Feasibility of Turing-Style Tests for Autonomous Aerial Vehicle "Intelligence"

Larry A. Young[*]

Flight Vehicle Research and Technology Division

NASA Ames Research Center, Moffett Field, CA, 94035-1000

A new approach is suggested to define and evaluate key metrics as to autonomous aerial vehicle performance. This approach entails the conceptual definition of a "Turing Test" for UAVs. Such a "UAV Turing test" would be conducted by means of mission simulations and/or tailored flight demonstrations of vehicles under the guidance of their autonomous system software. These autonomous vehicle mission simulations and flight demonstrations would also have to be benchmarked against missions "flown" with pilots/human-operators in the loop. In turn, scoring criteria for such testing could be based upon both quantitative mission success metrics (unique to each mission) and by turning to analog "handling quality" metrics similar to the well-known Cooper-Harper pilot ratings used for manned aircraft. Autonomous aerial vehicles would be considered to have successfully passed this "UAV Turing Test" if the aggregate mission success metrics and handling qualities for the autonomous aerial vehicle matched or exceeded the equivalent metrics for missions conducted with pilots/human-operators in the loop. Alternatively, an independent, knowledgeable observer could provide the "UAV Turing Test" ratings of whether a vehicle is autonomous or "piloted." This observer ideally would – in the more sophisticated mission simulations -- also have the enhanced capability of being able to override the scripted mission scenario and instigate failure modes and change of flight profile/plans. If a majority of mission tasks are rated as "piloted" by the observer, when in reality the vehicle/simulation is fully- or semi- autonomously controlled, then the vehicle/simulation "passes" the "UAV Turing Test." In this regards, this second "UAV Turing Test" approach is more consistent with Turing's original "imitation game" proposal. The overall feasibility, and important considerations and limitations, of such an approach for judging/evaluating autonomous aerial vehicle "intelligence" will be discussed from a theoretical perspective.

Nomenclature

$\mathbf{G}_i$	Cooper-Harper handling quality rating for the i[th] mission task for a generic aircraft model
$\mathbf{H}_i$	Cooper-Harper handling quality rating for the i[th] mission task for the subject aircraft
N_M	Number of mission simulations
N_T	Number of mission tasks being rated
s_A	Mean mission success (from several mission simulations) of an autonomous system guiding and controlling a subject aircraft
s_P	Mean mission success for human operator controlling subject aircraft in same set of mission simulations as autonomous system
$\mathbf{S}$	Array of individual mission success estimates
$\mathbf{T}_i$	Reviewer rating as to "intelligence" guiding the vehicle for i[th] mission task, for subset of tasks performed by autonomous system
$\aleph$	Level of autonomy
ι^*	"Mechanistic approach" intelligence metric, $0 \le \iota^* \le 10$
ι_T^*	UAV Turing test intelligence metric

[*]AHS International Specialists' Meeting on Unmanned Rotorcraft, Chandler, AZ, January 23-25, 2007.

I. Introduction

THERE continues to be ongoing debate as to how to define, measure, and evaluate key metrics as to autonomous aerial vehicle performance. This includes, of course, fundamental questions, and measures, as to aerial vehicle *autonomy* and *intelligence*. This is not a wholly academic question; autonomous aerial vehicles are steadily being introduced and finding great utility in society. References 1, 2, and 3, for example, discuss in considerable detail some of the societal benefits that could be derived from the widespread usage of autonomous aerial vehicles, with special emphasis on the benefits of autonomous vertical lift and/or rotary-wing vehicles. Figure 1 illustrates some of these possible mission or functional capabilities. To a considerable degree, though, the rate of UAV introduction can be considered contingent upon the relative maturity of emerging autonomous system technologies. From an engineering perspective it is difficult to develop a technology wherein fundamental questions as to its optimum functioning is still undecided. From an operational perspective it is difficult to define mission requirements as well as acquire and effectively use a new system if key performance metrics are only nebulously understood.

Does defining and measuring intelligence for embodied (i.e. robotic) intelligent systems, such as autonomous aerial vehicles, have special importance, or consequence, as compared to other establishing metrics for other intelligent systems? The answer is, of course, yes. For embodied intelligent systems, such as UAVs, actions can have dramatic consequences in the real world. UAVs can crash; they can collide in the air or on the ground with other vehicles or objects. They can fail to sense, and appropriately deal with, contingencies and mission uncertainties that a pilot onboard a manned aircraft might otherwise be able to deal with.

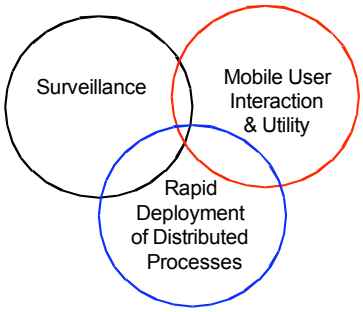

Fig. 1 – "Unmanned Rotorcraft" and Other Autonomous Aerial Vehicle Applications

An example of an unmanned rotorcraft performing a rapid deployment of distributed processes, as per Refs. 3-4, is given in Fig. 2, a forest service "Sentinel" fire spotter small autonomous rotary-wing vehicle. The forest fire tracking application has received wide-spread attention by the UAV research community, e.g. Refs. 5-6.

Fig. 2 – "Sentinel" fire-spotter

But why worry specifically about defining metrics and tests for machine intelligence for unmanned rotorcraft? It is generally recognized that rotorcraft are special -- and unique -- vehicles as compared to conventional fixed-wing aircraft. Mastery of rotorcraft technologies is demanding, challenging, and inherently complex and multidisciplinary; this is especially true for unmanned rotorcraft. One brief example, unmanned rotorcraft may not only have to have software capability for high levels of mission planning but may also have the capacity for controlling high-frequency (n-per-rev) on-blade active rotor controls as a function of some real-time operating condition(s). Many other coupled aeromechanics and autonomous system technology issues may need to be considered in future unmanned rotorcraft design. Control of variable geometry configurations for rotorcraft is yet another example. The design challenge becomes even more significant when a system of systems is being designed. In this case, perhaps, a collective of heterogeneous vehicles could be concurrently designed to cooperatively work together, or, alternatively, an automated base camp might be designed to service and maintain unmanned rotorcraft. A gamut of these and other possibilities, in terms of a multiplicity of intelligent systems and functions – for future unmanned rotorcraft is shown in Fig. 3.

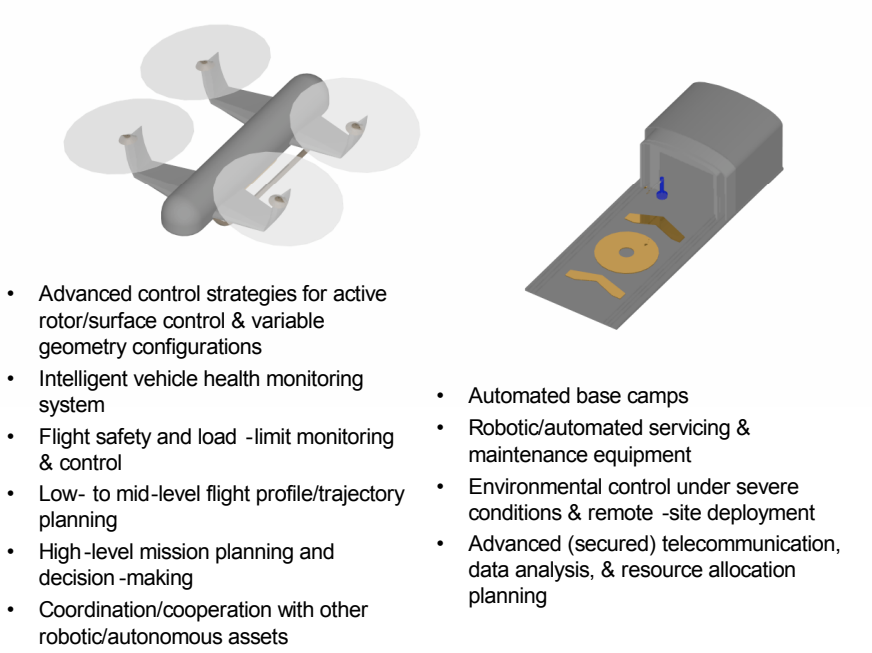

Fig. 3 – Unmanned Rotorcraft and a Potential Multiplicity of Intelligent Systems and Functions

As it is essential for successful vehicle development and operational usage to define metrics for key aspects of autonomous aerial vehicle performance – this, of course, includes, in addition to the familiar rotorcraft design aeromechanics parameters, new metrics for autonomy, intelligence, and others. If you cannot define and measure something, you cannot effectively expend effort to physically realize or improve something. Given this pressing need, why is it, then, so difficult to define and devise metrics for autonomy and intelligence? Two reasons, perhaps. First of all, it is never an easy process to define engineering standards for an emerging technology or a competitive research field of study. Second, *intelligence* continues to be an intangible/indefinable, though obviously innate, quality to understand in humans let alone defining, devising, and measuring it in machines. Fortunately, human beings are quite adept at forging ahead -- despite intangible, even metaphysical, concepts and questions -- pragmatically working around such issues/questions as need be. In this regards, *intelligence* falls within a special class of intangible concepts, or things, that could be collectively known as "I know it when I see it." It is the contention of this paper, that no single machine intelligence metric can be fully successful if it does not in some manner recognize and draw, in part, upon this very human characteristic of the intuitive grasp of the intangible. The trick, of course, is to merge qualitative with quantitative attributes, to arrive at practical measures that can be used to engineer complex systems. This is where the heritage of handling quality requirements comes into the forefront of enabling the definition of intelligence metrics for autonomous aerial vehicles -- more to follow later.

Prior to proposing specific example of machine intelligence metrics, it is crucial to address the question as to what are the minimum general attributes of a good intelligence metric. It is proposed that there are four essential attributes for good machine intelligence metrics: their formulation must be intuitive, their estimates must be generalize-able and predictable (from test-to-test, mission-to-mission, and operational-environment-to-operational-environment), they can be tailored to specific application domains but must be at least broadly applicable within that domain, and they must have a graduated (near-continuous and not discrete) scale. Intelligence metrics must be intuitive in the context that both the intelligent system user/customer and research communities must be able to quickly grasp, positively respond to, and concur with the key conceptual underpinnings of the intelligence metric. Similarly, a good intelligence metric must be both generalize-able and predictable such that a result, stemming from a subset of tests or estimates, must be consistent when applied to a wider and more diverse set of tests, and/or test conditions. The reliability and utility of a metric is greatly diminished if the metric results vary wildly from one test, or estimate, to another -- i.e. it cannot provide a generalize-able result. Further, a metric utility is also diminished if all key governing influences/factors cannot be established and

accounted for in the test/estimation methodology inherent in the metric -- i.e. the metric cannot be considered predictable.

The proposed UAV Turing test's (UTT) greatest strength lies in its intuitive formulation. This is in large part because the test draws upon long established, or heritage, concepts and practices from the handling qualities research community. The proposed UAV Turing test is inherently tailored to the autonomous aerial vehicle application domain, but how broadly applicable can it be defined such that the same resulting metric(s) might be applied to the wide range of aerial vehicles that might be considered UAVs? In other words, can these same metrics be applied to a wide range of UAVs that includes remotely piloted, tele-operated, optionally piloted, semi- and fully-autonomous platforms? For more discussion, as to *autonomy* versus *intelligence* for aerial vehicles, see Refs. 7-8. In particular, Ref. 7 presented a level-of-autonomy scale, ℵ, that is defined in terms of ground-station operator workload; refer to Fig. 4. Other level-of-autonomy scales have been defined in the literature, notably Refs. 9-12. In particular, a level of autonomy scale for spacecraft systems, and planetary aerial vehicles in particular, was defined and expanded upon in Refs. 8 and 13. The key difference in the work of Refs. 7, 8, 13 -- versus perhaps other work in the literature -- is the emphasis on attempting to integrate the defined metrics into aircraft and spacecraft conceptual design and system analysis processes.

The proposed UAV Turing test has to be carefully structured so as to address the question of graduated intelligence metric scaling. The original "imitation game" version of the Turing test is a discrete metric. (Yes/no, is it a machine or a human being that is being interacted with or observed?) Such a simple yes/no discrete metric is inadequate as an intelligence metric for UAVs. This quandary will be discussed further later in the paper.

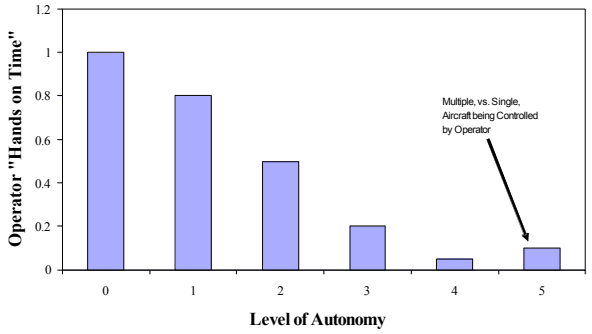

Fig. 4 – Levels of Autonomy in Terms of Ground-Station Operator Workload (from Ref. 7)

II. The Basic Proposal

Perhaps what is required is a "Turing Test" for UAVs. This idea was first suggested in Ref. 7. The Turing Test was first described in Ref. 14 – referred therein as the "imitation game" (perhaps this was a unfortunate label as it seems to compound the continuing debate as to whether the Turing test is a valid measure by which machine intelligence can be judged, e.g., Refs. 15-16). In short, the Turing test for evaluating machine intelligence can be posed as follows: during the course of a blind-test general (non-constrained with respect to subject matter) conversation via teletype, could a machine's response be made to be indistinguishable from a person's? If so, Turing argued, by reason of this inability to distinguish between the human and machine, the machine would have to be successfully judged as capable of human-like "thinking." In this regards, a similar kind of question can be posed as to UAVs. In effect, can a UAV, given its associated (semi or fully) autonomous systems, be made to fly so well, under realistic missions and operating conditions, that it appears to be (indistinguishable from) flown by a pilot or human operator (onboard or remotely piloting the vehicle)?

Compare and contrast this approach to "challenge"-style evaluations of autonomous systems such as the AUVSI (Association for Unmanned Vehicle Systems International) annual international aerial robotics competitions (Ref. 17) or the DARPA "Grand Challenges" (Ref. 18). The weakness of the challenge approach to testing autonomous systems is perhaps the results cannot be generalized. The inherent risk for such autonomy challenges is that they may be too tailored to specific mission types, and the operational environments employed, during the tests.

Finally what is also required to help evaluate UAV autonomous system capabilities is essentially a "Turing Test" (Ref. 14) for autonomous aerial vehicles. Such testing would have to be conducted by means of extensive mission simulations of the vehicle under the guidance of its autonomous system software. Such autonomous vehicle mission simulations would also have to be benchmarked against missions "flown" with pilots/human-operators in the loop. In turn, scoring criteria for such testing could be based upon 1. overall mission success metrics and 2. by "handling quality" metrics similar to the well-known Cooper-Harper pilot ratings, Ref. 19, used for manned aircraft. Autonomous aerial vehicles would be considered to have successfully passed this "UAV Turing Test" if the aggregate mission success and handling qualities for the autonomous aerial vehicle matched or exceeded the

equivalent metrics for missions conducted with pilots/human-operators in the loop.

Table 1. Sample Notional "UAV Turing Test" Checklist

Flight Phases	Mission "X" – Autonomous	Mission "X" – "Piloted"
Take-off (with x% probability – Gaussian distribution – of runway abort)		✓
…		
Navigating and flying waypoint-to-waypoint trajectories within prescribed precision		✓
…		
Landing (with z% probability of final approach abort)	✓	

Alternatively, a third-party knowledgeable observer could provide the "UAV Turing Test" ratings of whether a vehicle is autonomous or "piloted" (i.e. "check" the boxes in Table 1 for each pertinent mission task element). This observer, or reviewer, would also have the additional role of being able to override the scripted mission scenario and instigate failure modes and change of flight profile/plans. If the majority of tasks are rated as "piloted" by the observer, when in reality the vehicle/simulation is fully- or semi-autonomously controlled, then the vehicle/simulation "passes" the "UAV Turing Test." In this regards, this UTT approach is more consistent with Turing's original "imitation game" proposal (Ref. 14).

The advantages of the UTT are: its artificial intelligence (AI) heritage with respect to the classic "imitation game" Turing test; its fundamentally intuitive nature; its heritage with respect to the aeronautics handling qualities community, as to accounting for "pilot rating" of aircraft, aircraft systems (both optimal and "degraded"), mission operations, and operating environments. Because of this aforementioned heritage with respect to the handling qualities community, and by extension the broader aviation/user community, the proposed UAV Turing test has the greater potential for general adoption by that community. It is essential, though, to define a graduate scale for the UTT. A simple yes/no or pass/fail criterion, such as classic Turing test, is inadequate for autonomous aerial vehicle applications. Addressing how to best implement a gradated-scale -- versus a discrete yes/no or pass/fail -- intelligence metric is one of the chief objectives of this paper. It is perhaps this lack of a graduated-scale that is one of the two key disadvantages of the classic "imitation game" Turing test that continues to foster considerable debate – and sometimes acrimony – within the artificial intelligence research community. The other key disadvantage of the classic Turing test is perhaps in the unfortunate choice of the term "imitation game" that Turing used to introduce the concept. The use of the term imitation, to many AI researchers, seems to by definition imply that the Turing test can never hold any validity as a test for machine intelligence. More discussion will follow later in the paper on this debate within the AI community. If, as will be suggested, an acceptable graduated-scale can be devised for the UTT this will no doubt be of considerable enhancement to its utility as an intelligence metric.

III. Why Care about this Issue? The Justification for Measuring/Judging UAV Intelligence

What are the key justifications for developing robust intelligence metrics for UAV? Primarily, the justification fall in the following categories: development, acquisition, and operation. In all three areas technical strides must be made in defining and evaluating autonomy and intelligence metrics.

From a development perspective several issues stand out as to the imperative for defining robust and utilitarian metrics for autonomy and intelligence. For example, autonomy as an emerging design driver -- in particular, in the context of a new technical discipline yet to be incorporated into aerospace multidisciplinary design and optimization and analysis – has been briefly discussed in Ref. 20. Further, the implications of autonomy and intelligence metrics on the system analysis of aerial vehicle and spacecraft system has been discussed in Refs. 2, 7, 11-13; the emphasis of this work being primarily on assessing and prioritizing autonomous technology portfolios for given domain applications. The technology portfolio tools outlined have continued application throughout the development cycle of aerospace systems. Intelligent system metrics are discussed in Ref. 1 in the context as being a crucial element of defining design functional requirements, performing and evaluating conceptual designs, and aiding in the overall conceptualization process.

As mission capability becomes more heavily influenced by aerial vehicle autonomy and intelligence,

then respective user-community acquisition departments will require more demanding criteria to discriminate between autonomous platforms being submitted by offerors. This is especially true for autonomous aerial vehicles. As the number of platforms, vendors, and missions increase, more rigorous autonomy and intelligence metrics, and associated test and evaluation criteria, will become essential. This, in particular, is an important point: autonomy and intelligence metrics have to be testable, ideally in a manner consistent with traditional aerospace test and evaluation practices. Additionally, the upgrade or modernization programs of early generation UAV assets will focus inevitably not only on more capable sensor/payload packages but also on the caliber of "brains" flying the platforms.

The current primary utility for UAVs is surveillance. From an operations perspective, to enable other, more challenging, missions will dictate higher levels of autonomy and intelligence. Further, the user-community for such platforms will need to be presented convincing demonstrations of sustained reliability, mission effectiveness, and system "trustworthiness" before the widespread acceptance of autonomous aerial vehicles for the more challenging missions. Life cycle cost effectiveness for highly autonomous systems will be a major concern/issue for not only acquisition departments but the ultimate users/operators of such systems. There are three key cost-effectiveness assumptions that underlie the current popularity of UAVs: first, UAVs can potentially have a lower per unit cost than a manned aircraft, second, under special circumstances and missions UAVs can be considered more expendable than manned aircraft, and, third, use of UAVs can significantly reduce operational costs. As autonomous aerial vehicle are fielded these assumptions will come under increasing scrutiny as to their validity. Therefore, any and all autonomy and intelligence metrics will have to be of a general utility so as to be ultimately incorporated into life cycle cost estimation methodologies.

IV. The UAV Turing Test and Other Intelligence Metrics

There seems to be three general approaches to estimating, or rather judging, machine intelligence, denoted herein this paper as: *mechanistic*, *emergent*, and *empirical*. The *mechanistic* approach estimates machine intelligence through prescribed functional relationships based on innate parametric characteristics of the intelligent and/or autonomous systems being studied. Such innate parameters include number of sensors or input data provided to the system, number of lines of software defining the systems, etc. Thus the mechanistic approach reasoning goes: the more complex the machine the more capable and therefore the more intelligent (though not necessary computationally efficient or, rather, *elegant*) the system. The mechanistic approach has been examined in some depth, e.g. Refs. 7-8,13. The *emergent* approach seeks to evaluate machine intelligence in terms of initiating or observing complex intelligent system behavior that is *a priori* unpredictable and/or nondeterministic from the initial set of fundamental rules/behaviors instantiated in the system. In some regards this is the "complexity from simplicity" school of thought popular in recent artificial intelligence research. An example of this approach can be found in Ref. 21. Finally, the *empirical* approach seeks to validate or quantify autonomous and/or intelligent system performance in the context of, ideally physical but also simulated, demonstrations and field trials. It is in this later category whereby the concept of robotic or autonomous vehicle competitions or challenges comes into play. The UAV Turing test is but one example of the empirical approach to defining, or otherwise establishing, machine intelligence.

V. "Imitation Game" Version of UAV Turing-Style Tests

In order to continue with the discussion regarding the UTT, it is necessary to discuss some terminology. First, it is necessary to clarify the distinction between "piloted" versus "autonomous" (whether semi- or fully-autonomous) operation of an aerial vehicle. An aerial vehicle can be considered "piloted" in either the case where the pilot, or aircraft operator, is physically onboard the vehicle or is remotely, but directly providing the real-time flight control inputs, operating the aircraft. A couple of examples for clarification are provided. First, an aircraft, with or without passengers, with an operator (either onboard or remotely) providing only high-level -- neither continuous nor real-time input – mission commands and flight guidance should be considered as being at least semi-autonomous. Second, a manned aircraft, carrying passengers but having no operator (onboard or remotely) providing continual real-time flight control inputs, should be considered to be fully autonomous. Therefore, in this context most fielded UAV flying today would be considered to be "piloted," albeit remotely, except for perhaps for a subset of mission tasks. Considerable discussion is devoted in Ref. 7 to the topic of defining UAV autonomy levels.

Aeronautical design standard ADS-33E-PRF, Ref. 22, defines a fundamental subset of mission task elements (MTE) for military rotorcraft. These mission tasks include: hover, landing, slope landing, hovering turn, pirouette, vertical maneuver, depart/abort, lateral reposition, slalom, vertical re-mask, acceleration and deceleration, sidestep, deceleration to dash, transient turn, pull-up/pushover, roll reversal, turn to target, high yo-yo, and low yo-yo. Several of these mission tasks should ideally performed in both good and degraded visual environments, as well as subject to identified system/control failures. Obviously both other and/or additional mission tasks can also be considered in defining a UTT, depending on the type of vehicle and mission being considered.

It is important to rank hierarchically such mission tasks into sub-groups of autonomous system "complexity." This is a refinement of concepts related to mission operational and environmental characterization as were introduced in Refs. 7-8,13. To some degree, the likes of documents such as the U.S. Army's aeronautical design standard ADS-33E-PRF for military rotorcraft handling qualities, Ref. 22, has already accomplished this.

It is recommended that following principles be considered in attempting to perform the notional UAV Turing test: conduct the test to minimize observer/reviewer biases; conduct an adequate pre-test mission screening process to ensure that extraneous or inconsequential mission tasks and operational constraints are not introduced during the testing; take steps to ensure that a test is conducted such that a genuine blind-test is conducted; gather pilot or operator post-mission narrative comments; gather, in addition to the numerical rating data, observer/reviewer post-UTT narrative comments. Human observers/judges are intrinsic to the UTT; psychology and social science studies into observer biases and behaviors needs to be accounted for, or accommodated, in the UTT experiment planning and conduct. This question of observer biases and behaviors is a familiar one in sociological and psychological research; e.g. Ref. 23. Both the pilot/operator and observer/reviewer narrative comments, though not integral to defining UTT-derived intelligence metrics as soon will be seen, are nonetheless vital in refining and improving subsequent UTT exercises. Such narrative comments, in addition to numeric ratings, are very much consistent with the practices of the handling qualities community.

Some of the cues that might be used to differentiate between whether an aerial vehicle is being operated as a "piloted" or "autonomous" vehicle include: slowness of mission task execution; unsteadiness of flight maneuvers; use of two- versus three-dimensional rectilinear versus curvilinear trajectories; discrete or step-like, versus continuous and smooth, incremental attitude or position changes; (lack of) precision of flight maneuvers; severe or abrupt changes in attitude or position; failure to complete flight maneuvers or mission tasks; (poor) situational awareness of hazards, obstacles, and other aircraft flying in close proximity to the evaluated aircraft; manifestation of inadequate, or inappropriate, flight behaviors in response to (pre-flight) unplanned/unanticipated changes in the mission tasks, or scope, and the operational environment.

Three notional test/assessment protocols are now suggested for conducting the UTT. The protocols are listed in an increasing order of complexity/effort. The chief reason for multiple protocols is to attempt to decouple or delineate aircraft characteristics from autonomous system mission execution and decision-making performance.

Protocol # 1 "Observational Only" –

This is the closest UTT analog to the classic Turing "imitation game." A random but comprehensive series of mission tasks are "flown," or rather visually presented, to a group of subject matter expert reviewers (SME, in the case of autonomous aerial vehicles, would be manned-aircraft pilots and UAV operators), acting as "ground-observers," via either simulation or flight tests. An equal percentage of the "presented" mission tasks will in actuality be executed by pilots (either onboard or remotely piloting the aircraft, given the aircraft type/nature) or intelligent systems. Care must be taken in the protocols used to sanitize data and visual images presented to the reviewers so as to not bias the information with non-critical cues. The reviewers will would rate each mission task, $\mathbf{T}_i$, as being performed, in order of perceived "intelligence" guiding the vehicle, as follows: ($\mathbf{T}_i = 1$) by a semi-autonomous system controlling the aircraft (whereby some high-level authority or decision-making is exerted by human operators); ($\mathbf{T}_i = 2$) by a fully-autonomous system (whereby not authority or decision-making is exerted by human operators during the course of the mission or task); ($\mathbf{T}_i = 3$) by a junior/inexperienced pilot (either onboard or remotely piloting the aircraft); ($\mathbf{T}_i = 4$) by a senior/experienced pilot (onboard or remotely piloting). This spectrum of reviewer responses may seem overly convoluted but, in fact, allow the SME reviewers to deal in "shades of grey," rather than absolutes, in their responses. (Note, in the above, that it is a (minor) debatable

point whether the rating assignments for the semi- and fully-autonomous systems should be swapped.)

Protocol # 2 "Active Participation" –

To partially decouple aerial vehicle handling qualities from the autonomy and intelligence assessment, this suggested protocol would have reviewers (ideally, though not necessarily, the same, or all of the, reviewers performing the UTT) participate in complementary remotely-piloted handling quality simulations/evaluations of the subject aircraft prior to performing the above described UTT evaluations. The resulting handling quality assessments, i.e. Cooper-Harper ratings, would have two functions. First, it would provide the UTT reviewers their own personal qualitative benchmarking of the difficulty or ease of flying the vehicle. Second, the Cooper-Harper handling quality ratings could be quantitatively incorporated in the final derivation of intelligence metrics from the UTT evaluations.

Protocol #3 "Benchmarked" –

As an additional step towards attempting to decouple the aerial vehicle characteristics from the autonomous system technology evaluation, the following protocol is suggested. In addition to reviewers evaluating the piloted versus autonomous status of a subject aerial vehicle, the reviewers would also perform the same evaluations, for the same mission tasks, against a generic aircraft model maintained as a benchmark model for sustained autonomy and intelligence evaluations.

Inevitably the proposed UAV Turing test is a holistic assessment of the aircraft flight characteristics, the sensor or instrumentation implementation, and the autonomous systems employed. This coupling can be moderated to some degree by adopting the second or third UTT protocols as suggested above. However, this vehicle, sensor, and autonomous-system coupling inherent in the UTT assessment is not intrinsically undesirable. For the foreseeable future the trend will be to acquire complete or integrated aircraft solutions for unmanned rotorcraft and other autonomous aerial vehicles.

The results for the above protocol suggestions can be rolled up into one notional UTT intelligence metric, ι_{T}^{*}, as has been previously suggested. Alternate definitions of metrics, of course, could be proposed based on UTT results; however, the following expression is a reasonable foundation for future study of this issue. Equation 1 summarizes a definition for the proposed metric consistent with the first UTT protocol ("Observational Only") noted in the above. The inherent assumption in Eq. 1, and the first UTT protocol, is that all tasks are equally weighted with respect to difficulty and the need/requirement for intelligence guiding the vehicle through an individual mission task. The suggested second and third protocols attempt to, among other things, take into account differing levels of task difficulty for both for the "guiding intelligence" as well as the intrinsic handling qualities of the vehicle itself.

$$\iota_{\mathrm{T}}^{*}/(1+\aleph) = \frac{a}{4N_T}\sum_{i=1}^{N_T}\mathbf{T}_i \tag{1a}$$

And

$$\tau = \frac{1}{N_T}\sum_{i=1}^{N_T}\mathbf{T}_i \tag{1b}$$

Where the i^{th} mission task receives a mean rating (the average based on the aggregate of all observers or, rather, reviewers) of $\mathbf{T}_i$; N_T is the number of mission tasks being rated; a is a prescribed constant. The parameter $\aleph$ is the stated level of autonomy of the subject autonomous aerial vehicle. It is important to note that, though all mission tasks performed as a part of the UTT are rated and, further, UTT tasks are performed by both pilots, or human operators, and by autonomous systems, only the subset of tasks performed by the autonomous system (as known only by the UTT organizers) are incorporated in the Eq. 1 and, later, Eq. 2 intelligence metric estimates by means of the rating array, $\mathbf{T}$. Finally, note that the mean rating, τ, falls within the range $0 \le \tau \le 4$.

A linear trend is assumed between the mean observer/reviewer UTT ratings, τ, and the UTT intelligence metric, ι_{T}^{*}. This linear trend described by Eq. 1 can be seen in Fig. 5. As can be seen in Fig. 5 and Eq. 1 there is an assumed dependence of aerial vehicle intelligence on the same vehicle's level of autonomy. In other words, the higher the level of autonomy, the greater the vehicle intelligence required in order to successfully conduct missions. One consequence of Eq. 1 is that one might question why if the level of autonomy is zero. In other words, as the

vehicle is requiring the full attention of a human operator to maintain real-time control of the vehicle, how can the aerial vehicle be still considered to exhibit some small modicum of intelligence? The answer lies in the considering aspects of aerial vehicle control that occur outside of operator conscious control or reaction times; this would include such things as stability augmentation systems and high-frequency active rotor or surface controls which arguably provide some semblance of intelligence to an aircraft even if the vehicle trim state, flight path trajectory, and overall mission planning and decision-making are fully under the control of a human operator.

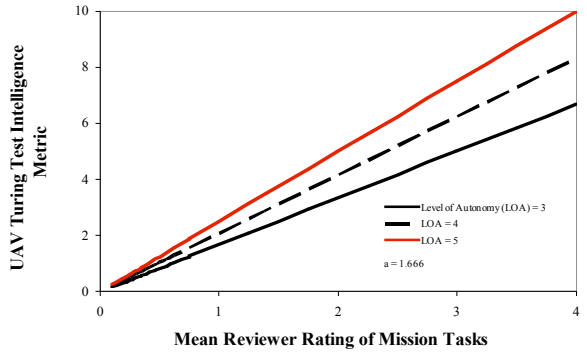

Fig. 5 – First Protocol Functionality

A somewhat more complicated expression, Eq. 2, can be defined for ι_{T}^{*} to be consistent with the second ("Active Participation") and third ("Benchmarked") suggested UTT protocols.

$$\iota_{\mathrm{T}}^{*}/(1+\aleph) = \frac{a}{4N_T}(g/h)\sum_{i=1}^{N_T}\frac{\mathbf{H}_i\mathbf{T}_i}{\mathbf{G}_i}$$

$$g \equiv \frac{1}{N_T}\sum_{j=1}^{N_T}\mathbf{G}_j$$

$$h \equiv \frac{1}{N_T}\sum_{j=1}^{N_T}\mathbf{H}_j \qquad \text{(2a-c)}$$

Where $\mathbf{H}_i$ and $\mathbf{H}_j$ is the Cooper-Harper handling quality rating for the i$^{\text{th}}$ and j$^{\text{th}}$ mission task for the subject aircraft respectively. Further, $\mathbf{G}_i$ and $\mathbf{G}_j$ is the Cooper-Harper handling quality rating for the i$^{\text{th}}$ and j$^{\text{th}}$ mission task for the generic/benchmark aircraft model respectively. The above weighting, in the form of $\mathbf{G}$ and $\mathbf{H}$ terms, implies that greater weight in the intelligence metric assessment is given to tasks that are harder to accomplish for a given mission (as represented by the set of tasks evaluated in the UTT) and a given subject aircraft (versus the generic, or benchmark, aircraft model).

Equation 2 is fully consistent with the suggested UTT protocol #3. Equation 2 devolves into a form compatible with UTT protocol #2 when the Eq. 3 constraint is applied. Finally, Eq. 2 reduces to Eq. 1, the protocol #1 form, given both the constraints noted in Eqs. 3 and 4.

For protocol #2, then the following holds

$$\mathbf{G}_{N_T} = \mathbf{G}_{N_T-1} = \mathbf{G}_{N_T-2} = \ldots = \mathbf{G}_2 = \mathbf{G}_1 = 1 \qquad \text{(3)}$$

For protocol #1, then the following also applies

$$\mathbf{H}_{N_T} = \mathbf{H}_{N_T-1} = \mathbf{H}_{N_T-2} = \ldots = \mathbf{H}_2 = \mathbf{H}_1 = 1 \qquad \text{(4)}$$

Finally, given the above, it is assumed that the intelligence metric derived from the UTT, ι_{T}^{*}, can be related to the alternative (mechanistic approach) intelligence metric, ι^{*} (Refs. 7-8,13), by means of Eq. 5a-b.

$$\iota_{\mathrm{T}}^{*} \propto \iota^{*}$$

Or, to a first order,

$$\iota_{\mathrm{T}}^{*} = b\iota^{*} \qquad \text{(5a-b)}$$

Where b is a prescribed constant such that $0 < b \le 1$. The ability to straightforwardly map one set of metrics against another alternate set of autonomy and intelligence scales/metrics is an important attribute.

VI. Mission Simulation and UAV Turing Tests

The handling qualities community has long recognized the mutual importance, and interdependence, of simulation (with varying levels of modeling fidelity) and flight testing to arrive at satisfactory design solutions for rotorcraft stability and control. Often understated in the design process is the utility and overall importance of "mission simulation" to defining design functional requirements. "Mission simulation" can be considered necessarily distinct from aerial vehicle simulations used to evaluate the detailed or final vehicle designs. Mission simulation can use low-fidelity models for the aerial vehicles as long as high-fidelity modeling is performed as regards the high-level mission tasks and operational and environmental constraints. The goal of the mission simulation is to evaluate the suitability of identified subject systems in expediting, enabling, or improving the performance of the mission. Ultimately the mission simulation is performed to evaluate the magnitude and probability of mission success. Mission success has to be defined in the context of the particular application or mission that is being performed. References 7 and 8 provided several examples of mission success metrics for high altitude long endurance UAVs and planetary aerial vehicles respectively.

Given the inherent power of mission simulation tools, the following 4th protocol () is proposed with respect to a UAV Turing test for defining autonomous aerial vehicle intelligence metrics.

Protocol # 4 ("Fly-Off") –

> Autonomous vehicle mission simulations and flight demonstrations would also have to be benchmarked against missions "flown" with pilots/human-operators in the loop. In turn, scoring criteria for such testing could be based upon both quantitative mission success metrics (unique to each mission) and by turning to analog "handling quality" metrics similar to the well-known Cooper-Harper pilot ratings used for manned aircraft. Autonomous aerial vehicles would be considered to have successfully passed this "UAV Turing Test" if the aggregate mission success metrics and handling qualities for the autonomous aerial vehicle matched or exceeded the equivalent metrics for missions conducted with pilots/human-operators in the loop.

A final conjectural expression, Eq. 6a-c, can be defined for ι_{T}^{*} to be consistent with this fourth protocol. This metric definition embodies the following functional assumptions: (1) the intelligence metric can be assumed to be directly proportional to both the level of autonomy and the relative mission success ratio, i.e. $\iota_{\mathrm{T}}^{*} \propto \aleph$ and $\iota_{\mathrm{T}}^{*} \propto s_A / s_P$, and (2) the intelligence metric asymptotically approaches some constant value as autonomous system enabled mission success approaches some very large (relative) value, i.e. $\iota_{\mathrm{T}}^{*} \to constant$ as $s_A / s_P \to \infty$. Both constraints are accounted for in the definition given in Eq. 6a-c.

$$\iota_{\mathrm{T}}^{*} / (1 + \aleph) = c\left(1 - e^{-d(s_A / s_P)}\right)$$

Where

$$s_A \equiv \frac{1}{N_M} \sum_{k=1}^{N_M} \mathbf{S}_k \Bigg|_{Autononomous}$$

$$s_P \equiv \frac{1}{N_M} \sum_{k=1}^{N_M} \mathbf{S}_k \Bigg|_{Piloted} \qquad \text{(6a-c)}$$

Note that, in the above, that c and d are prescribed constants and $\aleph$ is the vehicle's stated level of autonomy. It should be further noted that adherence, or consistency, with the stated level of autonomy should be checked/validated during the same set of mission simulations that define the mean value of mission success, s_A. If human operator (hands on) interaction with the aerial vehicle is greater than that allowed by the stated level of autonomy, $\aleph$, (refer, for example, to Fig. 4) then either the set of mission simulations would be at least partially invalidated or the level of autonomy would need to be revised downward. Figure 6 illustrates the basic functional properties of Eq. 6a-c.

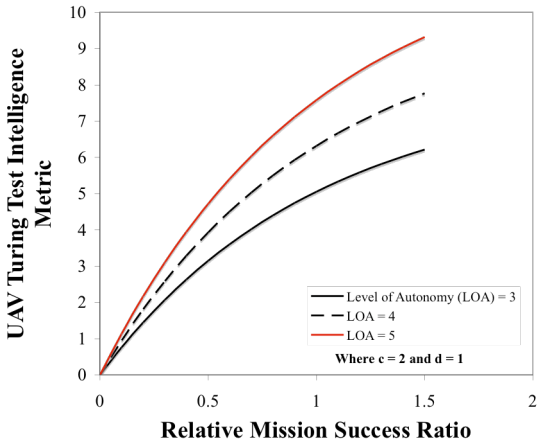

Fig. 6 – Fourth Protocol Functionality

The fourth suggested UTT protocol is perhaps the best choice, of those presented in this paper, for a metric to assess missions, and mission tasks, dominated by higher-level mission complexity and required decision-making as well substantial operational uncertainty. Use of the relative mission success ratio, s_A/s_P, decouples the vehicle aeromechanics characteristics from the autonomous system characteristics. Alternatively, measurement of the absolute values of s_A and s_P can give insight into the suitability of a subject aircraft – irrespective of the intelligent system guiding/operating it – performing a given mission; i.e. if s_A and s_P are both lower than the required target values then the aerial platform itself might not be suitable for the given mission studied. In this regards, the third and fourth protocols are nominally equivalent in the sense that the performance – and hence the intelligence assessment – of the aerial vehicle's autonomous systems is being evaluated and not other potentially extraneous factors. The fourth protocol can potentially find great utility in the early phases of vehicle and autonomous systems development.

VII. Potential Objections to Turing-Style Tests for Intelligent Systems: Practical and Philosophical

The classic Turing test has come under considerable scrutiny in the AI community over the last couple of decades. The strongest objections of "strong" AI (can machines, or rather machines, be made to ultimately to "think," particularly using "symbolic" types of approaches) and the Turing test, in particular, stems from arguments first presented in Ref. 15. This is unfortunate as it potentially reflects a shift from the type of empiricism required for the UTT from within the AI community.

Those AI researchers who debate the question of whether the Turing test truly measures machine intelligence (or embodies the key aspects of "thinking" machines) are missing the point entirely, or indeed whether machines can ultimately be made to think at all, e.g. Ref. 15. As Ref. 16 has pointed out the Turing test represents one of many potential empirical criterion for studying machine performance, for one or more applications, against a (penultimate) benchmark. Some researchers, for example Ref. 24, have gone to so as to argue that other benchmarks may, at least in the interim, be valuable in evaluating the progress of AI efforts; such benchmarks being animals and insects. For micro aerial vehicles, biomimetic, and/or "morphing" aerial vehicles, is it not plausible to, in fact, argue that human operators are not the pertinent benchmarks for the performance (from a flight control perspective) of such vehicles but that benchmark should instead be "lower-order" animals? Be that as it may, we are currently discussing autonomous aerial vehicle performance versus human operators or pilots, not against the flying skills of avian or flying insects.

The above not withstanding, let us stick to the empirical and the pragmatic. Let us leave the (AI) philosophy, as it has been said, to the philosophers. It is wholly appropriate to consider the Turing test as a conceptual model for defining a series of tailored empirical tests and test methodologies, one example being the use of independent observers, to judge the performance, utility, and overall effectiveness of the emerging application domain of autonomous aerial vehicles. Other research considering the feasibility of Turing-style tests for UAVs can be found in Refs. 25-26.

VIII. Additional Considerations

Will a UAV Turing test truly be able to meet the technical and programmatic requirements typical of rotorcraft development efforts as to successfully aid in the safe, timely, and efficient introduction of autonomous aerial vehicles into the national airspace? This is, of course, yet to be determined. In the above discussion a number of attributes for a "good" set of autonomy and intelligence metrics were discussed. Further, any such set of metrics needs to consider not only these attributes for good metrics as well as the goals and objectives, in terms of vehicle development, acquisition, and operations, in employing said metrics. Until then considerable discussion and research will continue as to defining and using autonomy and

intelligence metrics for autonomous aerial vehicles and other "intelligent systems" – see for example Refs. 27-29.

As mission complexity demands greater mission planning sophistication, then the UAV Turing test will become less of a satisfactory standalone intelligence metric. Additional metrics will likely need to be incorporate to assess the satisfactory performance of aerial vehicle autonomous systems embodying high levels of mission planning. Such potential metrics include the mechanistic autonomy and intelligence metrics of Refs. 7-8,13. Alternatively, though, the set of evaluated mission task elements in the UTT can be expanded with tasks of increasing complexity and sophistication as partial compensation and, therefore, extend applicability of the UTT derived metrics.

Though the focus of this paper has been on intelligence metrics for UAVs, it should be readily apparent that similar Turing-style tests could be devised for other intelligent system application domains. At first, envisioning the resulting diverse collection of metrics and test methodologies for different domains may seem counterproductive. Ideally, those who might argue, it would seem far better that a general unifying set of metrics should instead be the goal rather than a tailored set of domain-specific metrics. If indeed such a general unifying set of metrics is ultimately devised then, probably, it should be adopted. But, until then, taking a pragmatic engineering perspective, there is nothing fundamentally wrong with domain-specific intelligence metrics – as long as they address the key questions regarding aerial vehicle functional requirements, as affected by vehicle intelligence, that are/will be demanded by developers, users/customers, and regulatory bodies. (Some of these questions having been discussed earlier in the paper for autonomous aerial vehicles.)

Concluding Remarks

Likely more than one set of intelligence metrics (based perhaps on a combination of mechanistic, emergent, or empirical – of which, the UAV Turing test is one example – estimation approaches) will be applied in assessing the intelligence of autonomous aerial vehicles. This result, a multiplicity of metrics, is not necessarily an undesirable outcome. But, in proposing an intelligence metric that has twin parentage from classic work from both the aeronautics handling quality community, e.g. Cooper-Harper pilot ratings, and the AI research community, the classic "imitation game" Turing test, it is anticipated that the proposed "UAV Turing test" can potentially have broad appeal and utility.

In the end, though, it is not solely a question of what is the optimum set of autonomy and intelligence metrics to apply to unmanned rotorcraft, and autonomous aerial vehicles in general, but rather how will such metrics be effectively employed to arrive at real engineering solutions to development of these aircraft. Therefore, the definition of such metrics must be tailored so as to find this utility throughout the complete aircraft development cycle – including, perhaps most importantly, early incorporation into the system analysis and conceptual and preliminary design stages.

References

[1]Young, L.A., "Aerobots as a Ubiquitous Part of Society," AHS Vertical Lift Aircraft Design Conference, San Francisco, CA, January 18-20, 2006.

[2]Young, L.A., "Future Roles for Autonomous Vertical Lift in Disaster Relief and Emergency Response," AHS International Specialists Meeting on Advanced Rotorcraft Technology and Life Saving Activities, (Heli-Japan 2006), Nagoya, Aichi, Japan, November 15-17, 2006.

[3]Young, L.A., Aiken, E.W., Johnson, J.L., Demblewski, R., Andrews, J., and Klem, J., "New Concepts and Perspectives on Micro-Rotorcraft and Small Autonomous Rotary-Wing Vehicles," AIAA 20th Applied Aerodynamics Conference, St Louis, MO, June 24-27, 2002.

[4]Aiken, E.W., Ormiston, R.A., and Young, L.A., "Future Directions in Rotorcraft Technology at Ames Research Center," 56th Annual Forum of the American Helicopter Society, International, Virginia Beach, VA, May 2-4, 2000.

[5]Casbeer, D.W., et al, "Forest Fire Monitoring with Multiple Small UAVs," Proceedings of the 2005 American Control Conference, June 8-10, 2005.

[6]Freed, M., et al, "Human-Interaction Challenges in UAV-Based Autonomous Surveillance," Proceedings of the 2004 Spring Symposium on Interactions between Humans and Autonomous Systems over Extended Operations, American Association of Artificial Intelligence (AAAI), Publications, 2004.

[7]Young, L.A., Yetter, J.A., and Guynn, M.D., "System Analysis Applied to Autonomy: Application to High-Altitude Long-Endurance Remotely Operated Aircraft," AIAA Infotech@Aerospace Conference, Arlington, VA, September 2005.

[8]Young, L.A., Pisanich, G., and Ippolito, C., "Aerial Explorers," 43rd AIAA Aerospace Sciences Meeting, Reno, NV, January 10-13, 2005.

[9]Clough, B., "Metrics, Schmetrics! How the Heck Do You Determine a UAV's Autonomy Anyway," Proceedings of the Performance Metrics for Intelligent Systems Workshop, Gaithersburg, Maryland, 2002.

[10]Clough, B., "Unmanned Aerial Vehicles: Autonomous Control Challenges, A Researcher's Perspective," Journal of Aerospace Computing, Information, and Communication, Vol.2 No.8, 2005, pp. 327-347.

[11]Proud, R.W., Hart. J.J., and Mrozinski, R.B., "Methods for Determining the Level of Autonomy to Design into a Human Spaceflight Vehicle: A Function Specific Approach," NIST Workshop on Performance Metrics for Intelligent Systems, 2003.

[12]Proud, R.W. and Hart, J.J., "FLOAAT, A Tool for Determining Levels of Autonomy and Automation, Applied to Human-Rated Space Systems," AIAA Infotech@Aerospace Conference, Arlington, VA, September 26-29, 2005.

[13]Young, L.A., "System Analysis Applied to Autonomy: Application to Human-Rated Lunar/Mars Landers," AIAA-2006-7516, AIAA Space 2006, San Jose, CA, September 19-21, 2006.

[14]Turing, A.M., "Computing Machinery and Intelligence," Oxford University Press on behalf of MIND (Journal of Mind Association) – A Quarterly Review of Psychology and Philosophy, Vol. LIX, No. 236, pp. 433-60, 1950.

[15]Searle, J.R., "Minds, Brains, and Programs," Behavioral and Brain Sciences, Vol. 3, No. 3, pp. 417-457, 1980.

[16]Harnad, S., "Minds, Machines and Searle," Journal of Theoretical and Experimental Artificial Intelligence, Vol. 1, pp. 5-25, 1989.

[17]Association for Unmanned Vehicle Systems, International (AUVSI) Web-Site: http://www.auvsi.org/.

[18]DARPA Grand Challenge Web-Site: http://www.darpa.mil/grandchallenge/index.asp

[19]Cooper, G.E. and Harper, Jr., R.P., "The Use of Pilot Rating in the Evaluation of Aircraft Handling Qualities," NASA TN D-5153, April 1969.

[20]Young, L.A. and Aiken, E.W., "Exploration: Past and Future Contributions of the Vertical Lift Community and the Flight Vehicle Research and Technology Division," AIAA 1st Space Exploration Conference, AIAA-2005-2707, Orlando, FL, January 1, 2005.

[21]Brooks, R.R., "Stigmergy – An Intelligence Metric for Emergent Distributed Behaviors," National Institute of Standards and Technology (NIST) Workshop on 'Performance Metrics for Intelligent Systems,' Gaithersburg, MD, August 14-16, 2000.

[22]Anon., "Performance Specification: Handling Qualities Requirements for Military Rotorcraft," Aeronautical Design Standard ADS-33E, December 1999.

[23]Nunnally, J.C. and Bernstein, I., *Psychometric Theory, 3rd Ed.*, McGraw-Hill, 1994.

[24]Franklin, S.P., *Artificial Minds*, MIT Press, Cambridge, MA, 1995.

[25]Duke, E., Vanderpool, C., and Duke, W., "A Turing Test for UAV Operations," U.S. Air Force T&E Days, AIAA 2007-1622, February 13-15, 2007.

[26]Mehra, R.K. and Prasanth, R.K., "Quantification and Measurement of Autonomy for UAVs using Human Operator Modeling," 2nd AIAA "Unmanned Unlimited" – Systems, Technologies, and Operations, AIAA 2003-6605, San Diego, CA, September 15-18, 2003.

[27]Huang, H-M, Pavek, K., Novak, B., Albus, J., and Messina, E., "A Framework for Autonomy Levels for Unmanned Systems (ALFUS)," Proceedings of the AUVSI's Unmanned Systems North America 2005, Baltimore, MD, June 2005.

[28]Huang, Hui-Min, et al, "Autonomy Levels for Unmanned Systems (ALFUS) Framework: An Update," 2005 SPIE Defense and Security Symposium, Orlando, FL.

[29]Huang, Hui-Min, Ed., "Autonomy Levels for Unmanned Systems (ALFUS) Framework, Volume I: Terminology (Version 1.1)," NIST Special Publication 1011, September 2004.