by: Liam G. McCoy, Jacquelyn Burkell, Dallas Card, Brent Davis, Judy Gichoya, Sophie Le Page and David Madras

Reflections on the Process:

Our team at the Summer Institute was diverse in both skills (including technical computer science, cognitive science, systems innovation, and radiology expertise) and career stage (including faculty, graduate students, and a medical student). We were brought together at the 'pitch' stage by a mutual interest in human-machine partnerships in complex, high-stakes domains such as healthcare, transport, and autonomous weapons. We began with a focus on the topic of "meaningful human control" – a term most often applied in the autonomous weapons literature, which refers broadly to human participation in the deployment and operation of potentially autonomous artificial intelligence (AI) systems, such that the human has a *meaningful* contribution to decisions and outcomes.

We began from an applied perspective, with a how question: how might meaningful human control be created in high-stakes domains beyond autonomous weapons, and what threats might be faced during implementation? We worked backwards, and markers in hand we filled the windows of the large AMII boardroom with necessary preconditions for meaningful human control – focusing on challenges that could render human control *meaningless*, cause meaningful control to be *inhuman*, or cause *control* to be lost altogether. We continued to cover windows with specific design issues related to human factors, machine systems, and deployment environments that would influence "meaningful human control" in human-machine partnerships.

As the hours passed, our conversations grew more productive and contentious. We had made progress in understanding the design principles behind the a *priori* goal of meaningful human control, but questions remained. What did the human add to the equation? What did the human take away? As incredible progress is made in the

by: Liam G. McCoy, Jacquelyn Burkell, Dallas Card, Brent Davis, Judy Gichoya, Sophie Le Page and David Madras

space of AI, *is meaningful human control something we even want at all?* If so, what is the continuing value of meaningful human control? Amidst the discussion a clear consensus emerged: the core of our paper should focus on the importance and implications of meaningful human control, rather than the conditions required to attain it.

We are still committed to the framework that so painstakingly painted the windows, but in this initial paper from the Summer Institute we seek to answer the *why* before we get to the *how*. For this rapid web publication (in line with the concrete impact focus of the Summer Institute that kept us motivated all along), we have generated a preliminary draft of the former. We will define the concept of meaningful human control more precisely, explore background literature, and briefly outline our framework for the upsides and downsides of the concept in high-stakes environments. This preliminary publication will be updated with a link to the full work as it emerges in the fall.

Introduction:

The autonomy of artificial agents is an important aspect of defining machine-human partnerships. There is no clear consensus definition for the concept of *meaningful human control* (MHC) (Santoni de Sio et al., 2018). Broadly it connotes that in order for humans to be capable of controlling – and ultimately responsible for – the effects of automated systems, they must be involved in a non-superficial or non-perfunctory way. In particular, the concept of MHC emphasizes the "threshold of human control that is considered necessary" (Roff et al., 2016, p.1), to go beyond the ambiguous concept of humans "in-the-loop", or merely setting initial parameters and providing no ongoing control. Furthermore, the concept of MHC rests on the assumption that

by: Liam G. McCoy, Jacquelyn Burkell, Dallas Card, Brent Davis, Judy Gichoya, Sophie Le Page and David Madras

humans have exercised control over systems such as weapons in the past and present, and is concerned with maintaining this assumed human control (Ekelhof, 2018). Within this paper we aim to explore the concept of meaningful human control and its value as a concept for evaluating the balance of autonomy throughout the broader landscape of human-machine partnerships in high-stakes environments.

Background:

The concept of "meaningful human control" is most closely associated with lethal autonomous weapons systems, where there is general agreement that autonomous weapons capable of taking human life should not operate without human participation in or oversight of the decision-making process (see, e.g., Crootof, 2016; Roff and Moyes, 2016; Santoni de Sio et al., 2018). The term first appeared in this literature in 2001, in an article that discussed the inevitable rise of autonomous weapons systems and the accompanying challenge to meaningful human control of those systems (Adams, 2001). There are three broad and inter-related classes of reasons for requiring human participation in machine decision making. The first is rooted in the (arguably) unique capacity of human beings to make moral/ethical decisions, based in empathy and compassion (e.g., Asaro, 2006; Docherty, 2018). The second focuses on the ascription of legal responsibility or accountability (Hubbard, 2014; Calo, 2015; Scherer, 2016). The third class of reasons concerns performance – specifically system redundancy, error detection, and recovery – on the premise that humans can (at least for now) do some things, or do things in some ways, that machines cannot. These issues, of course, are not limited to lethal autonomous weapons: the issue of meaningful human control arises in other, typically high-stakes, contexts. Examples include the operation of autonomous vehicles (Heikoop et al., 2019) and surgical robots (Ficuciello et al., 2019). MHC has also been cited as a key

by: Liam G. McCoy, Jacquelyn Burkell, Dallas Card, Brent Davis, Judy Gichoya, Sophie Le Page and David Madras

challenge in the continuing development of robust and humane artificial intelligence systems (Russell et al., 2015; Stephandis et al., 2019).

Meaningful Human Control: Pros

The reasons meaningful human control is desirable in automated systems can be broadly divided into the categories of *performance-related* and *responsibility-related* – concerned with how well the autonomous system is able to perform the desired action, as opposed to the process of action selection itself.

Humans are adaptable, which can improve performance particularly on unusual inputs. Because machine learning-based systems are built to perform well on a pre-specified training set, they may underperform on novel or atypical inputs. These inputs may be benign (simply *out-of-distribution*), yet still yield harmful outcomes. For instance, an automated debt calculation system ("robo-debt") run by the Australian government frequently overestimated debts for people with highly variable income streams, who were not considered in the algorithmic design (Henriques-Gomes, 2019). Inputs can also be atypical in malicious ways – *adversarial examples* are a known vulnerability of computer vision systems (Akhtar and Ajmal, 2018). These are intentionally constructed to "fool" computer vision models into making incorrect classifications, yet appear unremarkable to the human eye (Goodfellow, 2014). In both these cases, human control over the system's response to the "atypical" input would allow for superior performance of the human-machine partnership.

Another important motivation for meaningful human control is adding redundancy to an otherwise automated system. Even on tasks where machine errors are highly infrequent, the character of their errors may differ greatly from human errors, in

by: Liam G. McCoy, Jacquelyn Burkell, Dallas Card, Brent Davis, Judy Gichoya, Sophie Le Page and David Madras

ways that can lead to catastrophic outcomes. Human oversight introduces heterogeneity to the decision-making progress, which can mitigate these risks. Airplane flight provides an example of a well-studied human-machine partnership which displays this characteristic. Airplanes are mostly guided by highly effective automated systems. Yet it is widely considered essential to have pilots "behind the wheel" to oversee the autopilot, who are able to select between different levels of control in case of system failure (Sheridan, 1987; Mindell 1999). This is the case despite the downsides: human pilots are a frequent cause of accidents (Shappell, 2017), and can lose skills over time if they are infrequently used (Arthur Jr et al., 1998), as they may be under a regime of widespread automation.

The most technologically intractable reasons for meaningful human control are moral. Human decisions are imbued with a moral weight that we do not accord to machines, and we commonly rely on humans to interpret vague rules in determining real-world actions in a way that is sensitive to context and human values (Russell, 2015). Humans are seen as having a capacity for moral judgment and empathy beyond any advanced AI. Domains such as legal decision-making (e.g. sentencing, bail, and parole) call for meaningful human control due to their moral dimension, despite some evidence that algorithms can predict recidivism as well as or better than expert human decision-makers (Kleinberg, 2017).

Automated systems with no human control also raise concerns about legal liability and accountability. For example, if a robot harms a person, who should be held responsible and liable for compensation? Possibilities include the manufacturer, the programmer(s), the user, and the robot itself. This is a real-world scenario, which courts have already addressed to some extent (Calo, 2015), but the prospect of increasingly intelligent, autonomous, interacting systems – especially those capable of ongoing learning from their environment – will create many legal and financial

by: Liam G. McCoy, Jacquelyn Burkell, Dallas Card, Brent Davis, Judy Gichoya, Sophie Le Page and David Madras

uncertainties. Under American law, for example, the situation of an autonomous system causing harm in a way that was not intended or foreseeable falls into a lacuna, in which it is unclear who, if anyone, would be liable (Hubbard, 2014; Calo, 2015; Scherer, 2016).

Meaningful Human Control: Cons

Meaningful human control also has costs, which again can be divided into *performance-related* and *responsibility-related* types. These costs, which can range from the minor to the substantial, must be weighed against the benefits when considering implementing MHC within a given context. The "handoff problems" associated with moving from a fully autonomous system to a human-machine partnership may be substantial (Mindell, 2015). It may also be possible that, in certain cases, human decisions are consistently inferior to or more biased than the machine's choices.

Many tasks are designed to be machine-driven precisely because of their superior performance or efficiency. Adapting a system for meaningful human control requires creating a monitoring apparatus, and potentially pausing automated routines to insert decision points. This paper focuses on "high-stakes" domains, where the consequences of errors can be substantial. Yet there are also many tasks for which each decision is so trivial that the loss of performance or efficiency outweighs the potential benefits of human involvement. Networking equipment, for example, autonomously performs repetitive tasks rapidly and accurately, with little perceived need for meaningful human control.

The variability and adaptability of human input interferes with predictability and consistency. This is particularly true in highly coupled, tightly interacting systems.

by: Liam G. McCoy, Jacquelyn Burkell, Dallas Card, Brent Davis, Judy Gichoya, Sophie Le Page and David Madras

Consider the case of an integrated autonomous driving network, in which vehicles hurtle past each other through an intersection at high speed. Safety and predictability are tightly linked in such a scenario, and the uncertainty introduced by the possibility for human control would have cascading effects. Instead of *knowing* what each other actor will do and planning accordingly, agents would be forced to *monitor*, *project*, and *react* to others' behavior under uncertainty. Contexts can be conceived in which, even if any given human decision was more appropriate than its automated counterpart, the downsides of this decoupling far outweigh the benefits.

There are also cases where human decision-making may be undesirable due to humans' risk of bias (intentional or unintentional) or ulterior motive. Certain autonomous systems – such as autonomous arbitration systems or escrow services – could derive their usefulness precisely from the lack of human control. The potential for bias in human decision-making may provide an additional impetus for developing autonomous systems without MHC – though precaution must be taken to ensure that the system does not merely perpetuate and mask existing biases with a veneer of algorithmic objectivity.

Preliminary Thesis:

Meaningful human control is important to consider in the context of machine human partnerships in high-stakes domains. Human involvement may improve system performance by way of redundancy and increased adaptability, and plays an important role in ensuring ethical and legal responsibility. These benefits do not come without downsides, however, including both the potential for improper human choices and the efficiency losses associated with decoupling complex autonomous systems. Finding the context-specific balance between these trade-offs is essential

to ensuring effective, robust, and ethical performance in cases of autonomous human-machine partnership.

References:

Adams, T. K. (2001). Future warfare and the decline of human decision-making. *Parameters*, 31(4), 57-71.

Akhtar, Naveed, and Ajmal Mian. "Threat of adversarial attacks on deep learning in computer vision: A survey." *IEEE Access* 6 (2018): 14410-14430.

Arthur Jr, Winfred, et al. "Factors that influence skill decay and retention: A quantitative review and analysis." *Human performance* 11.1 (1998): 57-101.

Asaro, P. M. (2006). What should we want from a robot ethic?. *International Review of Information Ethics*, 6 (12):9-16.

Calo, R. (2015). Robotics and the Lessons of Cyberlaw. *California Law Review*, 103(3).

Crootof, R. (2016). A Meaningful Floor for Meaningful Human Control. *Temp. Int'l & Comp. LJ*, 30, 53.

Docherty, B. (2018). Statement on meaningful human control, CCW meeting on lethal autonomous weapons systems, April 22, 2018. Retrieved from https://www.hrw.org/news/2018/04/11/statement-meaningful-human-control-ccw-meeting-lethal-autonomous-weapons-systems, July 31, 2019.

Ekelhof, M. (2018). Autonomous weapons: Operationalizing meaningful human control.

Ficuciello, F., Tamburrini, G., Arezzo, A., Villani, L., & Siciliano, B. (2019). Autonomy in

surgical robots and its meaningful human control. *Paladyn, Journal of Behavioral Robotics*, 10(1), 30-43.

Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." *arXiv preprint arXiv:1412.6572* (2014).

Heikoop, D. D., Hagenzieker, M., Mecacci, G., Calvert, S., Santoni De Sio, F., & van Arem, B. (2019). Human behaviour with automated driving systems: a quantitative framework for meaningful human control. *Theoretical Issues in Ergonomics Science*, 1-21.

Henriques-Gomes, L. Labour calls on government to scrap 'malfunctioning' robodebt scheme, July 30, 2019. Retrieved from https://www.theguardian.com/australia-news/2019/jul/30/labor-calls-on-government-to-scrap-malfunctioning-robodebt-scheme, August 8, 2019.

Hubbard, F. P. (2014). 'Sophisticated Robots': Balancing Liability, Regulation, and Innovation, *Florida Law Review*, 66(5).

Keeling, G., Evans, K., Thornton, S. M., Mecacci, G., & de Sio, F. S. (2019, July). Four perspectives on what matters for the ethics of automated vehicles. In *Automated Vehicles Symposium* (pp. 49-60). Springer, Cham.

Kleinberg, Jon, et al. "Human decisions and machine predictions." *The quarterly journal of economics* 133.1 (2017): 237-293.

Mecacci, G., & de Sio, F. S. (2019). Four Perspectives on What Matters for the Ethics of Automated Vehicles. *Road Vehicle Automation* 6, 49.

Mindell, David A. (2015). *Our Robots, Ourselves: Robotics and the Myths of Autonomy*. Viking.

Roff, H. M., & Moyes, R. (2016). Meaningful human control, artificial intelligence and autonomous weapons. In Briefing Paper Prepared for the Informal Meeting of Experts on Lethal Autonomous Weapons Systems, UN Convention on Certain Conventional Weapons.

Russell, S., Dewey, D., & Tegmark, M. (2015). Research Priorities for Robust and Beneficial Artificial Intelligence. *AI Magazine*, 36.

Santoni de Sio, F., & Van den Hoven, J. (2018). Meaningful human control over autonomous systems: a philosophical account. *Frontiers in Robotics and AI*, 5, 15.

Scherer, M. U. (2016). Regulating artificial intelligence systems: risks, challenges, competencies, and strategies. *Harvard Journal of Law & Technology*, 29(2).

Sheridan, Thomas B. "Supervisory control." *Handbook of human factors* (1987): 1243-1268.

Shappell, Scott, et al. "Human error and commercial aviation accidents: an analysis using the human factors analysis and classification system." *Human Error in Aviation*. Routledge, 2017. 73-88.

Stephanidis, C. C., et al., (2019) Seven HCI Grand Challenges, I*nternational Journal of Human–Computer Interaction*, 35(14), 1229-1269, DOI: 10.1080/10447318.2019.1619259