The Renovated 中文 *Room*: Ethical Implications of Intentional AI in Learning Technology

Jeanine A. DeFalco [1], Shelly Blake-Plock [2], & Andrew J. Hampton[3]
[1]U.S. Army CCDC–Soldier Center, [2] Yet Analytics, [3] University of Memphis

INTRODUCTION

In the (1980) essay, *Mind, brains, and programs*, John Searle maintains that an artificial intelligence (AI) program could not be realized to create a condition of understanding, perception, action, learning, and other intentional phenomena because only causal powers and the actual properties of synaptic sequences can instantiate the intentionality necessary for human-like understanding. Searle maintained that mere engineered instructions that manipulated formal symbols could not bring forth intentional understanding because symbol manipulations by themselves do not have intentionality, meaning, or a consciousness to make plans and achieve goals. Searle goes on to illustrate this argument with his *Chinese Room* thought experiment. Here he argues that inserting something that has intentionality (e.g., a person) into a system, but restricting their observable behavior by means of a formal program, essentially obviates that intentionality.

New developments in AI are beginning to suggest that intentionality could at some point be a feature—perhaps a core feature—of an artificially intelligent system. As intentionality implies an autonomous decision-making capability, it would stand to reason that such a system would on occasion be required to navigate ethical decisions—or decisions with ethical implications. If systems were designed with intentionality—meaning, a consciousness to make plans and achieve goals—how would we ensure its ethical nature, particularly if the AI system is intended to be autonomous?

The goal of this paper is not to suggest when intentionality will become a feature of AI. Rather, the authors have intended to provide a number of considerations with regard to ethics that may be beneficial to the design of AI-enabled systems in the present by considering the current state of personalized remediation practices. This includes the capability of leveraging reinforcement learning techniques in, for example, GIFT (the Generalized Intelligent Framework for Tutoring), as well as imagining what future ethical dilemma could look like if (or when) AI becomes capable of intention. By anticipating and designing for potential future ethical accidents or threats, one may improve upon the capabilities (and considerations that result in capabilities) of the systems feasible in the current technological paradigm. Therefore, the authors' position regarding this approach argues for a more deliberate articulation and standards-driven method of establishing an industry wide ethical framework and recommended practices that will inform the design and execution of AI driven systems where those systems may feature intentionality. In designing an ethical framework and processes for ethical risk assessments, the authors recommend designing for the eventuality of intentional AI. We feel this is particularly important as it relates specifically to AI-driven adaptive instructional systems (AISs).

INTENTIONALITY IN AI

Philosopher Robert Sparrow, in his paper *Why machines cannot be moral* (2021), insists that ethical reasoning will remain in the domain of human beings because ethics is inherently personal, subjective, and contextualized in a way that AI can never replicate; that the fundamental *personal* way in which we respond and reason about ethical dilemmas requires the *intention* to engage in ethical reasoning and behavior. Sparrow goes on to argue that this intention is grounded in individual subjectivity, and cumulative historical experiences that inform a person's response and responsibility in answering ethical dilemmas. This intentional and personal nature of ethical engagement cannot be realized in a reliable AI calculation, Sparrow argues,

because fundamentally, AI systems have neither the capability to experience emotional remorse that shapes ethical thinking, nor the established moral authority acquired over a history of realized ethical behaviors necessary for solving ethical dilemmas. Further, Sparrow maintains, "moral machines" would be incapable of identifying and resolving ethical dilemmas even if trained on datasets of ethical texts and judgements… No scientific calculations could ever adequately simulate the necessary associative personal regret that hallmarks the incentive for ethical behavior. In short, Sparrow maintains that a moral machine cannot be realized because engineers and designers do not (and presumably could not) understand the nature of ethics—which is fundamentally and uniquely shaped by the subjective personal stance and intentions of an individual.

Where Sparrow's analysis fails, as it fails similarly with Searle, is in the presumption that AI systems have a pre-ordained limitation in their capability to replicate intentional decisions. However, Sparrow's analysis exceeds Searle's constrained analysis in his assertion that ethical responses are defined by their associated affective responses by which people are incentivized to engage in ethical behavior. The authors of this paper counter this position by arguing that affective responses do not in and of themselves ensure nor promote ethical behavior, and to tie the potentiality of ethical behavior to the affective realm is, from a cognitive psychology perspective, in error. Chen et al. (2019) note that while Bandura (1999) theorized that internalized social moral standards regulate behavior through shame, guilt, or remorse, this self-regulation process can in fact be bypassed through cognitive mechanisms, e.g., employing disengagement strategies to evade self-condemning reactions: "The use of moral disengagement strategies enables individuals to engage in unethical behavior without self-disapproval," (Chen et al., 2019). This is an important distinguishable element particularly when seeking to model ethical processes in AI systems: because while self-deception and affective disengagement are elements that may interfere in the execution of an ethical behavior in humans, AI systems can maintain a consistency of ethically constrained actions by design.

But perhaps more importantly, Sparrow's analysis is short sighted in that—if followed to its natural conclusion—the dismissal of the possibility of intentional ethical AI systems risks abandoning any effort toward designing moral or ethical machines. To the contrary, the aim should not be an all or nothing venture. Even if a perfectly designed ethical AI system is not something presently realized, the authors of this paper argue that the aim should be to consistently re-train and provide data to support the continued development of the self-improving, intentional capabilities of an AI system. Sparrow's dismissal, in fact, creates a fertile ground for allowing invisible ethical threats to become manifest, as AI systems will increasingly be used to execute tasks independent of a human in the loop, and these tasks may contain ethical risks that threaten human flourishing—ensuring human flourishing, we argue, as the first principle of ethical AI (Stahl et al., 2021).

In short, advances in the design, function, and breadth of capability within AI has changed its fundamental nature. This level of complexity argues against a simplistic dismissal of the potentiality of devising ethical machines, and dismissal of AI as mere processes or symbol manipulation as was the case for Searle's ignorant man in the Chinese Room, or constraining ethical behavior as being regulated by affective responses according to Sparrow. Whether or not this advancement constitutes "true" intentionality is essentially irrelevant and missing important assessments of ethical risk. The fact remains that AI systems often (and increasingly) function as semi-independent agents, beyond the immediate control or understanding of even their designers. Decision-making and other intelligently executed actions thus demand careful consideration in both design and evaluation. Simply stated, if we intend to let machines make their own decisions, we need to know they do so within an accepted ethical framework.

IMPLICATIONS FOR GIFT COMMUNITY

AISs have demonstrated utility as training and education systems, as seen most evidently in the GIFT community. GIFT and other AISs systems have been deployed in military, academic, and commercial versions. These systems AISs have the potential to significantly improve both quality and scale of learning across many sectors and to be the minders of ethical values within the experiential space of learning for both their human and machine interactants, especially as they employ the scale resources of commercial AI

infrastructure and the opportunity of Linked Data and an ever evolving semantic data pool. As many AIS deployments now reside in or interact with the cloud, two market needs have been created: one for transparency concerning the operation, features, functionality, and use of AI in these systems; and one for the interoperable exchange of data with other learning and enterprise systems.

Focusing on the need for transparency, when we consider, for example, GIFT's ICAP pedagogical model that supports personalized remediation practices for individual learners, it has been designed to apply Markov Decision Processes and reinforcement learning techniques to establish remediation policies that determine what learning concepts and remediation content to deliver (Goldberg et al., 2020). This process that leads from presentation of information to gathering of evidence ultimately aimed at asserting competency is currently a tightly scoped and highly defined process. However, we would expect that subsequent generations of AI in GIFT will contribute to the automation of that process and that the process itself may become increasingly subject to the ongoing development and maturity of the decision-making capability of AI. It is reasonable to presume that GIFT will expand the scope of the automation of pedagogical decisions for learners that will have a direct downstream effect on business factors such as competency assertions.

Yet, the efficacy of ethical guidelines or codes as a basis for ethical decision-making for software engineers is effectively nonexistent (McNamara et al., 2018). Enforceability of aligning AIS design decisions to ethical frameworks will not occur by mere consensus across private and public sectors. Rather, policy makers must establish safeguards through legal measures and standards that incentivize compliance. Importantly, the consideration of ethics in AI needs to be reframed from a negative, restrictive mindset and rather as risk assessments that actually improve the scope of action, uncover blind spots, promote autonomy and freedom, and foster self-responsibility (Hagendorff, 2020). Within this context, we see the GIFT community as instrumental in contributing to the establishment and reinforcement of ethical risk assessment norms in their current and future design and implementation of GIFT and comparable AISs.

While trust in technology has been a longstanding area of concern with new and emerging technologies, it is important to highlight that one of the central functions of ethical thinking and reasoning is as a tool to bolster trust in a system that identifies blind spots and unanticipated threats as it relates to human flourishing. In complex component systems, risk assessments of engineering design is key to the engineering processes as it is an established principle that a system of significant scale will produce "normal accidents," (Williams & Yampolskiy, 2021). Similarly, in anticipation of a continued expansion and complexity of AI "components," risk assessments need to include an assessment beyond the mere mechanics of a system and include thorough analyses of ethical risks that threaten or even simply impede human flourishing. For example, unethical AI could include conducting unmonitored forms of AI experiments on society without informed consent, collateral damage form data breaches, biased and unfair algorithms, hiding harmful or flawed AI functionalities under the guise of trade secrets, vulnerabilities to cyberattacks, identity theft, disclosing personal data via machine learning applications, attacks on IT infrastructures, misinformation to perpetuate fraud or social engineering (Hagendorff, 2020).

More relevant to the AIS domain and the GIFT community of users, there are several types of ethical decision-making events that are specific to the learning and training domain which are likely to occur as AI matures in the educational domain. Not surprisingly, many of these events have corollaries in the ethical judgement process made by human instructors in the course of everyday work. They include:

- Identifying and subsequently dealing with cases of cheating
- Making adaptations to deadlines and schedules based on unforeseen or developing circumstances
- Allowing work to be turned in late for one of many reasons
- Deciding to issue or include a trigger warning with specific content
- Handling a student request to be excused from engaging with certain content
- Designing fair and equitable groupings of learners both homogeneous and heterogeneous
- Playing up a falsehood for the purpose of eliciting an instructional response

- Judging winners and losers of zero-sum and non zero-sum games, including in subjective events such as awards for artistic achievement
- Knowing when to alert authorities to a situation
- Deciding how to handle a parent or third-party request
- Choosing not to provide information upon determining a request from a bad actor
- Making decisions with life or death consequences in high stakes training environments

Additionally, there are decision-making events with ethical implications that fall outside of the normal course of human instructor experience. They may now or in the future include:

- Auto-scaling of one region versus another in the midst of a service disruption during high-stakes assessment delivered via distributed means
- Automated decision to share biometric or educational data with another application when those data are undefined per a service license or policy
- Privacy scope protocols when leveraging Linked Data across the internet
- Inherent bias amplified due to the nature of training data sets
- Dynamically evaluating decision criteria for A/B testing of features or algorithms that advantage one group over another.
- Pushing versus delaying updates based on the relative importance of optimization versus standardization of instruction.

Of more immediate concern in regards to GIFT is in the decisions that will drive competency assertions. In the case wherein decisions that GIFT makes in terms of what to provide to the learner, the experience provided holds the key to whether competency could be ascertained and therefore will have a direct outcome as to whether that learner should or can be asserted to have a competency. Currently, that decision is made in a virtual handshake between GIFT and an LRS upon the identification of target xAPI statements entering the data pool. In future iterations of GIFT, however, that decision making process has the potential to become increasingly automated--and when it is, eventually the automation is going to carry decision-making intentionality. For example, GIFT could decide to block a learner from accessing content or experiences necessary to demonstrate competency for a variety of decision-related reasons with ethical implications including risk management, prediction based on prior learning and history recorded in the learner's social or knowledge graph, and preference for promoting a learner with one behavioral profile over another regardless of the potential for training success. It is reasonable, then, for GIFT course authors to conduct ethical risk assessments when considering the effects and implications of regulating activity in such a way that a competency can or cannot be derived. And in the future it may mean that humans-in-the-loop will need to act as de facto referees of decisions made that have ethical leanings or moral consequences.

All of these and more dilemmas foreshadow the types of decisions which may fall into the hands (and artificial minds) of intentional AI systems in the future, including GIFT. In short, without anticipating the ethical threats that could occur from self-improving, intentional AI driven AISs, this oversight could cause loss and long term negative effects for individuals and society more broadly. Designing AISs that pass judgments and become the gatekeepers for growth and advancement of individuals without orienting that social power towards ethical principles such as honesty, justice, courage, empathy, care, civility, or magnanimity, could result in both short and long term societal detriments, including advancement for few and autonomy for none.

FUTURE CONSIDERATIONS

While contemporary AIs themselves are not capable of moral acts, that is not to say either that they will not be in the future nor that there may be acts carried out in the future that are not (by present definitions) considered moral acts. Hew (2014) points out that in a universe of ethical decisions composed of a rule set designed by humans, it stands to reason that any decision resultant in such a system be construed as an outcome of human design factors. But such a system is constrained by present knowledge of what such a system

contains—namely human-centric designs and expectations. In the same breath that we say that a future machine may write its own code, we can catch a whisper of what such machine-centric design may mean for ethical understanding. And such a shift in authorship, ownership, and indeed *intention*, could have the effect of creating a parallel set of ethics—or agreed upon rule sets with moral implications—created not by humans for human purposes, but by machines for machine purposes. As a field, AI developers must ultimately decide whether or not such parallel sets of ethics can co-exist, and if not, how to align such rules under the broader first principle of ensuring human flourishing. But this requires deliberation and foresight in initial engineering design planning.

If we were to follow Searle's reasoning that intent is the result of a causal power willed or otherwise negotiated by an instantiating process driven by synaptic sequences, then could not something like a drone programmed to leverage a neural network for decision-making purposes, such as in the context of a killchain, be deemed to be acting with intent as it executes its task? And if such a drone were to erroneously kill a civilian based on that decision, would it be the fault of a programmer who at no point during the decision-making process was ever in the loop? And if the neural network is constantly being updated by the delivery and gathering of data and generating decisions based on that constant flow of data, can we say that the decision occurs in any finite state? These arguments concerning intention, responsibility, and finite state seem as though written for an era prior to the one in which we find ourselves.

And there is the matter of who contrives the ethical universe. If we can agree that at some point in the future machines will be able to write their own code (and thus make their own decisions as to what to value in that design), then in the same way that we can note that an ethical universe may be designed by a human, so too a universe could be created by a machine. The ethical value of any system of rules within this artificially comprised universe could be indistinguishable from any such as created by humans. Once this leap of faith is made, can we then say that the machine is responsible for the decisions that it has made in its own ethical universe? Rather than wallow the philosophical muck of an ethical Turing Test, and rather than attempt to pin down blame for activities that will occur at a rate of speed and scale beyond human capacity to negotiate, we might be better off considering the design of AIs which themselves can act in this future artificial universe of machine-derived ethical rules as ethical referees among other AIs. In this way, we answer the question: "Are AIs capable of moral acts?" by asking the question: "Are AIs capable of minding other AIs?"

THE TASK GOING FORWARD

One of the central tenets of this paper is the notion that when considering ethical implications for AISs and GIFT in particular, our aim should not rest on the capabilities of AI as it is now, but anticipating what could be. Technological innovations, even if designed purposefully for human flourishing, still contain disruptive potentials (Hagendorff, 2020) that challenge our preconceptions on the stability of agency between humans and technology (Fischer & Wenger, 2021). A common refrain is that ethics is a process and not a solution (Boe et al., 2013; DeFalco & Hampton, 2020; Hagendorff, 2020). Stahl et al. (2021) asserts that the attempt to establish a stable definition of AI or the related ethical issues is misguided, and we should rather understand that ethics is dynamic and based on process and change wherein the integration of new technologies in society requires ongoing negotiations of facts and values.

These ongoing negotiations require regulatory governance, the adoption of legal frameworks, independent auditing of technologies, an investment in education that integrates ethics and technology, and standardization initiatives (Hagendorff, 2020). Standardization initiatives such as the IEEE P7000 family of standards and IEEE's P2247.4's working group that is establishing recommended practices for ethical considerations of AISs, are actively working on creating documents of consensus that can provide guidance to product developers and consumers. In addition, establishing mechanisms for assessing ethical risks—e.g., data protection, ethical, social, and human rights impact—within private corporations, as well as mandating ethical risk assessments for publicly funded acquisitions, would further the path of reinforcing normative AI ethics in the absence of legislation (Stahl et al., 2021). Ethical risk assessments could be informed by guidance from reports such as the *Ethical Framework for AI in Education* (2021) published by the Institute for Ethical AI in

Education at the University of Buckingham. This report identifies specific ethical concerns that should be considered when acquiring an AI-enabled learning capability:

- demonstrating efficacy in helping a learner to achieve educational goals
- implementation of a broad range of forms of assessment
- increasing administrative and workload capacity while respecting human relationships
- insurances of equity in learning
- enhancing autonomy of learners
- enforcing privacy
- transparency and accountability where humans are ultimately responsible
- informed participation by all constituencies
- adherence to best practices in ethical designs

These ethical concerns are important in the AIS domain if simply because the fundamental principle in devising a learning system is oriented specifically for human flourishing, and to omit proactive analyses to anticipate even unanticipated harm for its target population would be nonsensical. By designing ethical frameworks with an eye towards the possible world(s) fostered by an intentional AI future, we may protect the population engaged with the AIs of contemporary learning without putting unnecessary limitations on our ability to carry the philosophical and practical conversation into whatever the future may hold. Whereas limiting the ethical conversation to the AI capabilities of today may have the undue consequence not only of ill preparing our moral conversation, the saddling the policy on which our ethical values may be implemented with the ethical equivalent of technical debt.

CONCLUSION

We anticipate AISs such as GIFT will continue to make inroads as a central path for education and training for civilians and military personnel, as well as expand the capabilities of AI driven decision-making capabilities. AI technologies will continue to shape the evolution of AISs, and ethical considerations must acknowledge the growing complexity of AI and its increasing decision-making autonomy. This increasing decision-making autonomy of AI concerns decisions that an AI can take on its own with little or no prior human approval, intervention, or supervision. Whether or not an intentionally ethical or moral machine can be realized is almost an irrelevant speculation. What is relevant is that AI driven systems, and in particular AISs, will engage in decision-making that affects human flourishing, and it is to that point that organizations should assess ethical threats and establish processes to anticipate the unexpected. As part of the IEEE effort to establish recommended practices for ethical considerations in AISs, we suggest it would be beneficial to both GIFT and to the AIS field at large if GIFT stakeholders are actively taking part and contributing to the development of that standard, as well as make concerted efforts to establish normative ethical risk assessment processes in future design and implementations of GIFT. Preparing for what could be is perhaps the most ethical decision we can make.

REFERENCES

Bøe, T. D., Kristoffersen, K., Lidbom, P. A., Lindvig, G. R., Seikkula, J., Ulland, D., & Zachariassen, K. (2013). Change is an ongoing ethical event: Levinas, Bakhtin and the dialogical dynamics of becoming. *Australian and New Zealand Journal of Family Therapy*, *34*(1), 18–31.

Brendel, A. B., Mirbabaie, M., Lembcke, T. B., & Hofeditz, L. (2021). Ethical management of artificial intelligence. *Sustainability, 13*(4), 1974.

DeFalco, J. & Hampton, A., J. (2020). Dewey's ethics of moral principles and deliberation: Extending IEEE's ethics initiative for adaptive *i*nstructional systems. *International Conference on Human-Computer Interaction*. Springer, Cham.

Fischer, S. C., & Wenger, A. (2021). Artificial intelligence, forward-looking governance and the future of security. *Swiss Political Science Review*, *27*(1), 170–179. Retrieved at https://onlinelibrary.wiley.com/doi/epdf/10.1111/spsr.12439

Goldberg, B., Brawner, K., & Hoffman, M. (2020). The GIFT Architecture and Features Update: 2020 Edition. *Proceedings of the 8th Annual GIFT Users Symposium (GIFTSym8).* Retrieved at https://gifttutoring.org/attachments/download/3708/giftsym8_proceedings.pdf

Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. *Minds and Machines, 30*(1), 99–120. Retrieved at https://link.springer.com/article/10.1007/s11023-020-09517-8

Hew, P. C. (2014). Artificial moral agents are infeasible with foreseeable technologies. *Ethics and Information Technology, 16*(3): 197–206. doi:10.1007/s10676-014-9345-6

Institute for Ethical AI in Education. (2021). *The Ethical Framework for AI in Education.* University of Buckingham. Retrieved at https://www.buckingham.ac.uk/wp-content/uploads/2021/03/The-Institute-for-Ethical-AI-in-Education-The-Ethical-Framework-for-AI-in-Education.pdf

Mittelstadt, M. (2019.) Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*. *1*:501–507.

Searle, J. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences, 3*(3), 417–424. doi:10.1017/S0140525X00005756

Sparrow, R. 2021. Why machines cannot be moral. *AI & Society: Journal of Knowledge, Culture and Communication.* Published Online: 21 January, 2021. DOI: https://doi.org/10.1007/s00146-020-01132-6.

Stahl, B. C., Andreou, A., Brey, P., Hatzakis, T., Kirichenko, A., Macnish, K., ... & Wright, D. (2021). Artificial intelligence for human flourishing–Beyond principles for machine learning. *Journal of Business Research*, *124*, 374–388.

Williams, R. (2021). Understanding and Avoiding AI Failures: A Practical Guide. *arXiv preprint arXiv:2104.12582*.

Zhang, B., Anderljung, M., Kahn, L., Dreksler, N., Horowitz, M. C., & Dafoe, A. (2021). Ethics and governance of artificial intelligence: Evidence from a survey of machine learning researchers. arXiv preprint arXiv:2105.02117.

ACKNOWLEDGEMENTS

The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

ABOUT THE AUTHORS

Jeanine A. DeFalco, PhD, *is a Research Psychologist at the U.S. Army Combat Capabilities Development Command – Soldier Center, SFC Paul Ray Smith Simulation & Training Technology Center in Orlando, FL. She serves as chair for IEEE working group P2247.4, Recommended Practice for Ethically Aligned Design of Artificial Intelligence (AI) in Adaptive Instructional Systems.*

Shelly Blake-Plock, *is President and CEO at Yet Analytics, Inc. He is an officer of the IEEE Learning Technology Standards Committee.*

Andrew J. Hampton, PhD, *is a Research Scientist Assistant Professor at the Institute for Intelligent Systems & Department of Psychology, within the University of Memphis. He also serves as chair of the IEEE Standards Association working group P2247.1 for Adaptive Instructional Systems.*