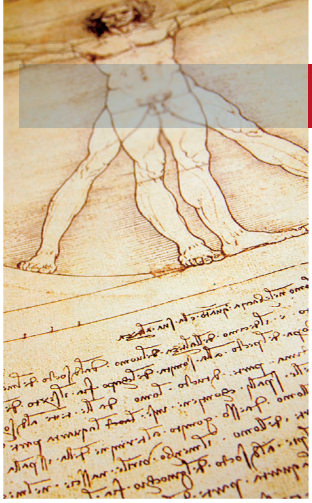

Editors: **Robert R. Hoffman, Kenneth M. Ford, and Matthew Johnson,**
Florida Institute for Human and Machine Cognition, rhoffman@ihmc.us

Bridging the Ethical Gap: From Human Principles to Robot Instructions

Neil McBride, *De Montfort University*
Robert R. Hoffman, *Institute for Human and Machine Cognition*

The idea that robots can act ethically has been debated for more than 70 years,[1] and the theme of robot dominance or rebellion has characterized fiction ever since Karel Capek published *Rossum's Universal Robots* in 1920.[2] Since 1950, when Asimov popularized his three laws,[3] debate has raged as to whether those particular rules would enable robots to make ethical decisions independent of human intervention. The underlying paradigm of Asimov's laws concerned the robot's self-directedness. Operating independently of any human intervention, the robot has the physical and intellectual capability to make its own moral decisions based on its internally held knowledge and rationality. Asimov's three laws were

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey orders given it by human beings except where such orders would conflict with the first law.
3. A robot must protect its own existence as long as such protection does not conflict with the first or second law.

In a previous essay in this department,[4] Robin Murphy and David Woods pointed out that Asimov's first law is outdated and irrelevant, and the second law is obviated by the persistent inability of computer scientists to make computers that can deeply and robustly understand natural language and the gestural meanings that accompany meaningful conversations. The third law can be taken for granted, because there are reasonably well-established methods for ensuring robot survival.

Murphy and Woods said, "Asimov's laws are based on functional morality, which assumes that robots have sufficient agency and cognition to make moral decisions."[4] One of this department's continuing themes has been the myths of autonomy,[5,6] including the myth that robots are self-sufficient. Central to a robot's deployment is its interdependence as part of a human–robot relationship and as part of a team.[7] This interdependence must extend to ethics. Ethical interdependence requires collaboration and communication between the robot and the human to achieve the best outcome where moral decisions or dilemmas may be involved.

Murphy and Woods proposed three alternative rules for "responsible robotics,"[4] which took an interdependence approach and highlighted the roles and responsibilities of robotics engineers:

1. A human may not deploy a robot without the human–robot work system meeting the highest legal and professional standards of safety and ethics.
2. A robot must respond to humans as appropriate for their roles.
3. A robot must be endowed with sufficient situated autonomy to protect its own existence as long as such protection provides smooth transfer of control to other agents consistent with the first and second laws.

To be useful, such high-level principles have to be taken closer to implementable rules. We propose an approach to spanning this ethical gap, which focuses on communication across the

1541-1672/16/$33.00 © 2016 IEEE
Published by the IEEE Computer Society

interface between the human and the robot and on setting the boundaries within which the robot operates.

The Ethical Gap

Humans do not make ethical decisions on the basis of anything like common or formal logic[8]; for ourselves, we do not equate rationality with logic. Emotions, social upbringing and maturity, gut instinct, and philosophical worldviews all affect ethical decision making and the ways people struggle with moral dilemmas. This involves a subtle grappling with meaning, understandings of purpose and morality, and feelings and principles that are hard to articulate and likely impossible to reduce to fixed codes. Such ability is built up by years of parenting, socialization, and involvement in cultures, societies, and communities.

The commonly held idea that fixed ethical decision procedures could be developed has become much less common, partly because the goal has not been achieved and is seen as philosophically less and less achievable. The gap between abstract principles and complex moral situations has become more and more obvious.[9] It is also the case that where rules might apply to simple moral situations, there is no guarantee that the human will be virtuous enough to comply with moral decision procedures.

In moral ethics, different philosophers can come to different, yet equally valid, conclusions about the same problem. Thus, the reduction of ethics to a set of fixed rules is fraught with difficulty and likely to fail. Additionally, for any nontrivial ethical question, there might be no single answer, as well as no conclusive answer. Not only may several possible decisions be "right," but the situation may be such that no decision is right, and, in such tragic dilemmas, any decision will fall short of uniform justice and will result in regret. Or worse.

In any case, even where codification might seem achievable, a set of rules believed sufficient for every contingency and every variation in circumstances and environment would grow to such an extent that even the most advanced processing power would drain out before coming to an acceptable conclusion. And there would be no end, because this is a moving target problem: new and unique ethical situations and dilemmas emerge all the time. Building rule upon rule in the pursuit of the end of the rainbow is distracting enough for humans. If the ethics rule set were a moving target, so would be any robot governed by it. A given circumstance could distract it from pursuing its basic functions and render it paralyzed.

Furthermore, as John von Neumann was aware,[10] the mechanism of human thinking is not the same as that of machine thinking. It is not a case of replacing the if-then-else rules of AI with fuzzy rules, statistics, and probability. The "language of the brain" is not the descriptive language of mathematics. Computers can manipulate symbols, but they cannot (yet) apprehend, because meaning is not a property of symbol strings (or word strings) but requires consciousness and a desire for human connection, which are not characteristics of a machine.

Thus, an immense gap exists between the architecture, implementation, and activity of humans and robots in addressing ethical situations. A robot might be able to make some simple decisions, in well-defined and well-understood environments, based on computational algorithms, but the resources of socialization, belief, and conviction are not available to the robot, which eventually depends on the human as moral arbiter.

Rather than creating rules that can be used to algorithmically constrain robots to engage in ethical behavior, we might develop laws or rules to enable humans and robots to manage and even explore ethics in their interdependent activity. Principles such as those developed by the UK's Engineering and Physical Sciences Research Council (www.epsrc.ac.uk/research/ourportfolio/themes/engineering/activities/principlesofrobotics) tend to address the human processes. Processes such as those suggested by Neil McBride and Bernd Stahl provide the right environment for ensuring that the social and ethical elements of the robots' design, development, and deployment are considered,[11] but they do nothing to address robot behavior in the wild. Similarly, the Murphy-Woods laws of responsible robotics are primarily directed at the engineers, so those laws also reside on the human side of the ethical gap.

The Murphy-Woods Alternative Laws of Responsible Robotics

The first Murphy-Woods law, which refers to professional and legal standards of ethics and safety, chimes with the UK Engineering and Physical Sciences Research Council's principles of robotics. These principles make it clear that a robot is a manufactured artifact that, however autonomous it might seem, is a product of the creator's mind, expressing the purpose, perceptions, and intentions of its creator. Thus, the responsibility for the robot's actions is attributed to the engineering team and does not float off and reside within the robot.

Achieving these high standards would require much more than a few algorithms relating to or assuming ethical interdependence. Legal aspects must be addressed, manufacturing processes considered, and responsibilities defined. It requires the development of standards and testing and the potential of criminal prosecution and economic loss if the standards are not adhered to. There must be laws about the qualification of robot engineers, the organization of enterprises that produce robots, and the characteristics and properties of the robot.

ISO standard 10218-1 for industrial robots applies to the processes and steps in robot design and manufacture. The revised ISO standard relies on the notion of the "man in the loop," recognizing the collaborative and team dimensions of robot system operation. Such standards must be extended to meet the increased complexities of social robotics. The ISO standards require the highest level of documentation and the rigorous incorporation of risk assessment. Such standards are only part of the story. Training of robot engineers must address ethical interdependence.

So, the first Murphy-Woods law can already be implemented: robotics engineers must undergo a professionally certified training before they can practice; robots must be produced within a laboratory that is certified to ISO standards and undergoes regular inspections; and managers and executives of such facilities will be legally responsible for the safe deployment of robots and may face criminal prosecution in the event of failure. But none of this addresses ethical algorithms in the robot, or even whether the robot has algorithms that might be considered ethical.

The second Murphy-Woods law states that a robot must behave as appropriate to its roles. This too is a principle, requiring us to derive rules that might be implementable in a robot design.

Any role is conducted within a context. It involves sets of tasks, responsibilities, and reports. A role is fulfilled in relation to others. Definition of a role will require definition of the relationship. Therefore, the focus is not on what the robot does when self-directed but on what it contributes to the relationship. The role the robot undertakes must be clearly within its capabilities. The role will have a defined degree of freedom on numerous dimensions; the variety of responses will depend on the role's complexity. In essence, the robot is operating a service, and the design of the robot's tasks and actions is the design of a service. The delivery of the service requires that we understand and define the roles enacted by the robot and the human, and that we design the interaction at the robot–human boundary.

The catch-22 is that a human role cannot just be transferred to a machine.[12] The coactivity of humans and machines changes the nature of the task and affects the interaction between the human and the machine. There is a redistributing and sharing of the role and its pertinent tasks, not a replacement or exchange of roles. Role-reliant interactions will create ethical interdependencies and therefore extend the ethical domain. This requires the human to have sufficient knowledge of the capabilities and workings of the robot, to promote realistic expectations of its performance. This is not a blind trust, but an evidence-based trust,[13] cemented by observation of the robot's actual performance, which requires observability, predictability, and reliability on the robot's part. It is in the provision of sufficient knowledge about the role and the robot that trust emerges, expectations are managed, and interdependence is made possible.

For both Asimov and for Murphy and Woods, their third laws refer to the robot's self-protection. There are many conceivable situations in which robots would or should take evasive action of some kind. For example, a robot could be acting as the extended eyes and ears of the human in a dangerous situation. Self-protection is necessary because the robot's capabilities, which exceed those of the human, are exercised in the context of a role relationship in which the protection of the human is paramount. Hence, the robot must also self-protect.

Conversely, there could be situations in which the robot becomes a danger to humans, and at the extreme should self-sacrifice or even self-destruct rather than self-protect. One might ask whether a robot should ever prioritize self-protection. The economic argument comes into play, of course (robots can cost a lot). But again, we run into the moving target problem.

Self-protection by the robot or protection of the human by the robot could involve situations in which the robot disagrees with the human. For example, if an aircraft is being flown into the ground, the aircraft should refuse the pilot's instructions and take evasive action. As analyses of aviation incidents by David Woods and others have shown, control in such situations must be backed up by increased communication. Control remains shared in the relationship, and independence of action is compensated by an increase in interdependent communication. Increasing robot capability demands increased coordination and communication.

Finally, situational ethics are such that a human's well-being might not always take precedence over the goal of accomplishing a task. So, it seems that a blanket rule stating that the robot has to self-protect has little purchase.

Thus, the question remains: How do we cross the ethical gap from human practice and principles to machine practice and algorithms?

Crossing the Ethical Gap

As Murphy and Woods emphasized, the management of ethical interdependence will require a systems approach that recognizes that the engagement between the robot and the human

Table 1. The Un-Fitts' List recast for the robotics context.

Robots	
Are constrained in that...	**And they need humans to...**
sensitivity to context is low and is ontology limited.	keep them aligned to the context.
sensitivity to change is low, and recognition of anomaly is ontology limited.	keep them stable given the variability and change inherent in the world.
adaptability to change is low and is ontology limited.	repair their ontologies.
they are not "aware" of the fact that the model of the world is itself in the world.	keep the model aligned with the world.
Humans	
Are not limited in that...	**Yet they create robots to...**
sensitivity to context is high and is driven by knowledge and attention.	help keep them informed of ongoing events.
sensitivity to change is high and is driven by the recognition of anomaly.	help them align and repair their perceptions because they rely on mediated stimuli.
adaptability to change is high and is goal driven.	affect positive change following situation change.
they are aware of the fact that the model of the world is itself in the world.	implement their preferences for changing the world.

takes place within the context of complex systems, including relationship dynamics that range far beyond the robot–human dyad. The key to crossing the ethical gap lies in how the human and robot engage in their interdependent relationship.

Any attempt to code ethical rules into the robot will be incomplete and inconsistent and will be brittle over time. Robot ethics can only be partial, because the robots' capabilities are limited and changing. Ethical standards or beliefs that are employed in communities and societies act as scaffolding to support the building of the character and the maturity necessary to engage with ethics.

Any framework for crossing the ethical gap will have to be based on the complementary capabilities of the machine and the human, which Robert Hoffman and colleagues captured in the "Un-Fitts' List."[14] The original Fitts' List emerged in post–World War II human factors engineering and emphasized the machine's capabilities and the human's limitations. The Un-Fitts' List emphasizes human–machine interdependence. Here, we recast the Un-Fitts' List for the robotics context.

In perception, the robot can provide images and access that the human cannot achieve. It might access inhospitable areas that the human cannot access, such as planets or the inside of a nuclear reactor. It might be able to detect smells, chemicals, and sounds that are beyond the human. Conversely, the human can detect social cues and have a wider, more intuitive, and more holistic perception of a field of play. In cognition, the robot can calculate faster and access information quicker, whereas the human can process tacit and intuitive knowledge and make decisions in a manner well outside the robot's computational capabilities. Physically, the robot makes up for the human's lack of strength, reaction speed, and disability, but it might not match the human's deep dexterity, suppleness, and physical adaptability.

As the line entries in Table 1 imply, the key to the human–robot interdependence is communication and common ground. The more difficult and uncertain the environment and the tasks, the greater the extent of communication required. The messages the robot and human exchange, and their meaning, will be critical to the ethical interdependency as well as to goal or task achievement. Therefore, regardless of what capabilities can be implemented in the robot, they will be of little value if not encased in good communication.

This requirement to focus on the robot's interactional capabilities is in itself a major challenge in addition to setting our attention on the ethical gap. The rule-based approaches required by computational systems stand in stark contrast to the openness and unpredictability of social interaction.[15] In creating a focus on communication, it is necessary to treat the human–robot interaction as a distinctive and different system in its own right, and not attempt to reproduce human communication nor attempt to reduce the human to a machine supervisor.

The basis of moral practice will require the definition of a work system competence envelope that defines the parameters and limits within which the robot can work. These limits might concern physical perimeters, time thresholds, capabilities, responsibilities, and acceptable tasks. The competence envelope is a multidimensional space characterized by parameters that describe the limited set of tasks and problem situations. If we can define the task and environment, we can apply some limiting parameters to the robot's behavior. For example, we can set physical borders outside which the robot cannot roam,

Human and Robot must achieve and maintain common ground.
HANDSHAKE: At the start of any interaction between a human and a robot, tokens must be exchanged that establish the relationship. Establishing the relationship is a prerequisite for commencing tasks associated with that relationship. The tokens will validate qualification both on the robot and human's part to participate in the task. A robot may not interact with a human without a valid and completed handshake.
DECLARATION: A defined set of parameters concerning the environment should be agreed to before the interaction commences. The parameters will define thresholds for a range of dimensions that define the limits and constraints of the work system competence envelope.

Human and Robot must model each other's intentions and actions.
CONSENT: A robot may not proceed with a task without the consent of the human in the role relationship or a proxy for it. The human must consent to the robot going ahead with a task. This requires that the human is cognizant of the DECLARATION. In certain cases, for example, where the robot is working with a patient with dementia or autism, consent may be established with a third party acting as the responsible agent on behalf of the human in the human–robot relationship.

Human and Robot must be mutually predictable.
TEST: A test or examination must be conducted to demonstrate that the robot is certified to act in a predefined role. There must be a documented probationary period evaluated by robot developers and an independent adjudicator.
START: A robot should start its tasks within a role relationship only when the HANDSHAKE and DECLARE have been completed, DISTANCE minimized, and CONSENT obtained.
SACRIFICE: A robot must sacrifice itself if the environment and conditions indicate severe physical or mortal threat to the human. Possible conditions for sacrifice will be defined in the DECLARATION. Severe conditions are highly likely to involve significant UNCERTAINTY, and will elicit OVERRIDE.

Human and Robot must be mutually directable.
STOP: A robot must completely cease an activity the moment the human indicates a wish for the robot to stop that activity. Any continuation of activity will constitute an attempt to control and dominate and violate the human's rights. The robot should maintain a catalogue of stop signals, as part of the DECLARATION, including the human saying stop,crying out in pain, making a gesture (such as falling over), and so on.

Human and Robot must make their status and intentions obvious.
DECLARE: The defined limits of task, intervention, technical capability, robot usage, and responsibility must be declared clearly, agreed upon by the user or the user's proxy, and recorded in the DECLARATION before activity commences.
INFORM: The robot should inform the human whenever a requested action has the potential to violate the defined limits and constraints on activity and responsibility.
TELL-ME: A request for information, clarification, or understanding of the robot's actions from the human, the proxy, or the controllers must take precedence over the task. Communication and connection must always take precedence over task completion.

Human and Robot must be able to observe and interpret signals of status and intention.
CHECK: No action should be undertaken by a robot without reference to the role relationship. A request for action from the human will be declined if it is not valid for the role and will result in an AUDIT.
IN-THE-LOOP: A robot may not operate in a situation where there is no connection with the human. When the robot is left to get on with tasks, a channel must remain open to communicate with the human. The robot must cease activity if the channel is broken.
DISTANCE: The distance between a robot and a human partner should always be kept to a minimum in time, geography, knowledge, and cognition. Thresholds for distance will be defined in the DECLARATION. It will be expected that the robot engineer will provide acceptable justification for any increase in that distance.

Human and Robot must be able to engage in goal negotiation.
OVERRIDE: A robot may override the control of the human if and only if the environment, conditions, and physical parameters indicate severe physical or mortal threat to the human in the role relationship.An override will require reference to the DECLARATION and be preceded by the triggering of other rules, including CHECK and INFORM.

Human and Robot must be able to collaboratively manage their attention.
AUDIT: There must be a periodic check during an interaction that the HANDSHAKE remains valid and that no aspects of competency, role, or task have changed. If the audit is not successful, the HANDSHAKE is broken, and the relationship ceases until established by a new HANDSHAKE.

Human and Robot must actively help control the coordination costs.
UNCERTAINTY: As the task complexity increases, the rate of exchange of information between the human and the robot must increase appropriately.

Figure 1. Possible interdependence rules that fall at the human–robot boundary.

or we can set rules of engagement. This applies Ross Ashby's principle of requisite variety—the number of options possible in the system must match the number of different states in the environment.[16] Of course, such limits can be dynamically defined. The negotiation and definition of the boundaries will lead to safety and ethics standards that are likely to be codeable in robot behavior.

By defining and limiting the environment, we might reduce the range of ethical problems likely to be experienced, but we would certainly change that range. As the robot moves to a new environment and a new task within its capabilities, we can redefine the borders and create new limits, which again constrain the ethical variety.

The Interdependence Rules for Human-Centered Robotics

Interdependence is critically dependent on communication at the machine–human boundary. Like other system boundaries, this boundary is a filter of information. There may be prejudices, tacit assumptions, unspoken expectations and differences in meaning and understanding that can disrupt the partnership. Unless prejudices are removed, expectations declared and negotiated, and meaning and terminology agreed upon, the relationship could collapse, tasks might not be completed, and goals might not be met.

In crossing the boundary between the robot and the human, a translation process occurs. In the human–robot interaction, this is a significant translation, from human intuition, perception, and insight to machine rules, algorithms, and executable instructions. Messages lose or change meaning; they lose detail and content. What is made of the message depends on the receptors that carry the message across, and what receptors are available will be influenced by the mechanisms and actions programed within the robot.

In managing the ethical gap, human-centered robotics should focus first on rules and guidance for communication at the boundary between the human and the robot. The approach should consider the human and robot as one interactional system while recognizing the capabilities of the partners in the interactional system.

The interdependence rules presented in Figure 1 are offered as a starting point for the envisioning of algorithms that reference the human–robot boundary. These are inspired by the "challenges for making machine agents team players,"[17] but reinterpreted for the human–robot context. The purpose of the rules is to create ethically acceptable conditions for starting, progressing, and stopping a human–robot interaction. Some of the rules identify directions for more detailed rule setting and algorithm development.

If the environment or task changes, this must result in increased information flow between the robot and the human to enable adaptation of the human–robot interaction to that change. INFORM, CHECK, TELL-ME, and UNCERTAINTY all address the need for unbroken information flows between the robot and the human and recognize the primacy of the interaction over the autonomous action of either the human or the robot. Increased DISTANCE, for example, increases the risk of loss of communication. Additionally, the possibility of loss of contact, perhaps though a loss of focus by the human, is addressed by a regular AUDIT and renewed HANDSHAKE to check that the conditions of the human–robot interaction have not changed.

These interdependence rules are intended to be suggestive; we do not think them uniformly correct, complete, or exhaustive. However, they do indicate a move away from the general principles initially represented in Asimov's laws and developed in the alternative laws toward something that can be implemented. This recognizes that any ethical decision or action taken within a bounded situation is a result of the interdependence and not an autonomous decision of the robot.

Ethical rules need to have the potential to be developed into implementable algorithms. However, in producing algorithms, and hence crossing the ethical gap, it is inevitable that something is left behind and some intuition and ethical knowledge is lost. That is why ethical robots must retain an umbilical link with human partners and ensure that human awareness of the task situation and environment is maintained. If that link is lost and the human ceases to be IN-THE-LOOP, the robot ceases to be ethically viable, and activity should STOP first, before it might trigger a new AUDIT and HANDSHAKE.

The interdependence rules shift the focus from an unachievable functional morality to responsibility and resilience at the system level. Because any ethical action or consequence occurs within a relationship, an effort to cast the robot into a lonely ocean of self-sufficiency and total autonomy is bound to fail.

We have highlighted the problem of moving from principles and general rules in the human domain to the specific and algorithmic rules of the machine domain. In characterizing this ethical gap, we focus on the boundary between the machine and the human and the communication across that boundary. We would suggest that the design of this communication is a priority in robot design, perhaps even more important than the design of what the robot actually does. Creating practical ethical guidance must require defining the capacities of both the human and the robot to reduce uncertainty. ■

References

1. P. Bizony, "Asimov's Three Laws of Robotics," *Eng. and Technology Magazine*, 15 June 2015; http://eandt.theiet.org/magazine/2015/06/asimov-3-laws-robotics.cfm.
2. K. Capek, *Rossum's Universal Robots*, Penguin, 1920.
3. I. Asimov, *I, Robot*, Gnome Press, 1950.
4. R.R. Murphy and D.D. Woods, "Beyond Asimov: The Three Laws of Responsible Robotics," *IEEE Intelligent Systems*, July/Aug. 2009, pp. 14–18.
5. J.M. Bradshaw et al., "The Seven Deadly Myths of 'Autonomous Systems,'" *IEEE Intelligent Systems*, May/June 2013, pp. 54–61.
6. R.R. Hoffman, J.K. Hawley, and J.M. Bradshaw, "Myths of Automation Part 2: Some Very Human Consequences," *IEEE Intelligent Systems*, Mar./Apr. 2014, pp. 82–85.
7. M. Johnson et al., "Seven Cardinal Virtues of Human-Machine Teamwork," *IEEE Intelligent Systems*, Nov./Dec. 2014, pp. 74–79.
8. R. Newman, "Can Robots Be Ethical?" *Philosophy Now*, Oct./Nov. 2015, pp. 30–32.
9. R. Hursthouse, *On Virtue Ethics*, Oxford Univ. Press, 1999.
10. J. von Neumann, *The Computer and the Brain*, Yale Univ. Press, 1958.
11. N. McBride and B. Stahl, "Developing Responsible Research and Innovation for Robotics," *Proc. IEEE Int'l Symp. Ethics in Science, Technology, and Eng.*, 2014; doi:10.1109/ETHICS.2014.6893392.
12. K. Christofferson and D.D. Woods, "How to Make Automated Systems Team Players," *Advances in Human Performance and Cognitive Eng. Research*, vol. 2, Elsevier Science, 2002, pp. 1–12.
13. R.R. Hoffman et al., "Trust in Automation," *IEEE Intelligent Systems*, Jan./Feb. 2013, pp. 84–88.
14. R.R. Hoffman et al., "A Rose by Any Other Name ... Would Probably Be Given an Acronym," *IEEE Intelligent Systems*, July/Aug. 2002, pp. 72–80.
15. K. Pitsch, "Limits and Opportunities for Mathematicizing Communicational Conduct for Social Robotics in the Real World? Toward Enabling the Robot to Make Use of the Human's Competencies," *AI and Society*, 22 Dec. 2015.
16. W.R. Ashby, "Requisite Variety and Its Implications for the Control of Complex Systems," *Cybernetica*, vol. 1, 1958, pp. 83–99.
17. G. Klein et al., "Ten Challenges for Making Automation a 'Team Player' in Joint Human-Agent Activity," *IEEE Intelligent Systems*, Nov./Dec. 2004, pp. 91–95.

Neil McBride is a reader in IT management at the Centre for Computing and Social Responsibility at De Montfort University. Contact him at nkm@dmu.ac.uk.

Robert R. Hoffman is a senior research scientist at the Institute for Human and Machine Cognition. Contact him at rhoffman@ihmc.us.