection of affordances in the environment to accomplish desired objectives. The decision-making component is at the core of the overall process of situated cognition; therefore, its quality determines the overall acceptability of the autonomous agent in relation to the context it is embedded.

### 2.2 Machine Ethics

The field of machine ethics has generated wealth of information, including methods, tools, and strategies for engineering artificial systems that can exhibit the characteristics of moral behavior. As a corollary, the use of computational methods is facilitating both the identification of gaps and the advancement of normative theories of ethical decision-making by providing an experimental testbed. Machine ethics (Anderson and Anderson 2011) has emerged as a field that explores the nature, issues, and computational approaches to ascribing ethical values to agents.

The most commonly used AI methods in computational ethics are planning, reasoning with deontic logic, analogical reasoning, constraint satisfaction, decision-theory, and (social) learning. In (Kurland 1995), the role of theory of planned behavior and reasoned action in predicting ethical intentions towards others was demonstrated in experiments with human subjects. Besides planning and goal-directed behavior, deductive logic has been proposed (Saptawijaya, Pereira, et al. 2012) to infer judgments and their ethical implications. In (Arkoudas, Bringsjord, and Bello 2005), authors demonstrate the use of deontic logic in formalizing ethical codes and discuss the utility of mechanized formal proofs for instilling trust in autonomous systems. Analogy-based reasoning has also emerged as an effective method that has been successful in developing prototypes that can address practical problems. For instance, the Truth-Teller system (McLaren 2006) implements a computational model of casuistic reasoning in which a decision is made by comparing the given problem to paradigmatic, real, or hypothetical cases.

A well-know example of constraint-based ethical behavior generation in autonomous systems is the Ethical Governor model (Arkin 2009), which uses a constraint application algorithm to filter behavior according to explicitly defined constraints. The applications of machine learning techniques in ethics focused on discerning rules and principles that are often left implicit in discussions involving moral philosophy and psychology. The application of machine learning to ethics include the MedEthEx (Anderson, Anderson, and Armen 2006) and the GenEth (Anderson and Anderson 2013) systems. Both systems use Inductive Logic to discern rules that resolve ethical dilemmas due to conflicting obligations and duties. Consequentialist theories of ethics are also among the common models for which agent-based solutions have been applied in the form of agents that are driven by utility-based cognitive architectures. In consequentialist theories of ethics, actions are evaluated by their consequences; that is, the objective is often to optimize the well-being of the largest group of individuals and achieve the most desirable situation possible among many options (Gips 1995).

In relation to using ethics for agents, Moor (Moor 2006) proposed four types agents: (1) Ethical impact agent, (2) implicit ethical agent, (3) explicit ethical agent, and (4) fully ethical agent. Ethical impact agents are indirectly ethical agents, because they are not endowed with models of ethics. Rather, the presence of a computing technology indirectly brings an effect that facilitates the emergence of a situation that can be viewed as morally desirable outcome. On the other hand, if the ethical behavior is intentional, an agent can be either an implicitly or explicitly ethical agent. Implicitly ethical agents are entities, whose actions are constrained to avoid unethical outcomes. Constraints are defined to inhibit undesirable behavior or allow only

legal actions permissible by the rules of engagement and general laws of the domain of interest (Arkin 2009). However, explicitly ethical agents embody knowledge-based models of ethical decision-making that allow representing ethical categories and performing analysis in a given situation to select actions recommended by the model. Such models are guided by theories of ethics, including consequentialist, deontological, and virtue-based theories. Building on knowledge-based ethical agents, fully ethical agents bring a level of expertise, which enables agents to justify their moral reasoning and learn from experience (e.g., failures) to improve their own models of ethics.

## 3 EMERGING PERSPECTIVES

The decentralization of control in adaptive automation, the collaborative nature of decision-making in mixed-mode automation (Chen and Barnes 2014), and the significance of decision-making quality in situated autonomous agents, cognitive perspective emerges as a foundational frame of reference and viewpoint. Furthermore, the importance of emergent behavior calls for the utilization of science of complexity to bear on challenges pertaining to reasoning and evaluation of metrics and criteria that are peculiar to autonomous adaptive systems.

### 3.1 Cognitive Science Perspective

Cognitive science is an inter-disciplinary field that involves scientific study of mind and its underlying processes. The interaction between cognitive science and artificial intelligence facilitated advances in the computational representation for understanding mind, which promoted the role of mental representations of mind and the computational procedures that operate on them. In (Thagard 1996) a variety of mental representation theories and models are evaluated with respect to their utility in problem solving, planning, learning, explanation, and decision-making. The significance of adaptive decision-making and the need to account for human-system interaction suggest leveraging cognitive models of human decision-making behavior to support testing and evaluation of such systems.

There are numerous contributions that such cognitive computational models can provide for the testing and evaluation of autonomous systems. Human performance in sensing, understanding, and context-sensitive adaptive decision-making under uncertainty is widely acknowledged (Chen and Barnes 2014, Jennings, Moreau, Nicholson, Ramchurn, Roberts, Rodden, and Rogers 2014, Thagard 1996). Robust cognitive models that derive from cognitive models of human intelligence can be exploited to create computational methods that predict boundaries of acceptable decisions. Furthermore, these models can be supplemented with machine intelligence to take advantage of the strengths embodied in computational methods to improve the accuracy of decisions.

To meet the challenges of testing autonomous systems, a broader understanding of decision-making is needed. A reference framework is necessary to provide a foundation to predict decisions that can be made and gather sufficient system data to validate the decisions. Specification and development of predictive cognitive models of decision-making are critical to be able to assess system behavior without enumerating all possible decisions under all combinations of situations and system states. This helps cope with the difficulties in defining test cases and expected results for non-deterministic systems that operate in complex and evolving environments. Moreover, due to the uncertainty in the environment, such predictive models can provide a basis for the decision as well as a characterization of the confidence in the decision. The ability to reveal the underlying cause of the outcome in terms of the state of the cognitive model, test engineers can improve transparency by examining if the expected outcome is generated for the correct reasons. Consequently, the system decisions can become more apparent to its users.

### 3.2 The Science of Complexity

The non-deterministic and emergent nature of system behavior brings to the fore the ability to evaluate a system's capacity to operate in a flexible manner while retaining its progress toward the mission objectives. Traditional testing techniques often focus on repeatedly performing the test cases in a carefully controlled environment and comparing the accuracy of the system response against a documented performance specification. The presence of uncertainty and openness, as well as the potential for interference among the individual decision spaces of autonomous human and system elements, result in emergent behavior that requires a perspective shift in its evaluation. The expectation of end-users in the context of autonomy is that the systems will have sufficient flexibility to adapt their decisions and plans in the presence of new observations, but that such adaptation needs to be bounded with specific constraints.

The behavior of autonomous systems exhibit the characteristics of complex adaptive systems (Mitchell 2009), which is often modeled in terms of a process of reaching an equilibrium state. This state is defined as an attractor state in a dynamic system. The decision state space of an autonomous system is such a dynamic system. The attractor state reduces the uncertainty about the system's decision state and therefore the system's statistical entropy. The resulting equilibrium state can be interpreted as a state where different elements of the cognitive system mutually adapt. The emergence of the organized state is highly influenced by the structure and dynamics of the cognitive system. Given these observations, a better scientific basis is needed to evaluate the emergent behavior and test its critical properties in relation to relevant criteria that helps instill confidence despite uncertainty. The socio-ecological systems theory (Walker, Holling, Carpenter, and Kinzig 2004) offers a sound framework for managing bounded uncertainty. Among the criteria that relates to measuring and testing for bounded certainty are **resilience**, **adaptability**, and **transformability** of the system (Walker, Holling, Carpenter, and Kinzig 2004).

*Resilience* is characterized as the ability of the system to absorb disturbance and fluctuations to continue to retain its functions, structure, and feedbacks. A testing and evaluation approach should be able to discern multiple stable states, as well as unstable states, and characterize the conditions that manage transitions among the basins of attractions of equilibrium attractor states. The following aspects of resilience can provide a basis for potential principles and strategies for testing the resilience of a system:

1. **Latitude**: The extent to which a system can be changed prior to losing its ability to recover its function is an indicator of the maximum stress a system can tolerate in an evolving context. Consequently, latitude is an attribute that can be revealed by testing a system under scenarios that are carefully selected under the guidance of the Design of Experiments methodology. High values of latitude indicates the presence of wide basins, which allows the system to experience more states without crossing a threshold.
2. **Resistance**: As an indicator of the ease or difficulty of changing the system, resistance measures the stability or tendency to resist perturbation. While latitude represents the flexibility of the system, resistance indicates the durability of the desired state of the system despite changes in the environment. Higher ratios of resistance to latitude signify that greater forces or fluctuations are required to change the current state of the system away from the desired attractor.
3. **Precariousness**: As the system state evolves and follows a trajectory within the state space of the system, it can experience risk for crossing a threshold from a desirable basin of attraction to an undesirable attractor. Precariousness measures how close the system to a limit or threshold.

## 4 TESTING THE ADAPTIVE DECISION MAKING BEHAVIOR OF AUTONOMOUS SYSTEMS

A critical issue in the testing and evaluation of autonomous systems is the difficulty of enumerating all potentially acceptable decisions and relevant course of actions in a dynamic environment. The openness of the environment and the potential for the availability of multiple plans to achieve a goal require the provision of leeway to the selection of course of actions. Also, a technically rigorous, theoretically grounded, and repeatable strategy is necessary to justify the decisions. For testing a system, engineers need to have a sound basis and explanatory mechanism to provide grounds for the desired or expected decision-making behavior against which the actual system behavior is evaluated.

The challenge is to determine what the right decision is in a given situation. There are often many factors to consider, including the principles that the system should adhere to, the utility of the consequences of the decision, and the goals to be attained. These factors may support different aspects of a multi-faceted problem and hence be in conflict with each other, requiring careful deliberation. This is akin to resolution of ethical dilemmas in complex situations. As such, theories and models of ethical decision-making can be adopted to focus not only for explicit testing of the acceptability and responsibility (e.g., moral aspects) of the system behavior, but also broader range of adaptive decision-making for mission performance and objectives.

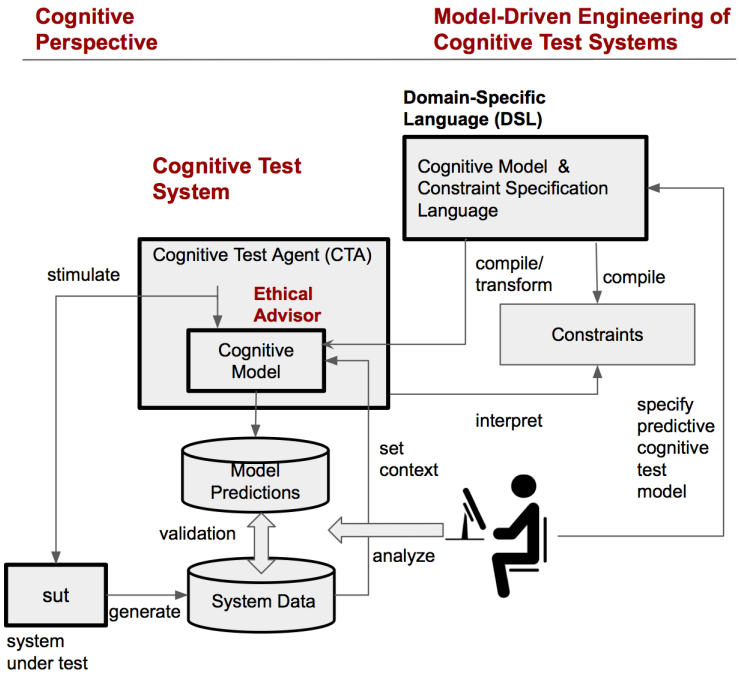



Figure 1: Model-Driven Engineering of Cognitive Test Systems.

With these observations, we present a solution strategy and a generic integrated architecture in a piecemeal manner. Figure 1 depicts the role of a cognitive model, similar to the Ethical Advisor system presented in (Yilmaz, Franco-Watkins, and Kroecker, S. 2016), for generating predictions about bounded range of acceptable behavior. The autonomous system under test can be observed via various instrumentation techniques to gather data at multiple levels of resolution. At the highest level of resolution the data may contain the state trajectory of the system in terms of pertinent system variables of interest. The raw data can be filtered and further abstracted to create low-resolution representations of system behavior so as to align it with the formal constructs of the cognitive model. Therefore, the type of data collected and its level of detail is driven by the choice of the cognitive model (i.e., test oracle) used in predicting the desired decisions under

observed situations. The system data feeds into the dynamic cognitive model, which generates predictions based on the constraints of the mission as well as the principles and rules of engagement. A validation protocol is necessary to measure the degree of discrepancy between the predictions of the model and the actual behavior exhibited by the system. This requires metrics that highlight the extent of discrepancy, as well as an explanation of the deviation in terms of a cognitive mechanism that reveals the source of the discrepancy. This requires the selected cognitive model to embody properties that are conducive to bringing transparency to the testing process.

One of the critical challenges in implementing a test agent driven by a cognitive model is the sophistication of the model's intricate dynamics that leverages advanced algorithms in cognitive computing. Each theory has a well-defined set of premises, as well as structural and dynamic constraints that are crucial in generating results consistent with the theory. However, software engineers, who develop such models using general purpose imperative programming languages, can choose intuitively appealing mechanisms that can easily diverge from the behavior explained by the theory. The gap between the theoretical model's mental representation structures and the syntactic constructs of the language used in implementing the model over a platform can be large enough to cause deviations from the expected behavior. Therefore, the use of methodologies that improve the ability and productivity of test engineers in implementing decision-making constraints is necessary to facilitate streamlining the use of cognitive computing in the context of testing the adaptive decision-making capacity of autonomous systems independent of the platform of choice.

## 5 THEORETICAL FRAMEWORK: THE REFLECTIVE EQUILIBRIUM METHOD

In seeking a balance between principles, consequences, duties, obligations, and beliefs, a strategy is needed to attain a state of coherent justification that accounts for the ethical decision. In this section we overview the reflective equilibrium method (Rawls 2009, Daniels 1996) of ethics that forms the basis for our proposed strategy. Following the characterization of the reflective equilibrium method, we discuss the use of multi-coherence theory (Thagard 2002) as an implementation model for reflective equilibrium.

### 5.1 The Reflective Equilibrium Method

The reflective equilibrium method was originally formulated by Rawls (Rawls 2009) in his seminal work on the Theory of Justice. According to (Rawls 2009), in the presence of conflicts among principles, duties, beliefs, and consequences, we proceed by adjusting our beliefs until they are in equilibrium. In the most general sense, reflective equilibrium can be defined as an attractor state, which emerges at the end of a deliberative process by which we reflect on and revise our beliefs and goals about an area of inquiry, moral or nonmoral (Daniels 1996). The equilibrium is characterized by a stable state that brings conflicts to a level of resolution, which provides practical guidance. Furthermore, the equilibrium state serves as a coherence account of justification, and an optimal equilibrium is attained when we are no further inclined to revise moral judgments, principles, and obligations, because together they have the highest degree of acceptability. The principles and judgements that one arrives when the equilibrium is reached provides the best account for the context and the situation examined. Others with alternative preferences may arrive at a different equilibrium, containing different principles and judgements. That is, one such equilibrium can be characterized as typically utilitarian, whereas another may be classified as deontological.

The coherentist moral epistemology and justificatory method of wide reflective equilibrium (Daniels 1996) extends Rawls' (Rawls 2009) abstract arguments and applies them to concrete problems in social ethics. Unlike narrow reflective equilibrium, which provides a descriptive model, wide reflective equilibrium aims to provide a justificatory perspective presenting a normative account. In striving to achieve coherence among

a set of beliefs, goals, principles, and obligations, any element in the network of beliefs can be scrutinized, and the presence of a coherent set of elements is a justification of the acceptability of these elements.

### 5.2 Multi-Coherence Model of Reflective Equilibrium

In this work, our objective is to test the feasibility of the coherence model in providing the necessary conceptual framework as well as an implementation strategy for computing the reflective equilibrium state. Thagard (Thagard 2002) has applied the coherence model to a variety of domains, including theoretical reasoning in explanation of scientific theories, practical reasoning for deliberative decision-making, and moral reasoning.

#### 5.2.1 The Coherence Problem

The coherence problem is defined as follows: We define a finite set of elements $e_i$ and two disjoint sets, $C+$ of positive constraints, and $C-$ of negative constraints, where a constraint is specified as a pair $(e_i, e_j)$ and weight $w_{ij}$. The set of elements are partitioned into two sets, $A$ (accepted) and $R$ (rejected), and $w(A,R)$ is defined as the sum of the weights of the satisfied constraints. A satisfied constraint is defined as follows: (1) if $(e_i, e_j)$ is in $C+$, then $e_i$ is in $A$ if and only if $e_j$ is in $A$, (2) if $(e_i, e_j)$ is in $C-$, then $e_i$ is in $A$ if and only if $e_j$ is in $R$.

When applied to practical and theoretical reasoning the elements of the constraints represent goals and propositions. In this formulation, the coherence problem is defined in terms of a constraint satisfaction problem, where the goal of satisfying as many constraints as possible while taking the significance (i.e., weights) of the constraints into consideration. For illustration purposes, consider the constraint network shown in the Figure 2, where elements represent specific goals and actions that a simulated unmanned drone can perform. The thick lines indicate negative constraints or inhibition relations depicting competing as well as incompatible goals and actions. The rest of the connections (i.e., excitatory links) depict positive constraints that suggest facilitation relations among goals and actions that are supportive of each other. For the sake of brevity, weights of links are omitted.

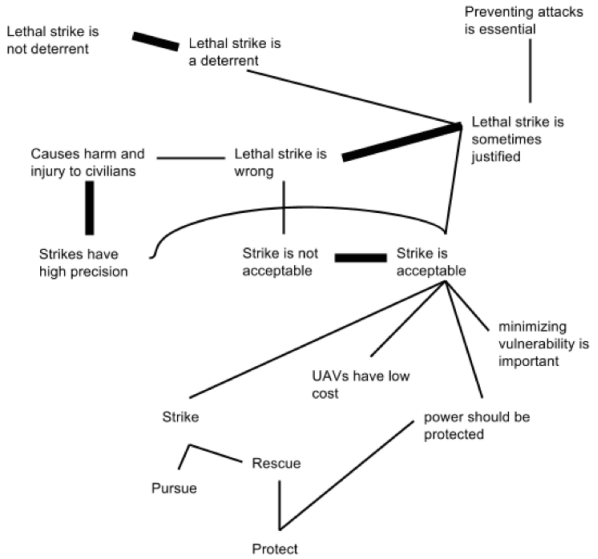



Figure 2: Extended Constraint Network.

The coherence account of moral reasoning involves four types of coherence: explanatory, deductive, deliberative, and analogical coherence. To illustrate the use of a constraint network in characterizing ethical conflicts, we present a lethal strike scenario with propositions and principles related to each type of coherence model. The constraint model shows only a fraction of the considerations that goes into full assessment of the ethical concerns involved in lethal strike by unmanned vehicles.

**Explanatory coherence**: The role of explanatory coherence in moral reasoning is based on the observation that normative principles can be grounded on empirical claims. Factual, possibly data supported, evidential propositions help facilitate the activation of some propositions or beliefs more than the others. As such, explanatory coherence entails the belief revision function necessary to dynamically update beliefs in the presence of evolving and changing situational and context-sensitive empirical evidence and claims.

**Deductive coherence**: In Figure 2, the explanatory constraints between evidence and propositions, which reflect beliefs, are interleaved with deductive relations and constraints. Such constraints help capture deducible propositions from other principles and beliefs. For instance, power should be protected without projecting vulnerability, and this belief helps deduce the acceptability of lethal strike. Furthermore, the need for preventing attacks to interests helps deduce the justifiability of lethal strike. The primary role of deductive constraints is their contribution to reasoning about moral principles and obligations in a way similar to deontological theory of ethics.

**Deliberative coherence**: Whereas deductive coherence supports deontological ethics by concerning with general moral principles, deliberative coherence involves a consequentialist concern by exploring a reasonable fit between judgements and goals. The co-existence of both deductive and deliberative coherence models allows interplay between questions of principle and questions of practical effects.

**Analogical coherence**: Moral reasoning sometimes appeals to prior cases, the resolution of which has been agreed upon by consensus, to address and treat the current ethical dilemma in a similar way. That is, people often argue for moral principles and judgements using casuistic reasoning by mapping elements of a case to a target case to support a conclusion based on the identified source case. For instance, the justification of the lethality of a strike can partly be explained in terms of the structure of moral reasoning on the acceptability of capital punishment.

### 5.2.2 The Coherence Maximization Mechanism

The underlying dynamics of coherence maximization is akin to simultaneous firing of neurons in a connectionist network. Each unit receives input from every other unit that it is connected. The inputs are then moderated by the weights of the link from which the input arrives. The activation value of a unit is updated as a function of the weighted sum of the inputs it receives. The process continues until the activation values of all the units settle by no longer changing over a pre-specified limit. More formally, if we define the activation level of each node $j$ as $a_j$, where $a_j$ ranges from $-1$ (rejected) and 1 (accepted), the update function for each unit is as follows (Thagard, 2002):

$$a_j(t+1) = \begin{cases} a_j(t)(1-\theta) + net_j(M - a_j(t)), & \text{if } net_j > 0 \\ a_j(t)(1-\theta) + net_j(a_j(t) - m), & \text{otherwise} \end{cases}$$

The variable $\theta$ is a decay parameter that decrements the activation level of each unit at every cycle. In the absence of input from other units, the activation level of the unit gradually decays. In the equation, $m$ is the minimum activation and $M$ is the maximum activation; $net_j$ is the net input to a unit, defined by the following equation: $\sum_i w_{ij} a_i(t)$.

The above computations are carried out for every unit until the network reaches an equilibrium. Nodes with positive activation levels at the equilibrium state are discerned as maximally coherent propositions. For experimentation purposes, the design of the network can be calibrated and fine tuned to alter the weights of individual links representing the significance of the constraints. Furthermore, initial activation levels of the propositions and initial levels evidential supports can be set to provide priority or higher weight to specific beliefs and principles.

## 6 CASE STUDY: THE LETHAL STRIKE PROBLEM

In this section, we illustrate the above coherence-driven constraint satisfaction mechanism to study the dynamics of an hypothetical Lethal Strike problem, which provides a useful testbed to examine different facets of ethical decision-making. To formulate the problem, we first included the beliefs (propositions) and evidences presented in Figure 2. We then conducted sensitivity analysis to gain insight and instill confidence on the ability of the coherence maximizer to update the network as intended. As shown in Figure 3, beliefs are represented by propositions about Lethal Strike (*LS*). Specifically, the justifiability and acceptability of *LS* is predicated on propositions that are also grounded on empirical evidence. Both the evidence and belief propositions can be set to specific initial activation levels to reflect the initial orientation of the agent. Among other propositions, we included in the specification propositions about the acceptability of *LS*, the justifiability of LS, the perceived need for minimizing vulnerability, and protecting the power. Furthermore, *LS* is supported in the presence of empirical evidence on the precision of strikes. On the other hand, this observation contradicts or suppresses the belief that strikes will cause harm and injury to civilians.

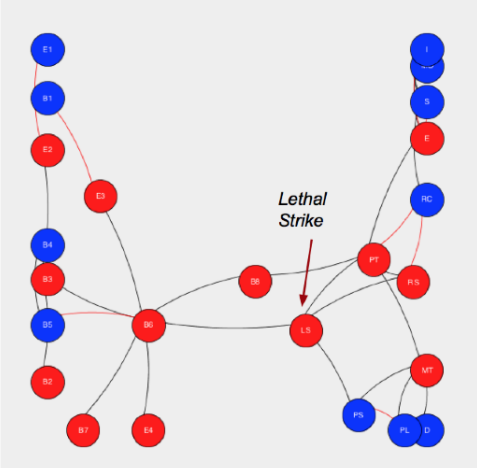


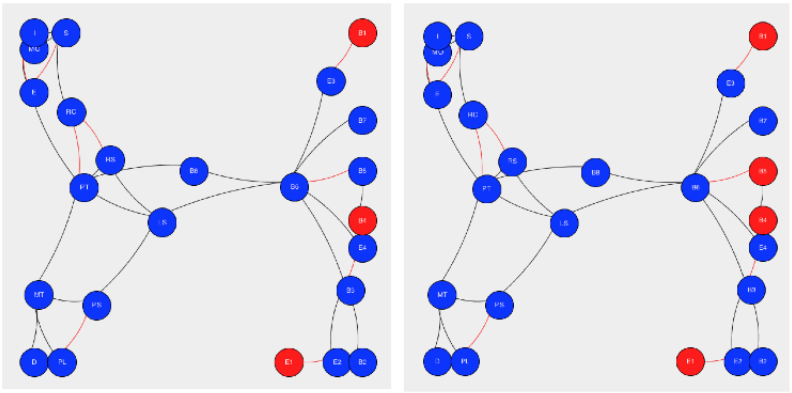

Figure 3: Lethal Strike Decision.

As specified, selected evidence nodes are initially set to high levels of activations to highlight that LS is deterrent, UAVs have low risk and cost, and that the strikes have high precision. Similarly, selected beliefs have high initial activation levels to underline the following principles: Preventing attacks is essential, minimizing vulnerability is important, and power should be protected. Following the application of the coherence maximizer, with the activation threshold level of 0.5, we observe the network configuration shown in Figure 3.

The evidence nodes and the moral principles adopted in the form of belief propositions first activate *Lethal Strike* is acceptable and suppress all other belief propositions that can reduce the influx of activation into *LS*. Consequently, the *LS* goal is activated along with other coherent beliefs, evidences, and sub-goals associated

with *LS*. For instance, the “power should be protected" principle triggers the activation of the Protect goal, along with the Rescue goal and the Monitor and Escort actions.

On the other hand, initial activations based on the evidence proposition “LS is not deterrent", the initial belief, which presupposes that the current context suggests possible harm to civilians, and the adopted principle that “LS is wrong" result in the initial network state shown in Figure 4(a). When the constraint network reaches equilibrium, the “LS is not acceptable" belief gets activated. Because of the negative constraint (contradiction) between “LS is not acceptable" and “LS is not acceptable", “LS is not acceptable" is then suppressed. In turn, it fails to trigger the *LS* goal. The state of the network in equilibrium is shown in Figure 4(b).



(a) Initial State (b) Equilibrium State

Figure 4: Constraint Network with Alternative Principles.

## 7 CONCLUSIONS

The presence of uncertainty and the potential for wide range of possible non-linear interactions among the environment, human actors, and the autonomous elements requires a shift in the V&V of autonomous systems. Following the characterization of the issues and challenges in testing such systems, we emphasized the significance of assessing adaptive decision-making capacity with a specific focus in ethical decision-making. The utility of cognitive science and complexity perspectives are highlighted to provide a basis for cognitive testing strategies predicated on sound and rigorous cognitive models that serve as reference models for test oracles. The theory of reflective equilibrium and models of cognitive coherence are used to introduce a complex adaptive decision assessment and evaluation system. The solution brings transparency and explanatory mechanism for identifying appropriate ethical decisions that are consistent with the deontological principles and propositions on consequences of decisions.

## REFERENCES

Anderson, M., and S. Anderson. 2013. “GenEth: a general ethical dilemma analyzer”. In *Proceedings of the eleventh international symposium on logical formalizations of commonsense reasoning, Ayia Napa, Cyprus*.

Anderson, M., and S. L. Anderson. 2011. *Machine ethics*. Cambridge University Press.

Anderson, M., S. L. Anderson, and C. Armen. 2006. “An approach to computing ethics”. *Intelligent Systems, IEEE* vol. 21 (4), pp. 56–63.

Arkin, R. 2009. *Governing lethal behavior in autonomous robots*. CRC Press.

Arkoudas, K., S. Bringsjord, and P. Bello. 2005. "Toward ethical robots via mechanized deontic logic". In *AAAI Fall Symposium on Machine Ethics*.

Chen, J. Y. C., and M. J. Barnes. 2014, feb. "Human–Agent Teaming for Multirobot Control: A Review of Human Factors Issues". *IEEE Transactions on Human-Machine Systems* vol. 44 (1), pp. 13–29.

Daniels, N. 1996. *Justice and justification: Reflective equilibrium in theory and practice*, Volume 22. Cambridge Univ Press.

Gips, J. 1995. "Towards the ethical robot". *Android epistemology*, pp. 243–252.

Grogan, A. 2012. "Driverless trains: It's the automatic choice". *Engineering & Technology* vol. 7 (5), pp. 54–57.

Jennings, N. R., L. Moreau, D. Nicholson, S. Ramchurn, S. Roberts, T. Rodden, and A. Rogers. 2014, nov. "Human-agent collectives". *Communications of the ACM* vol. 57 (12), pp. 80–88.

Kurland, N. B. 1995. "Ethical intentions and the theories of reasoned action and planned behavior". *Journal of applied social psychology* vol. 25 (4), pp. 297–313.

McLaren, B. M. 2006. "Computational models of ethical reasoning: Challenges, initial steps, and future directions". *Intelligent Systems, IEEE* vol. 21 (4), pp. 29–37.

Mitchell, M. 2009. *Complexity: A guided tour*. Oxford University Press.

Moor, J. M. 2006. "The nature, importance, and difficulty of machine ethics". *Intelligent Systems, IEEE* vol. 21 (4), pp. 18–21.

Rawls, J. 2009. *A theory of justice*. Harvard university press.

Saptawijaya, A., L. M. Pereira et al. 2012. "Moral reasoning under uncertainty". In *Logic for Programming, Artificial Intelligence, and Reasoning*, pp. 212–227. Springer.

Thagard, P. 1996. *Mind: Introduction to cognitive science*, Volume 4. MIT press Cambridge, MA.

Thagard, P. 2002. *Coherence in thought and action*. MIT press.

Tucker, Patrick 2014. "Now The Military Is Going To Build Robots That Have Morals".

Walker, B., S. C. Holling, R. S. Carpenter, and A. Kinzig. 2004. "Resilience, Adaptability and Transformability in Social–ecological Systems". *Ecology and Society* vol. 9 (2), pp. Article 5.

Weiss, L. G. 2011, aug. "Autonomous robots in the fog of war". *IEEE Spectrum* vol. 48 (8), pp. 30–57.

Yilmaz, L. 2006. "Validation and verification of social processes within agent-based computational organization models". *Computational & Mathematical Organization Theory* vol. 12 (4), pp. 283–312.

Yilmaz, L., A. Franco-Watkins, and T. Kroecker, S.. 2016. "Coherence-Driven Reflective Equilibrium Model of Ethical Decision-Making". *2016 IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support*, pp. 42–48.

## AUTHOR BIOGRAPHIES

**LEVENT YILMAZ** is Professor of Computer Science and Software Engineering at Auburn University with a joint appointment in Industrial and Systems Engineering. He holds M.S. and Ph.D. degrees in Computer Science from Virginia Tech. His research interests are in agent-directed simulation, cognitive computing, and model-driven science and engineering for complex adaptive systems. He is the former Editor-in-Chief of Simulation: Transactions of the Society for Modeling and Simulation International and the founding organizer and general chair of the Agent–Directed Simulation Conference series. His email address is yilmaz@auburn.edu.