Centre for Army Leadership

Leadership Challenges to Autonomous Weapons

June 2019

SLIDE 1 - INTRO

Good morning and thank you for giving me the opportunity to address you today on leadership issues thrown up by removing human supervision from weapon systems. My name is Paddy Walker. I am a rather newbe PhD, my title *Challenges to the Deployment of Autonomous Weapons*, having passed my viva only in April. But fortunately I am in front of you today in two other capacities. The first is as an ex-soldier and as a gunnery officer. I served in the 5th Royal Inniskilling Dragoon Guards in the 80s before disappearing to America to study Business. I have then spent the last three decades as an investor, often in those technologies that will underpin the same autonomy that we will be discussing today, either directly or through venture funds. In this capacity I also lead the Leon Group, a 4th generation family investment office based in Hampstead. Leon has an interesting charitable foundation and we sponsor several Third Sector initiatives, particularly in the human rights space. Much of our attention is directed at weapon matters, specifically the use of explosives in built-up areas, the appropriate accounting for casualties and, the subject for today, the deployment of weapons without meaningful human control.

SLIDE 2 - HRW

My second capacity today is that I co-chair the London Committee of charity Human Rights Watch. It was Human Rights Watch who were an early mover in pointing out pitfalls to removing weapon oversight. Their 2012 report on 'Killer Robots' built on the work undertaken by their Weapons Division under Steve Goose and Mary Wareham and has proved pivotal in

alerting society to the disruption and dangers that purportedly sentient weapon systems pose for already fragile ecosystems in this space.

Lethal autonomous weapons may be in their infancy, but I don't intend to spend any of my precious forty minutes justifying the prescience of the matter. It is useful, however, to spend a moment establishing common ground for our discussion. Today we are being quite specific in considering the following: 'Lethal weapons without meaningful human control in their processes' and where humans are out-of-the-loop. For our purposes this morning, I'm also discussing *wide* task and *wide* capability systems. This is then what I mean today when I use the term 'weapon'. We must also acknowledge that this is a '*future*-oriented' subject. Precursor weapons already exist that can impart violence independent of human involvement. But we are *very* early in the continuum for this phenomenon, but still a continuum that I will be arguing remains anchored in surprisingly rudimentary models where expectations already far outweigh what is currently deliverable from such a meld of technologies.

SLIDE 3 – SCHEDULE TODAY

In the interest of time and, given my particular interests in this arena, I am going to focus this morning on what are non-obvious and *technical* ramifications to leadership of deploying autonomous weapons. In doing so, I will also comment on the attributes required of – and pitfalls in front of – *leaders* in this emerging technological paradigm. This is generally a less trodden path and receives scant attention. I intend to restrict my observations to the macro, the advantage here being that broad technical constraints will apply to you regardless of the weapon's underlying technical spine, today the construct of machine learning. The nub of course

is this: As field commanders and those involved in the removal of supervision from weapon processes, you of course remain responsible for judging what constitutes appropriate compliance.

In giving this talk, I'm making various assumptions today. Given the pace we've seen in the deployment of unmanned, armed drones, I'm supposing that you are already well versed such weapon's *ethical*, *legal* and *moral* issues. I am therefore largely ignoring the key precepts of the Laws of Armed Combat. I am also assuming that we understand the several drivers to weapon autonomy - runaway budgets, the operational advantages of speed, of remote delivery and force multiplication. But it is also exactly this list that contributes to what London's Professor Sabin refers to as a *Revolution in Expectation*. My argument, *and* his, is that much of this is actually illusory given the cumulative costs of diluting human oversight. A 2nd issue for you, of course, is that such systems – regardless of deployment model - cannot themselves make legal determination and it is unequivocally the responsibility of what I call these weapons' *Delivery Cohort* upon which this falls. This Cohort therefore refers to the very many parties *leading* this deployment. From politicians to generals, soldiers on the ground to maintenance personnel, those in procurement, programming, the Press and those from the Third Sector. And it is certainly leadership that will be required to navigate these several competing constituencies. These constituencies also include *structural* roadblocks from deep-seated inertia, from complex rules of engagement and from challenges arising from multi-*service*, multi-*force* and multi-*national* deployment.

The final underplayed characteristic today will be the matter of *context*, to me the fundamental deployment challenge. As leaders, do you want great soldiers or the latest kit which, I hope to illustrate, can readily and enduringly be undone by low tech, by human ingenuity and the generally *fleeting* nature of technical advantage that Max Boot compellingly sets out in his theory of *nullification*.

SLIDE 4 - TERMINATOR

In order to pass useful comment on leadership in this digital age, let's also agree on what we're *not* talking about. First, our timeline today is near-term, today to 2025. The medium term here might extend to 2035. That's just 15 years with anything longer being conjecture and unusable. *Also* unusable are is the notion of Terminator-like structures holding your right flank. And within or without these timelines we are certainly not talking about weapon *sentience* or Artificial *General* Intelligence. Instead we are talking about machine substitution in ever-more engagement processes in what Human Rights Watch usefully calls 'the piece-meal ceding of human oversight in lethal engagement'.

_____

Likewise, we are talking today about machine learning and prediction algorithms that can uncover approximations and so automate decision-taking albeit in quite defined scenarios.

SLIDE 5 – PERHAPS

SLIDE 6 – DEPLOYMENT MODELS

And this is to be achieved using very large datasets of previous examples to train – although I don't like the word as it's too suggestive of solutions – to *train* the weapon system.

We are also discussing deployment models involving machine-human *teaming* where the *experience* of human soldiers is leveraged by companion autonomous technologies. But this creates our first leadership quandary. The human, after all, has *hitherto* been *operator, moral agent* and *failsafe* across an engagement's whole decision waterfall. It has been the *human* operator who has managed that engagement's intangibles such as task complexity, its cognitive workload, its number, type and duration. The leadership issue, taught every day here in Churchill Hall, has been to understand that tasking *in advance* of a mission and to recognise how states may change *within* the mission. Tasking, after all, is complicated by scale, margin of error, task space and colleague assets. As autonomy is introduced, slack time for these processes is reduced, scope for rule-bending and initiative is removed and, in our case this morning of machine learning, the leader's ability to *predict and* influence outcomes is lessened.

Before narrowing down onto the technical, it's also useful to raise some *behavioural* issues suggested by autonomy.

SLIDE 7 – BEHAVIOURAL ISSUES

In the first instance, fielding independent weapons certainly challenge traditional notions of what it is to be a combatant. This has several angles. As noted by Enemark, the longbow was outrageous at the time because of the new degree of *force* that it brought to the battlefield but also because the lowly peasant could now defeat the *aris*tocrat. Similarly, the drone operator *already* kills without any material *personal* risk. Autonomous weapons would appear to promise an even higher degree of user insulation. Perhaps, therefore, you and those

you command suddenly require *less* of the traditional courage that for millennia has distinguished the warrior profession from all others. But at least, I would argue, drone operators must meet a clear legal bar both through their tracking of targets and, arguably, the engagement protocols that must precede permission to engage. In removing oversight, where *really* does accountability rest in what has now become an unsupervised attack?

Enemark terms the notion 'Post heroic warfare', questioning the traditional construct of Nation States prepared to wage war on the basis of substantial casualties. A shift now to independent weapons suggests an even *wider* gulf between civilian and military values, a disappearing paradigm of *noble* death and the rise of a new norm which is based instead on remote warfare, few combatants and fewer casualties.

This certainly impacts those in your command. For those in battle, several *attributes* have long characterised combat soldiers across cultures and time. Endurance, courage, strength, skill and honour. Riskless war would appear to dilute these previous requisites of valour. No longer must you, the leader, be admired, respected, cheered… *mourned*. Remove human oversight and perhaps you weaken traditional notions of duty and self-sacrifice, the *covenants* binding soldiers together *and* to the society that they serve, the idea of subjugating self-preservation in the *present* to make life better in future. Riskless warfare, first the unmanned drone but soon the unsupervised weapon, certainly ends the concept of a soldier having a contract to kill that is based upon a risk of *being* killed.

And language here is interesting – you will know that a drone in colloquial Urdu is a *derogatory* term suggesting cowardice for sending robots to do man's work. But then there's also

the US Air-Force language of 'bugsplat' which seemingly belittles collateral damage in a remote strike.

SLIDE 8 – TECHNICAL ISSUES OVERVIEW

Ten minutes into the presentation on technical perspectives and only now can we start that technical review. First, ignore this slide which is intended only to convey quantum of what comprises this whole argument. And second, one further assumption. An autonomous weapon must work *first*-time, *every*-time. It must do this to be compliant, to enable you, its Delivery Cohort, to be conforming and, indeed, it's an obvious imperative if its deployment is to be a rational and an accretive addition to your arsenal. You do, after all, have choices here. It is for *this* reason that weapon efficacy boils down here to the weapon's *technical* spine and, currently, the fitness for purpose of *machine learning* as a replacement for human oversight in lethal engagements.

SLIDE 8 – MACHINE LEARNING ISSUES

It is useful to spend a moment to consider what comprises this *core* technology. How in weapons is machine learning currently envisaged (and note the word *currently* as the spine that we discuss today is far from set in stone and moves at alarming pace)? The overarching purpose of machine learning is to effect *deduction*, deduction and prediction whereby whatever the weapon has experienced in *prior* cases should inform what the weapon should expect now. Its purpose here is to process what is a *known case* whose relationships can then be carried over to the *present case*. An immediate challenge is that such architecture must also include quite *counterintuitive* capabilities such as detecting contradictions, evaluating significance and,

complicatedly, rejecting actions that leave the weapon with foreseen *unsatisfactory* outcomes. All of this requires a further and – as yet – quite *unforeseen* technical disruption to occur in software and hardware processes. Machine learning in its *current* guise is simply infeasible for these purposes and it is to this matter of *infeasibility* that I now turn.

Nearly all aspects of weapon operation require enduring *human* oversight and *human* calibration in order that they be properly useful to you as leaders in the Delivery Cohort. This is easy to demonstrate. What *tolerances*, for instance, will be required by you for a *defensive* weapon versus an *offensive* weapon that is spatially unbounded? And is this a linear relationship or one that changes with circumstance? A useful example is how you programme *out*-of-tolerance weapons to fail. Is it 'dangerously' whereby the weapon fights through malfunction and, without supervision, continues to engage targets that have been pre-authorised? Or is it 'deadly' where there is *no* off-switch? And is it *practically* feasible to deploy a model that is based upon *flexible* autonomy whereby humans and machines toggle control between parties? Humans, after all, perform erratically when required to intervene in moments of high stress or in situations of *limited* information.

The architectural intention is therefore that the weapon's artificial neural network *learns by example*. These models involve core applications such as pattern recognition, label matching and data classification in order to process outputs into a series of actions. So how might this work? Each neuron has a summation function *and* a threshold function. Should a battlefield signal exceed a threshold, which in theory has been defined you the Delivery Cohort, then it is

propagated forward on to other neurons in order to instigate a known action. The weapon *is* thus independent but based squarely upon human-set configuration, and *trained* rather than being explicitly programmed.

At this point it is useful to introduce some context. The very largest current artificial neural network today has some 16 million neurons. This is a laboratory example and equates perhaps to the cognitive function of a small frog. By way of reference, the human brain approaches some *hundred billion* such neurons, evidence (if it is required) that we are very *early* on this machine learning continuum. So what, then, are the *enduring* challenges here? Fundamentally, it is the management and application of machine learning's *prioritising* weights, without which the weapon's neural network is unable to derive *meaning* from newly sensed inputs. This propagation is a core function whereby the weapon's neurons must be dynamically coached to encourage specific behaviours and specific battlefield outcomes: If it's shaped roughly like a square, moves obliquely with a particular heat moniker then that is the target signature of a tank. To work , however, the model requires each string of data be processed on a near identical basis to that data which the weapon has *previously* encountered and it is *this* characteristic that leads to the acronym CACE - Change Anything and you Change Everything. New parameters, slightly heterogenous training data and then machine learning is empirically - and enduringly - compromised, particularly in what will be the dynamically changing nature of a battlespace and its prevalence of *hidden* or *partially observable* states.

There are myriad reasons why this methodology is not, can not be, fit for purpose in the field. The number of datapoints comprising an engagement sequence is intractably large and a material step-up from anything currently in the research laboratory. The leadership takeaway here is that maintaining human control is not simply a matter of ensuring compliance but is primarily a matter of feasibility and it's for this reason, I'm afraid, that it's useful to drill down further on the issues…. The management of individual neurons is governed by an *error function* and the goal of training is to *minimise* this function. This model, however, confines weapon learning to be an *approximator*, quite at odds and, I argue, quite inappropriate for compliance with the Laws of Armed Combat mentioned above. First, the process relies upon seamless data. Second, model performance systemically *plateaus* every time the weapon polls its sensors, making an ever-smaller change to the model's weights with every new iteration. And adding *new* training parameters and *new* layers to a network in order to broaden this data capture empirically means *extra* layers of non-linearity which makes optimizing the weapon's learning routines ever more challenging. A marginally different setup or a marginally different training dataset lead to very substantial differences in that weapon's decision-making. Machine learning empirically ignores features with only a *small* number of examples in a training set but ones that possibly account for features of critical battlefield importance. It is also fundamentally challenging for the weapon's sensors, its eyes and ears and its *sole* source of decision inputs, to garner appropriately *consistent* information. Smoke, reflectance, image echo as well as data intensity *and* saturation all comprise *practical* difficulties that hobble the model. And where that

data is noisy and indistinct, then class boundaries separating different class examples resist definition and become impossible to separate for *further* statistical analysis. Mismatches with its training set, basically anything out of the ordinary, will *confound* the weapon, whether the result of simple feint, by enemy surprise or by this insufficient data separation. The leadership knock-on for you is therefore to understand – and *own* - both the model's *sensitivity* (termed its detection rate) and its *specificity* (here, its false alarm rate). And if this worries you then you also need to watch that additional *supervision* doesn't introduce either bias or over-fitting to that weapon data.

SLIDE 9 – GENERIC ISSUES FROM ML

So what then are the generic leadership issues thrown up by this machine learning spine? A second quandary, of course, is that the unsupervised weapon cannot be sure it has found its *best* action for each state until it has tried *all* possible actions in *all* possible states. Once again, it is the *combinatory* requirement that this posits which is infeasibly large. Current learning models are empirically just *too* iterative to manage the parameter-rich procedure that is target selection, the more so without any appropriately definable end-state. Error, moreover, in the weapon's sensing of its *current* state will carry forward in that machine's *future* learning and *future* battlefield actions. At the same time, much of what the weapon has learnt may be invalid if its environment or its combat task *change* given what is a perpetual learning phase, the trade-off between a weapon that is 'constantly learning' as opposed to one that is using what is already known to work at the cost of missing out on further improvement. Indeed, the challenge of

incorporating *un-learning* into machine routines remains at best untested. Unsurprisingly, a key *in*efficiency therefore turns out to be the *preparation* of that weapon's input data: Fifty to seventy per cent of data analysis is currently taken up with data preparation, an inappropriate luxury on the quick-changing battlefield if the weapon is to be properly valuable yet still compliant.

What then is required for the weapon's *cognitive* ability? After all, the theory is that this should comprise *very many* processes including knowledge and memory management, the tenets of attention, judgment and evaluation, as well as problem solving and decision making. Without just *one* of these elements, is the weapon *actually* deployable? Can it really be useful and, for you, sufficiently *robust* to replace the human soldier? The difficulty, of course, is that cognition, even in its human condition, is challenging to define. It can be conscious or unconscious, concrete or abstract. It can also be intuitive (here, the 'knowledge' of a task) or conceptual (the 'model' of a task). For an autonomous weapon, the construct is also that its cognitive processes must use *existing* knowledge in order to generate *new* knowledge. Several leadership snags arise. Regardless of its data preparation, *current* collation models lack sufficient scripting differentiation. Weapon datasets must also be *incrementally* built as not all data is available to the weapon at the same time. Information arrives in waves and, empirically, the model falls over without seamless data-*priority*, data *allocation* and data *denial*. Little *transferable* research is going on around these central tenets. Finally to this point, the weapon's underlying data distribution must generally *evolve* with time and this contradicts the fundamental hypothesis of '*identically distributed data'* upon which machine learning and classic

data-mining algorithms rely.

How then can the introduction of autonomous practices be reconciled? In the first instance, future weapons will undoubtedly comprise an increasing number of independent *sub-parts*, each of which are capable of agency. Each and every weapon component may therefore be capable of autonomy in its *own* right and this raises additional complications. *Composite* agency, after all, complicates motivational selection and this complexity is unexpectedly fundamental to removing oversight as the motivations and therefore action selection of the *composite* weapon depends not only on the impulses of its constituent sub-agents but also on how those sub-agents are organised and coded.

SLIDE 10 –CODING CHALLENGES

Indeed, fitness for purpose is particularly compromised by *coding* challenges. It is coding, after all, that must express all of intentions *and* constraints of you, the Delivery Cohort. It is also coding that must *anchor* the weapon's goals and action selection. The challenge here is that the model is reliant upon frequent (possibly continuous) *scaling* factors and *conviction* weightings to work its sensed data. Empirically, data with *least variance* is given additional weighting in each new polling iteration. This has consequences. Confidence weightings have several unintended effects including data smoothing whereby that data - in the case of an engagement sequence, information on target signature, classification, location, threat and sensitivity to battlefield clutter - becomes inappropriately scaled to the mean to the extent of formlessness. War may be 9/10ths inactivity but sudden exogenous events are now prone to be

smoothed out of the weapon's calculations. The theory here may be that the weapon computes new probabilities for its immediate world, but the model is also based upon setting back to zero any *inconsistent* probabilities and then undertaking 'renormalizing' over the weapon's remaining possible outcomes. This is termed conditionalization whereby the weapon is calculating conditional probabilities for *each* set of possible causes *and* for each of its observable outcomes. Weapon operation is squarely based upon a series of posterior probability distributions to use as its *new* prior in every *next-time* step. On paper this may be fine but, in the case of the battlefield, the model is unsuitably prone to obsolescence. It is also *systemically* instable. An example here is in the weapon's movement and the unavoidable requirement that *all* relevant navigable space must first be identified, then processed and then made 'map-ready' for each and every of the weapon's representations. The resulting dataset must then be searched in real time to evaluate, first, *available* paths and then, second, the *best* path for the weapon *after which* the weapon's goals, values and action selection must all be revised to account for that newly selected path.

Staying a bit longer with coding challenges, the notion of meaning - here, the Cohort's intention –also creates challenge by the need to link data to a *particular* representation on one of the weapon's learning planes. Each such linkage must be coded so that *one* associatively-connected entity can evoke another. In this way, the meaning of an instruction, your instruction, should not be restricted to *just one* association. But coding's capture of abstracts is again *systemically* difficult. The information contained within your command must also be coupled to *previously* given information as well as to information that is to follow. This is difficult to test and

introduces further fragility. Nested structures and conditionals that characterize complex instructions similarly create *syntactic* issues. While it may be human practice to understand what has been directed without having to figure out exactly the *meaning* of the words, this doesn't translate across in machine coding. *Coding cannot capture context*…. here, the information in *adjacent* instructions and from *non-associated* routines. Instruction and analysis, after all, both use *different* categories of facts within their syntax; for instance, *indexical* facts, *normative* facts, strong convictions, observations and hints, clarifications, reinforcements as well as basic ontological factual statements. *All* of these inform the human decision. The challenge is also that such categorizations are *volatile* and change unpredictably according to new intelligence, feedback and input, of course, from weapon sensors.

All of this therefore requires arbitration. And arbitration can't simply be based on recency heuristic. What, after all, constitutes 'acceptable' delay? What for you as local commander represents urgency? Arbitration strategies are generally challenging. Averaging protocols that might select weapon actions according to where the *most* conditions have been satisfied (also referred to as 'longest matching') will similarly be inappropriate as, by definition, they are *approximations* and must reduce data precision. Two challenges therefore arise. The first relates to the *degree* of stepped change - the increment of learning termed 'anchoring' - that each follow-on process exerts on the weapon's immediately prior set of beliefs. A second challenge is that weapon actions must comprise the appropriate reaction to *every* relevant sensed stimulus. The weapon cannot offer patchy or erratic performance. To this point, coding for weapon 'curiosity' might be a composite of the two states of 'novelty' and 'attraction' but

each of these states has a specific and often conflicting action routine in the weapon. Thus, programming for 'astonishment' might be the combination of, say, 1/3rd 'attraction' – go forward - , 1/3rd 'withdrawal' – go backwards - and 1/3rd 'curiosity' – stay still and wait to garner additional information. This is also the case with the weapon's reward system. So how does weapon coding provide for the notion of a *delayed* response, a *measured* response, a slight *deferral* while additional information is sought or, more complicated, a *variable* response?

Current programming is particularly compromised when dealing with *ambiguity*. Human communication is not just about sending a message that can be recovered and enacted upon by the receiver. There is usually a yawning gulf between the spoken or written word and an intended message. First, *lexical* ambiguities tend to concern omitted or imprecise script. *Semantic* ambiguities concern interpretative uncertainty while *pragmatic* ambiguity hobbles communicating parties with mildly different contextual bases. All of these challenges raise basic ethical issues. Are 'occasional mistakes' acceptable on the basis that unsupervised violence is similarly susceptible to ethical deficiency as human soldiers? And is a lower legal bar appropriate for machines with *less* lethality?

A further coding difficulty arises from the *firing sequence* for such weapon instructions, both for rules-based but also for neural network-based AI models. Each routine (and the pattern created by those routines) will result in quite different outputs being enacted depending upon the *order* in which instructions are processed by the unsupervised weapon. This then requires weighted average or 'centre of gravity' routines intended to *optimise* action selection but leaving

unresolved a key limitation of machine learning that symbols are read one at a time and, moreover, with each such symbol being processed on information collected from previous symbols.

Indeed, it is widely acknowledged that algorithmic tools become *less* useful as uncertainty grows. Uncertainty, of course, is everywhere: The wheels of your weapons may spin, enemy forces dissemble and battlefield obstacles move unpredictably. It is here when *human* behaviours are required that can leverage experience, judgement and intuition. And it is exactly *here* where algorithms fall short of *you*, the human expert, who is instead able to make difficult decisions in a fast, frugal but *predictable* manner. The point is that at such higher levels of expertise, battlefield commanders do not even recognise that they are making decisions; rather, you are interacting with a changing situation and responding to patterns that you've long recognised. Training, experience, subjectivity and a wide grasp of context are all critical attributes that cannot currently be captured by code.

SLIDE 11 – DETERMINING WEAPON ACTIONS

In deciding to deploy weapons without human supervision, you also need to understand what will drive those weapon *actions*. Weapon sequences will be governed by what is termed an 'optimality notion', each unique to a weapon class and set by the Delivery Cohort as its initial set of decision rules. At each such step, the weapon is selecting an action with the *highest expected* utility. But various challenges arise from this utility model. The margin for error is very considerable and, in the independent weapon, obviously comes with the possibility of very high-

regret outcomes. To find the action with the *highest* expected utility, the weapon must first run an internal computation on *all* possible actions, a considerable task given the almost limitless number of battlefield parameters including legal status and risks arising from poor execution. The utility function must also regulate what constitutes *appropriate* use of force, the involvement of colleague assets, consideration of next steps, a data audit *ahead* of an action sequence as well as post-event communication of each engagement. You of course do this innately.

Similarly intractable is the Cohort's need to set *goals* to govern weapon priorities and action selection. Here, as leaders, you need to distinguish between values and goals in machine autonomy: Goals, I argue, prompt an intelligent weapon to *develop* plans of action while values enable it to *assess* the comparative *merits* of such plans. If this process is stunted or inappropriately undertaken, the weapon will either be illegal or useless. Goals concern what must be undertaken at once, what should be undertaken next, the resumption of a task that was previously discontinued and, more complex for the weapon, what actions should *subsequently* take place in order to *capitalize* on battlefield opportunities. Errors here may also have quite unforeseen battlefield consequences that include 'infrastructure profusion' where a machine unexpectedly allocates disproportionately large parts of its reachable resources into the service of some inappropriate internal goal.

What then is the theory in goal setting? The difference between a weapon's current and desired states will be the weapon's observed error. The goal of the action selection is thus to

minimize that error. Two issues arise for you. The *pace* of this error correction is not obvious. It depends, for instance, on how often the error is computed and how much correction is then made on each feedback loop. Loops, I should add, are significantly less adroit at modifying a weapon's *higher-level* action selection and here I refer to navigation, longer-term coordination and collaboration. Second, the weapon must also be *front-facing* and determine its system state based on sub-goals set for itself *ahead of time*. For this reason, weapon systems must be able to support *parallelism*, the complex ability to monitor and execute multiple actions at once. Sequential processing would otherwise risk missing events that might be critical to compliant and *winning* combat operation.

Lastly, goals and behaviours similarly require the weapon has a workable process of 'data forgetting' if the platform's behaviour is not to be compromised either by information overflow or the retention of sub-optimal (or wrong) data sequences. Very little progress that is relevant to weapons is evident here. The weapon, after all, should *not* generally forget *acquired* skills, termed 'catastrophic forgetting', a well recognised limitation of the neural network model. The issue, moreover, has several levels. How, for instance, should data *age* conflate with data *redundancy*?

I want in my last few minutes to turn to one further action precept, that of *attention* and the challenges of weapon focus and application. The human brain appears to be free to choose what it looks at, listens to and thinks about. In benign conditions, humans can focus their attention as they please. This drift of information, however, if not limited in any way, would lead

to memory overflow and to what Haikonnen terms in a machine as 'contradictory neural cacophony'. Similarly, the autonomous weapon must *actively* select both the *source and quantity* of its information, what to process, what to store and which *peripheral* information should be attenuated into its decision processes, the so-called 'cocktail party effect'. This is not obvious. Similar to cognition, attention is divided into voluntary and involuntary and no proven *statistical* model yet exists to determine that one such sensor input be preferred over others. The issue really distils into how the weapon's input *intensities* should be managed. After all, two variables that may be useless by themselves can be useful *together*. Similarly, a *single* variable that is useless by itself can then be instrumental *with* others. Nor can it be that the *loudest* signal is automatically the one upon which the weapon should focus. In this instance, the twin phenomena of data *sensitivity* and data *habituation* will be enduringly irksome to programmers in your Delivery Cohort.

SLIDE 12 – TECHNICAL DEBT AND HUMAN INTERVENTIONS

All of the foregoing evidences the issue of *technical debt* that lurks in autonomous weapon design. The notion of technical debt provides a useful metaphor that anchors much of this talk. It links the consequences of poor system design to accumulating a 'financial debt', the assertion here being that such 'debt' is particularly prevalent in the removal of battlefield supervision. Causes here of technical debt should concern you and arise from inappropriate architecture, from shortcuts resulting from commercial pressures, poor testing protocols and, more broadly, from poor *whole-system* understanding. It will arise from an overall lack of

ownership, from poor technical leadership and the consequences of pervasive specification changes. Debt here is also a consequence of 'counterparty development' where a weapon's disparate software routines, once developed, must eventually be merged into a single source base. *Scale* compounds technical debt both through the number of interactions *and* the number of interdependencies required among developers and those charged with deploying these weapons. You also need to be aware that such debt compounds as projects *evolve*, including management of the weapon's configuration, its integration, its verification and validation (a catch-all expression for its testing) and, in the case of machine learning, determining its 'logical completeness'.

Correction routines also deserve brief mention given that autonomous weapons are just that, weapons that are unsupervised and where remediation is correspondingly difficult. They also require that the correcting party factor both the *total* but also the *distribution* of system error. Error cascades are also generally prone to *deadlock* whereby the *local* optimum for the learning system quickly becomes *circular* and *iterative* so that neither the weapon component nor its attached routine can then be improved. This is also not helped by the weapon being a bundling together of disparate proprietary routines held together by 'glue' or 'spaghetti' code. Glue here relates to the quantity of *supporting* code that must be incorporated to allow data transfer within these routines. The phenomenon increases fragility, not least because glue code anchors the weapon to the idiosyncrasies and initial states of *each* autonomous component and so discourages experimentation, the intended essence of machine learning. The problem is well framed by Sculley's research that machine learning systems end up being five per cent executive

code and ninety-five per cent glue code.

My final observation relates to the *configuration* of unsupervised weapons. Configuration routines are by their nature ephemeral, less tested and, for you as leaders, a further source of poor predictability. There are several layers to this. These thresholds must be overseen by responsible, experienced humans. Indeed, part of leadership is that you *own* this process. Weapon configuration will be inherently complex and quite unsuited to any in-field, ad hoc adjustment where soldier ingenuity and experience have traditionally been relevant. If a weapon updates on new data, then its old manually-set thresholds may be invalid, the more so given your several weapons' different learning states. This also needs to be undertaken in real-time across the *whole* of the weapon system in order for that weapon to be both compliant and still valuable. The leadership challenge, moreover, is what *metrics* should be monitored given that a purpose of its machine-learning is actually to adapt overtime. Finally to this point, legal, social and political constraints will agitate that such limits are set *conservatively* which, should an action limit unexpectedly trigger and the weapon close down, will compromise its operational usefulness to the Delivery Cohort. The issue here is that weapon *sub*-systems then become '*undeclared consumers*', consuming the *output* of a particular prediction as *input* to another component of that sequence. As noted by UNIDIR, unintended feedback loops then form between weapon algorithms and the weapon's external world. Think of these loops as filter bubbles in social networks whereby noise suppression mechanisms inadvertently suppress nonconforming data. For a leadership angle, the phenomenon actually makes it problematic to make *any* changes to the weapon's firmware.

SLIDE 13 – CONCLUSIONS

My time is up but I hope that this has been a useful overview of these weapons' higher-order and technical pitfalls. And I think that a somewhat social science perspective is also useful in confirming their *enduring* persistence in the models that are currently posited for removing supervision. I bring you back to Sabin's Revolution in Expectation and the matter of feasibility. But I would also conclude by noting that leadership here is once again about *context*, the subject perhaps for a whole separate talk, and that the key leadership point to arise here is to question how useful wide-task, wide-capability autonomy *really* might be given the very many levers and assets, human assets, available to you as leaders to achieve aims and objectives.

Thank you.

SLIDE 14 – THANK YOU

SLIDE 15 - QUESTIONS

17/6 FINAL