Canonical instabilities of autonomous weapons under fog-of-war constraints: extending the Data Rate Theorem to cognitive systems

Rodrick Wallace

Canonical instabilities of autonomous weapons under fog-of-war constraints: extending the Data Rate Theorem to cognitive systems

Rodrick Wallace
Division of Epidemiology
The New York State Psychiatric Institute [*]

April 15, 2016

Abstract

Under fog-of-war conditions, skilled human operators of complex man-machine cockpit weapon systems have repeatedly been unable to avoid fratricide incidents and violations of the laws of land warfare. In spite of these continuing catastrophic failures, autonomous weapons have been proposed that entirely remove the man from the command loop. Formal analysis, via an extension of the Data Rate Theorem, shows that fully autonomous weapons under constraints on incoming environmental information will be subject to sudden, highly punctuated collapse to a dysfunctional ground state in which 'all possible targets are enemies'.

Key Words: control theory; fratricide; information theory; laws of land warfare; Rate Distortion Theorem; spontaneous symmetry breaking

1 Introduction

Autonomous real-time systems are widely proposed for both civilian and military enterprises. Since '94% of fatal traffic accidents are driver-related', the mass media argument goes, 'eliminating the driver from the loop will decrease traffic deaths by more than nine tenths'. This is eerily reminiscent of the famous beuracratic old saw that 'if a woman can produce a baby in nine months, nine women should be able to do it in a month'. In similar fashion, a rush-to-judgement argues that autonomous weapon systems can me made that adhere to the laws of land warfare, in spite of the infamous Patriot missile fratricides during the first Gulf war, and the manifold ongoing drone war catastrophies in

[*] Wallace@nyspi.columbia.edu, rodrick.wallace@gmail.com

which cognitive man-machine ‘cockpit’ systems were unable to operate successfully under fog-of-war conditions (Hawley 2006; Scharre 2016; Columbia 2012; Stanford/NYU 2012).

Here, we will extend the Data Rate Theorem that links control and information theories (Nair et al. 2007) to real time cognitive systems under fog-of-war constraints, highlighting the inevitability of their highly punctuated catastrophic failure, rather than ‘graceful degradation under stress’.

2 The Data Rate Theorem

The Data Rate Theorem (Nair et al. 2007) establishes the minimum rate at which externally-supplied control information must be provided for an inherently unstable system to maintain stability. Assuming a linear expansion near a nonequilibrium steady state, an n-dimensional vector of system parameters at time t, x_t, determines the state at time $t+1$ according to the model of figure 1, so that

$$x_{t+1} = \mathbf{A}x_t + \mathbf{B}u_t + W_t \tag{1}$$

where $\mathbf{A}, \mathbf{B}$ are fixed $n \times n$ matrices, u_t is the vector of control information, and W_t is an n-dimensional vector of white noise. The Data Rate Theorem (DRT) under such conditions states that the minimum control information rate $\mathcal{H}$ is determined by the relation

$$\mathcal{H} > \log[|\det[\mathbf{A}^m|] \tag{2}$$

where, for $m \leq n$, $\mathbf{A}^m$ is the subcomponent of $\mathbf{A}$ having eigenvalues ≥ 1. The right hand side of Eq.(2) is interpreted as the rate at which the system generates ‘topological information’. The proof of Eq.(2) is not particularly straightforward (Nair et al. 2007), and, via expansion of a simple correspondence reduction, we will use the Rate Distortion Theorem (RDT) to derive a more general version of the DRT.

3 An RDT approach to the DRT

The RDT asks how much a signal can be compressed and have average distortion, according to an appropriate measure, less than some predetermined limit D. The result is an expression for the minimum necessary channel capacity, R, as a function of D. See Cover and Thomas (2006) for details of the proof, and the Mathematical Appendix for a statement of the theorem. Different channels have different expressions. For the Gaussian channel under the squared distortion measure,

$$\begin{aligned} R(D) &= \frac{1}{2}\log[\frac{\sigma^2}{D}] \; D < \sigma^2 \\ R(D) &= 0 \; D \geq \sigma^2 \end{aligned} \tag{3}$$

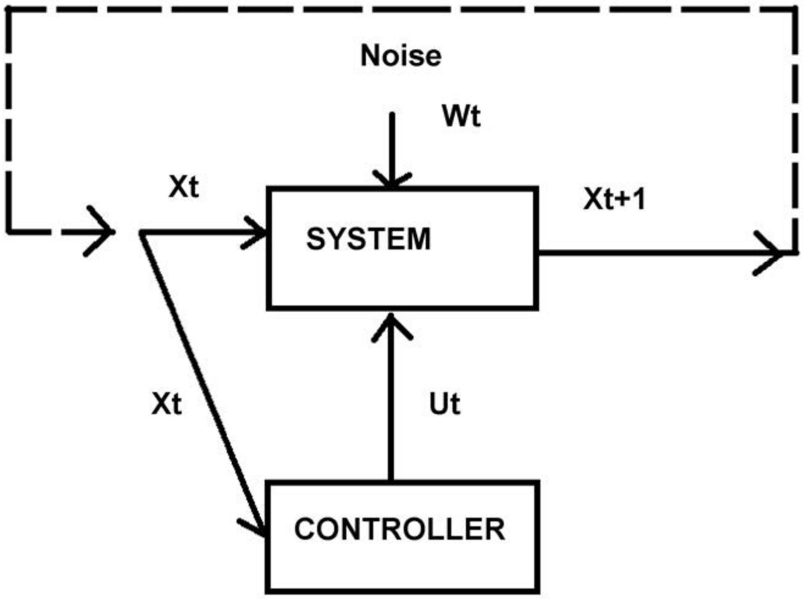

Figure 1: A linear expansion near a nonequilibrium steady state of an inherently unstable control system, for which $x_{t+1} = \mathbf{A}x_t + \mathbf{B}u_t + W_t$. $\mathbf{A}, \mathbf{B}$ are square matrices, x_t the vector of system parameters at time t, u_t the control vector at time t, and W_t a white noise vector. The Data Rate Theorem states that the minimum rate at which control information must be provided for system stability is $\mathcal{H} > \log[|\det[\mathbf{A}^m|]$, where $\mathbf{A}^m$ is the subcomponent of $\mathbf{A}$ having eigenvalues ≥ 1. For an autonomous cognitive system, the control signal u_t becomes the incoming 'target' data, in a large sense, which may be highly interactive for driverless cars and weapons closing on evasive subjects.

where σ^2 is the variance of channel noise having zero mean.

Our concern is how a control signal u_t is expressed in the system response x_{t+1}. We suppose it possible to deterministically retranslate an observed sequence of system outputs $x_1, x_2, x_3, ...$ into a sequence of possible control signals $\hat{u}_0, \hat{u}_1, ...$ and to compare that sequence with the original control sequence $u_0, u_1, ...$, with the difference between them having a particular value under the chosen distortion measure, and hence an observed average distortion.

The correspondence expansion is as follows.

Feynman (2000), expanding on ideas of Bennett, identifies information as a form of free energy. Thus $R(D)$, the minimum channel capacity necessary for average distortion D, is also a free energy measure, and we may define an entropy S as

$$S \equiv R(D) - DdR/dD \tag{4}$$

For a Gaussian channel under the squared distortion measure,

$$S = 1/2 \log[\sigma^2/D] + 1/2 \tag{5}$$

Other channels will have different expressions.

The simplest dynamics of such a system are given by a nonequilibrium Onsager equation in the gradient of S, (de Groot and Mazur 1984) so that

$$dD/dt = -\mu dS/dD = \frac{\mu}{2D} \tag{6}$$

By inspection,

$$D(t) = \sqrt{\mu t} \tag{7}$$

which is the classic outcome of the diffusion equation. For the 'natural' channel having $R(D) \propto 1/D$, $D(t) \propto$ the cube root of t.

This correspondence reduction allows an expansion to more complicated systems, in particular, to the control system of figure 1.

Let $\mathcal{H}$ be the rate at which control information is fed into an inherently unstable control system, in the presence of a further source of control system noise β, in addition to the channel noise defined by σ^2. The simplest generalization of Eq.(6), for a Gaussian channel, is the stochastic differential equation

$$dD_t = [\frac{\mu}{2D_t} - F(\mathcal{H})]dt + \beta D_t dW_t \tag{8}$$

where dW_t represents white noise and $F(\mathcal{H}) \geq 0$ is a monotonically increasing function.

This equation has the nonequilibrium steady state expectation

$$D_{nss} = \frac{\mu}{2F(\mathcal{H})} \tag{9}$$

measuring the average distortion between what the control system wants and what it gets. In a sense, this is a kind of converse to the famous radar equation which states that a returned signal will be proportional to the inverse fourth power of the distance between the transmitter and the target. But there is an even deeper result to be found.

Applying the Ito chain rule to Eq.(8) (Protter 1990; Khashminskii 2012), it is possible to calculate the expected variance in the distortion as $E(D_t^2)-(E(D_t))^2$. But application of the Ito rule to D_t^2 shows that *no real number solution for its expectation is possible unless the discriminant of the resulting quadratic equation is* ≥ 0, so that a necessary condition for stability is

$$\begin{aligned} F(\mathcal{H}) &\geq \beta\sqrt{\mu} \\ \mathcal{H} &\geq F^{-1}(\beta\sqrt{\mu}) \end{aligned} \tag{10}$$

where the second expression follows from the monotonicity of F.

As a consequence of the correspondence reduction leading to Eq.(7), we have generalized the DRT of Eq.(2). Different 'control channels', with different forms of $R(D)$, will give different detailed expressions for the rate of generation of 'topological information' by an inherently unstable system.

4 A DRT for autonomous cognitive systems

A deeper approach to the dynamics of real-time autonomous systems is via the 'cognitive paradigm' of Atlan and Cohen (1998), who recognized that the immune response is not simply an automatic reflex, but involves active choice of a particular response from a larger repertoire of those possible to it. Such choice reduces uncertainty in a formal manner, and implies the existence of an information source. See Wallace (2012, 2015) for details.

Given an information source associated with a rapid-fire autonomous system – characterized as 'dual' to it – an equivalence class algebra can be constructed by choosing different system origin states a_0 and defining the equivalence of two subsequent states at times $m, n > 0$, written as a_m, a_n, by the existence of high-probability meaningful paths connecting them to the same origin point. Disjoint partition by equivalence class, essentially similar to orbit equivalence classes in dynamical systems, defines a symmetry groupoid associated with the cognitive process. Groupoids represent generalizations of the group concept in which there is not necessarily a product defined for each possible element pair (Weinstein 1996). The simplest example would be a disjoint union of groups.

The equivalence classes define a set of cognitive dual information sources available to the autonomous system, creating a large groupoid, with each orbit corresponding to a transitive groupoid whose disjoint union is the full groupoid. Each subgroupoid is associated with its own dual information source, and larger groupoids will have richer dual information sources than smaller.

Let X_{G_i} be autonomous system's dual information source associated with the groupoid element G_i, and let $\mathcal{H}$ be the 'control' information rate associated with incoming environmental signals, in a large sense. Wallace (2012, 2015) shows how environmental regularities – road conditions, 'target' characteristics or evasive behavior – imply the existence of an environmental information source.

We can construct a Morse Function (Pettini 2007) as follows. Let $H(X_{G_i}) \equiv H_{G_i}$ be the uncertainty of the dual information source of the cognitive autonomous system. Define a Boltzmann-like pseudoprobability as

$$P[H_{G_i}] = \frac{\exp[-H_{G_i}/\kappa\mathcal{H}]}{\sum_j \exp[-H_{G_j}/\kappa\mathcal{H}]} \tag{11}$$

where κ is an appropriate constant depending on the particular system and its linkages to incoming information from the embedding environment that acts here as a 'control' signal for a DRT-like analysis, and the sum is over the different possible cognitive modes of the full system.

A 'free energy' Morse Function F can be defined as

$$\exp[-F/\kappa\mathcal{H}] \equiv \sum_j \exp[-H_{G_j}/\kappa\mathcal{H}] \tag{12}$$

Given the inherent groupoid structure, it is possible to apply an extension of Landau's picture of phase transition (Pettini 2007). In Landau's formulation of spontaneous symmetry breaking, phase transitions driven by temperature changes occur as alteration of system symmetry, with higher energies at higher temperatures being more symmetric. The shift between symmetries is highly punctuated in the temperature index, here the minimum necessary 'control' information rate $\mathcal{H}$ taken from the embedding environment. Typically, such arguments involve only a very limited number of possible phases.

Decline in $\mathcal{H}$, or in the ability of that incoming information to influence the system as characterized by κ, can lead to punctuated decline in the complexity of cognitive process possible to the autonomous system, driving it, in a highly punctuated 'symmetry-breaking' manner, to a ground state collapse in which, as with the Patriot missile fratricides and drone war massacres, 'all possible targets are enemies'. Indeed, these were 'best case' scenarios where skilled human operators had ultimate control.

Fog-of-war conditions are not like playing a slow-paced game of Go.

5 Discussion and conclusions

The kernel of the argument is that combat is an inherently unstable dynamic relation between competing cognitive entities. For autonomous weapons, information from the embedding adversarial environment acts as a control signal, u_t in the sense of figure 1, permitting a sequence of mid-course corrections enabling completion of the system's assigned function. Experience finds that high-end man-machine cockpit weapons systems repeatedly and persistently fail to avoid fratricide or violations of the laws of land warfare. The formal analysis here shows that autonomous weapons in a rapidly-changing environment – a fog-of-war limiting the magnitude of $\mathcal{H}$ – will not gracefully degrade under stress, but instead will display sudden punctuated collapse onto a dysfunctional, indeed murderous, ground state unable to differentiate friend or the innocent from foe.

It has been said that ‘the language of business is the language of dreams’. The assertion that autonomous weapons will avoid significant catastrophic mistargeting is a business-driven fantasy. A political leadership that embraces dreams in the conduct of war might well find itself embroiled in persistent nightmares of war crimes accusations and prosecutions.

Caveat Emptor

6 Mathematical Appendix: The RDT

Suppose a sequence of signals is generated by an information source Y having output $y^n = y_1, y_2, ...$ This is ‘digitized’ in terms of the observed behavior of the system with which it communicates, for example a sequence of ‘observed behaviors’ $b^n = b_1, b_2, ...$ Assume each b^n is then deterministically retranslated back into a reproduction of the original signal, $b^n \rightarrow \hat{y}^n = \hat{y}_1, \hat{y}_2, ...$

Define a distortion measure $d(y, \hat{y})$ comparing the original to the retranslated path. Many distortion measures are possible. For example, the Hamming distortion is defined simply as $d(y, \hat{y}) = 1, y \neq \hat{y}, d(y, \hat{y}) = 0, y = \hat{y}$. For continuous variates, the squared error distortion measure is $d(y, \hat{y}) = (y - \hat{y})^2$.

The distortion between paths y^n and $\hat{y}^n$ is

$$d(y^n, \hat{y}^n) \equiv \frac{1}{n} \sum_{j=1}^{n} d(y_j, \hat{y}_j) \tag{13}$$

The central characteristic of the Rate Distortion Theorem is that the basic result is independent of the exact distortion measure chosen (Cover and Thomas 2006).

Suppose that with each path y^n and b^n-path retranslation into the y-language, denoted $\hat{y}^n$, there are associated individual, joint, and conditional probability distributions $p(y^n), p(\hat{y}^n), p(y^n, \hat{y}^n), p(y^n|\hat{y}^n)$.

The average distortion is defined as

$$D \equiv \sum_{y^n} p(y^n) d(y^n, \hat{y}^n) \tag{14}$$

It is possible to define the information transmitted from the Y to the $\hat{Y}$ process using the Shannon source uncertainty of the strings:

$$I(Y, \hat{Y}) \equiv H(Y) - H(Y|\hat{Y}) = H(Y) + H(\hat{Y}) - H(Y, \hat{Y}) \tag{15}$$

where $H(..., ...)$ is the joint, and $H(...|...)$ the conditional, Shannon uncertainties (Cover and Thomas 2006).

If there is no uncertainty in Y given the retranslation $\hat{Y}$, then no information is lost, and the systems are perfectly synchronous.

This will almost never be true.

The rate distortion function $R(D)$ for a source Y with a distortion measure $d(y, \hat{y})$ is defined as

$$R(D) = \min_{p(y,\hat{y}); \sum_{(y,\hat{y})} p(y)p(y|\hat{y})d(y,\hat{y}) \leq D} I(Y, \hat{Y}) \tag{16}$$

Thus the minimization is over all conditional distributions $p(y|\hat{y})$ for which the joint distribution $p(y, \hat{y}) = p(y)p(y|\hat{y})$ satisfies the constraint of having average distortion $\leq D$.

The Rate Distortion Theorem states that $R(D)$ is the minimum necessary rate of information transmission which ensures the communication between the interacting entities does not exceed average distortion D. Thus $R(D)$ defines a minimum necessary channel capacity. Cover and Thomas (2006) provide details. The rate distortion function has been calculated for a number of systems, often using Lagrange multiplier or Kuhn-Tucker optimization methods. Cover and Thomas (2006, Lemma 13.4.1) show that $R(D)$ is necessarily a decreasing convex function of D for any reasonable definition of distortion. That is, $R(D)$ is always a reverse J-shaped curve. This is crucial: convexity is an exceedingly powerful mathematical condition, and permits deep mathematical inference on system dynamics.

References

Atlan, H., I. Cohen, 1998, Immune information, self-organization and meaning, International Immunology, 10:711-717.

Columbia University Law School Human Rights Clinic, 2012, Counting Drone Strike Deaths, http://web.law.columbia.edu/human-rights-institute.

Cover, T., J. Thomas, 2006, Elements of Information Theory, Wiley, New York.

de Groot, S., P. Mazur, 1984, Non-Equilibrium Thermodynamics, Dover Publications, New York.

Feynman, R., 2000, Lectures on Computation, Westview Press, Boulder, CO.

Hawley, J., 2006, Patriot Fratricides: The human dimension lessions of Operation Iraqui Freedom, Field Artillery, January-February.

Khasminskii, R., 2012, Stochastic Stability of Differential Equations, 2nd Edition, Springer, New York.

Nair, G., F. Fagnani, S. Zampieri, R. Evans, 2007, Feedback control under data rate constraints: an overview, Proceedings of the IEEE 95:108-137.

Pettini, M., 2007, Geometry and Topology in Hamiltonian Dynamics and Statistical Mechanics, Springer, New York.

Protter, 1990, Stochastic Integration and Differential Equations, Springer, New York.

Scharre, P., 2016, Autonomous weapons and operational risk, Center for New American Security, Washington, D.C.

http://www.cnas.org/autonomous-weapons-and-operational-risk.vtISHGO-RiY

Stanford/NYU, 2012, Living under drones:death, injury and trauma to civilians from US drone practices in Pakistan,

http://livingunderdrones.org/

Wallace, R., 2012, Consciousness, crosstalk, and the mereological fallacy: an evolutionary perspective, Physics of Life Reviews, 9:426-453.

Wallace, R., 2015, An Information Approach to Mitochondrial Dysfunction, World Scientific, Singapore.

Weinstein, A., 1996, Groupoids: unifying internal and external symmetry, Notices of the American Mathematical Association, 43:744-752.