Empirical Information, Potential Information, and Disinformation as Signatures of distinct Classes of Information Evolving Machines

Christopher Lee

Departments of Computer Science, Chemistry and Biochemistry, UCLA

Translating Information Theory to Information Production

- "God's View": information theory metric I(X;Y,Z) assumes knowledge of joint distribution of all system variables p(X,Y,Z)
- "data scientist's view": by contrast, in science, that is what we are trying to figure out, ideally through an efficient sampling and statistical inference process.
- "a bug's view": intuitively a bacterium seems to contain "information", but the exact mapping to classic information theory is non-trivial.

Statistical inference provides a nice bridge between "God's View" and a "bug's view", where we can look carefully at the process of information production.

Replication Factor Sampling

- say we have a population of elements $G \in \{g_1, g_2, ...\}$ with initial population values $\{p_1, p_2, ...\}$, that undergo a sequence of replication rounds each with replication factors $\{f_1(t), f_2(t), ...\}$. Then after time steps $t_1, t_2, ...t_n$ the population for g_i will be:

$$p_i(t_n) = p_i(t_0)f_1(t_1)f_1(t_2)...f_1(t_n)$$

- say for a given g_i, the $f_i(t)$ are drawn from the same distribution ("independent and identically distributed", IID). If we want an "average" for this process with a guaranteed convergence, we must convert the product into a sum via $L = \log F$. Then the Law of Large Numbers guarantees the sample average

$$\overline{L(g_i)} = \frac{1}{n}[\log f_i(t_1) + \log f_i(t_2) + ... + \log f_i(t_n)] \rightarrow E(L(g_i)) \text{ in probability}$$

Example: Bayesian Inference

- define our elements to be likelihood models $\psi_i(X) \equiv p(X|\psi_i)$ for some observable X.
- now draw an IID sample $\vec{X}^n = \{X_1, X_2, ... X_n\}$.
- define "replication factors" $f_i(t_j) = p(X_j|\psi_i)$
- then the posterior probability of a model given the observations is

$$p(\psi_i|\vec{X}^n) = \frac{p(\psi_i)p(\vec{X}^n|\psi_i)}{\sum_j p(\psi_j)p(\vec{X}^n|\psi_j)} = \frac{p(\psi_i)e^{n\overline{L}(\psi_i)}}{\sum_j p(\psi_j)e^{n\overline{L}(\psi_j)}}$$

- the *empirical log-likelihoods* $\overline{L}(\psi_i)$ are a sufficient statistic for this process.
- define the *empirical information* as the increase (or decrease!) in prediction power for the observable achieved by model ψ relative to the marginal distribution $p(X)$:

$$\overline{I_e}(\psi) = \overline{L}(\psi) - \overline{L}(p)$$

Mutual Information: the Big Picture

- define Ω as the true (but unknown) distribution of X, with prior distribution $p(\Omega)$.
- for a specific inference problem $\Omega = \omega^*$, repeatedly draw samples $\vec{X}^t, \vec{X}^n$ from it, to quantify how much we learn about the observable X from a training sample $\vec{X}^t$.
- say we compute the equation for $I(\vec{X}^t; X)$ on these data, what will this give us?

Mutual Information: the Big, Big Picture

- by definition $I(\vec{X}^t;X|\Omega=\omega^*)=I(\vec{X}^t;X|\Omega)=0$. The $\vec{X}^t;X$ are "conditionally independent given Ω".
- the definition of mutual information only applies over the complete joint distribution of $p(\Omega,X)$, the "big, big picture" over all possible universes Ω.
- a single instance of inference $\Omega=\omega^*$ is a "little picture" of this specific universe ω^*.
- is there a way to estimate the "big picture" information $I(\vec{X}^t;X)$ from a "little picture" sample for a single instance of inference $\Omega=\omega^*$?

Empirical Information: the Little Picture

- Say we train an inference process Φ with training data $\vec{X}^t$, so its resulting prediction is $\Phi(X|\vec{X}^t)$.
- Consider the expectation of $\overline{I_e}$ over all possible $\Omega, \vec{X}^t, \vec{X}^n$:

$$E(\overline{I_e}(\Phi(X|\vec{X}^t))) = E(\overline{L}(\Phi(X|\vec{X}^t))) - E(\overline{L}(p(X)))$$

$$= H(X) - H(X|\vec{X}^t) - E_{\vec{X}^t}(D(p(X|\vec{X}^t)||\Phi(X|\vec{X}^t)))$$

$$= I(X;\vec{X}^t) - E_{\vec{X}^t}(D(p(X|\vec{X}^t)||\Phi(X|\vec{X}^t)))$$

  where $D(\omega||\psi) \geq 0$ is the *relative entropy*, and vanishes iff $\omega(X) = \psi(X)$ everywhere.
- *sampleable I*: we can measure I_e on any inference problem, and its average over many problems converges to mutual information.

Biological Evolution

- likelihood of the observations $p(X|\psi_i) \in [0,1] \rightarrow$ fitness $f_i(p(t)) \geq 0$, which can be greater than 1, and can change depending on the total population distribution *p(t)*.
- Bayes' Law $\rightarrow$ Discrete Replicator Equation (Harper 2009)

$$p_i(t_1) = \frac{p_i(t_0) f_i(t_1)}{\sum_j p_j(t_0) f_j(t_1)}$$

- analogous extension to "multiple observations" $\{t_1, t_2, ..., t_n\}$

$$p_i(t_n) = \frac{p_i(t_0) f_i(t_1) f_i(t_2) ... f_i(t_n)}{\sum_j p_j(t_0) f_j(t_1) f_j(t_2) ... f_j(t_n)} = \frac{p_i(t_0) e^{n\overline{L(g_i)}}}{\sum_j p_j(t_0) e^{n\overline{L(g_j)}}}$$

- Again the empirical metrics $\overline{L}/\overline{I_e}$ are the sufficient statistic for this sampling process.

Local-Gradient Information Evolving Machine

- two basic ingredients:
  - *place your bets*: predictions of what's going to "replicate best";
  - *draw a test sample* of empirical replication factors.
- *hill climbing* (in the Fisher/Price sense): population tends to shift towards g_i with highest $\overline{f}, \overline{L}, \overline{I_e}$.
- although this model is general, for most systems, high rate of diffusion means information destroyed as fast as it's produced.
- strictly empirical, local process limited by finite sample of elements with $p(g_i) > 0$, which may well not include g^* with maximum fitness.
- "tunneling problem": if path to g^* poses an "activation barrier" (reduced fitness relative to current $\overline{f}$), transition rate to g^* will be exponentially slow.

Entropic Limits of IEMs: Empirical Information

- *model selection* is the process of finding model ψ^* that maximizes some metric $f(\vec{X}^t|\Psi = \psi^*) \geq f(\vec{X}^t|\Psi)$.
- Just "hill-climbing" by another name (local or global)?
- Many (e.g. in Computer Science) would regard this as a valid, general framework.
- *maximum likelihood* (ML) is equivalent to using $I_e(\Psi)$ as the model selection metric.
- Many other metrics (e.g. k-means) can be shown to be equivalent to this metric.
- Can explicit definition of IEM sampling process shed light on limitations of ML?

When is "Prediction Power" Predictive?

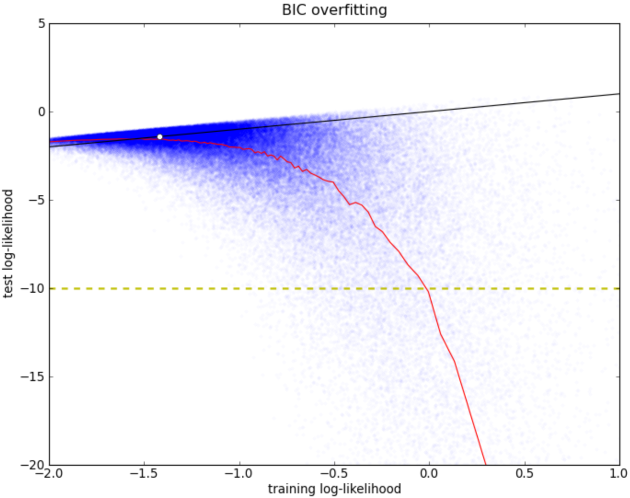
- the claim that $\overline{l}_e(\Psi)$ measures *future* prediction power depends on the Law of Large Numbers guarantee that for *one* model Ψ

$$p\left(|\overline{L}(\Psi) - E(L(\Psi))| \geq \sqrt{\frac{Var(L(\Psi))}{n\varepsilon}}\right) \leq \varepsilon$$

- but if we select from a huge effective number of models Ψ for maximum $\overline{L}(\Psi)$, we *expect* to find big deviations from $E(L(\Psi))$.
- ML "corrections" like the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) etc. help stave off some forms of this overfitting (number of degrees of freedom) but not others.

BIC Does Not Eliminate Overfitting

Draw three observations from a unit normal, choose $N(\mu^*, \sigma^*)$ with maximum BIC, measure L on training vs. test sample:



Optimal Inference

- What modeling procedure $\Phi(X|\vec{X}^t)$ will maximize our expectation prediction power? Recall

$$E(\overline{I_e}(\Phi(X|\vec{X}^t))) = I(X;\vec{X}^t) - E_{\vec{X}^t}(D(p(X|\vec{X}^t)||\Phi(X|\vec{X}^t)))$$

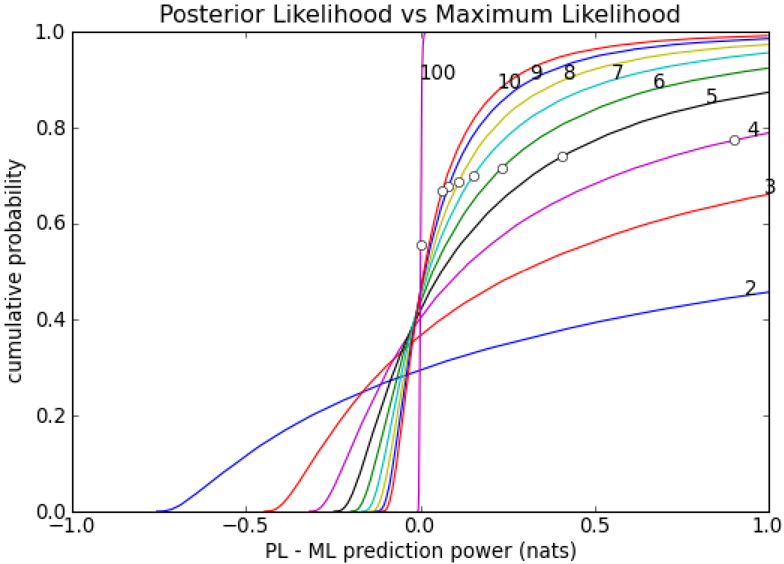
- maximized by the *posterior likelihood* (PL)

$$\Phi(X|\vec{X}^t) = p(X|\vec{X}^t) = \sum_{\phi} p(\phi|\vec{X}^t)p(X|\phi)$$

  where $p(\phi|\vec{X}^t)$ are just the posterior probabilities from Bayes' Law.

- any attempt to reduce entropy of Φ below that of the PL will *degrade* prediction power.

Better Prediction Power by Avoiding Overfitting



Training Data vs. Test Data

- The classic "last line of defense" against overfitting is to separate the data into *training data* (used to train the model) and *test data* (used to measure the model's prediction power).
- Is there a valid training + test procedure where *every* data point is used **both** to train the model *and* to test the model (i.e. compute an unbiased measure of the model's prediction power)?

Forward Log-Likelihood

- Yes, the posterior likelihood provides exactly this procedure: at each step of the observation process, it uses the previous $\vec{X}^i$ observations to predict the likelihood of the next observation X_{i+1}.
- define *forward log-likelihood* as

$$\overline{L_f}(\Phi) = \frac{1}{n}\sum_{i=i}^{n} \log \Phi(X_i|\vec{X}^{i-1})$$

- averages over the model's gradually improving prediction power and thus "lags" its true final prediction power.
- Note, by the chain rule this is a guaranteed valid way of factoring the total joint probability of the observations:

$$\log p(\vec{X}^n|\Phi) = n\overline{L_f}(\Phi)$$

Entropic Limits of IEMs: Potential Information

- Generally should assume IEM actual sample $\{\psi_i\}|p(\psi_i) > 0$ may not include true distribution ω.
- *Bayesian relativism*: in that case, computed $p(\psi_i|\vec{X}^t)$ will be off by unknown factor E, so strictly speaking all we can calculate accurately are *posterior odds* ratio $p(\psi_i|\vec{X}^t)/p(\psi_j|\vec{X}^t)$.
- However, the maximum prediction power possible for the observable can be estimated from its *empirical entropy* $\overline{H_e}$. So we define the *potential information* I_p as how far our current model Ψ falls short of that bound

$$\overline{I_p} = -\overline{H_e} - \overline{L}$$

  and more importantly get confidence bounds for our sample.
- Again, a Law of Large Numbers convergence guarantee

$$\overline{I_p} \to D(\omega||\Psi) \text{ in probability}$$

Empirical Density

- Concretely, "empirical entropy" means computing a model-free density estimator, via sampling

$$\overline{H_e} = \frac{1}{n}\sum_i -\log \rho_e(X_i) \to -E(\log \omega(X)) \text{ in probability}$$

- typically estimate $\rho_e(X)$ in higher dimensional spaces by measuring volume containing k-nearest neighbors (k-NN).
- Can a model ever exceed the prediction power of this empirical density $\rho_e(X)$?

Model Information

- No, for large $n \to \infty$ (assuming a good ρ_e estimator).

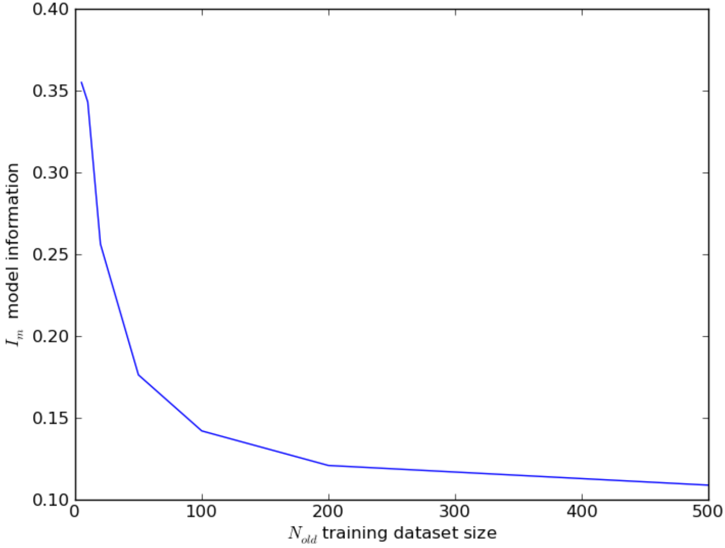
$$\overline{I_p} \to D(\omega || \Psi) \geq 0$$

- But initially, it is possible. Define the *model information* as

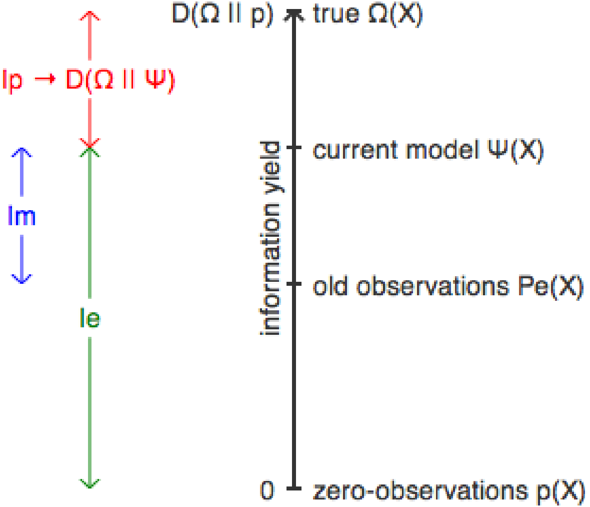
$$I_m(\Phi) = \frac{1}{n} \sum_i \log \frac{\Phi(X_i|\vec{X}^{i-1})}{\rho_e(X_i|\vec{X}^{i-1})} = \overline{L_f}(\Phi) - \overline{\log \rho_f}$$

- $I_m > 0$ is what we intuitively mean by "prediction value", i.e. the model tells us more than what we already knew empirically, via "better interpolation and extrapolation".

Model Information of the Normal PL



Ie, Ip, Im Partitioning



Example: Biological Sequences

- consider case where the observable X is itself a high-dimensional vector (e.g. protein sequence).
- naive k-NN empirical entropy estimator $\overline{H_e}$ based on short word length will outperform long word length (due to insertions / deletions not handled by naive estimator).
- identifies words with high potential information $\overline{I_p}(p) = -\overline{H_e} - \overline{L}(p)$ relative to marginal distribution *p(W)*.
- convert $\overline{I_p} \rightarrow \overline{I_e}$ by applying mutation model ψ.
- this is exactly what the standard BLAST algorithm actually computes, with the main difference that they don't call it $\overline{H_e}, \overline{I_p}, \overline{I_e}$.

"Ip Tunneling" Information Evolving Machine

- computes I_e just like local IEM.
- computes lower bound on I_p.
- if significant I_p detected, transition to "model extension" cycle where new model terms added either by search or interpolation / extrapolation.
- new model extensions given small initial prior, but can rise by better predicting subsequent observations.
- **Not** Model Selection, because extensions can only rise from "insignificant" to "dominant" on *independent* observations.
- has the potential to "tunnel" rapidly to true distribution ω via direct detection and modeling of I_p.
- of course, observed process may try to make I_p appear to be zero by pseudo-randomizing its output (encryption).

Disinformation

define *disinformation* as reduction in an IEM Φ 's *subsequent prediction power* due to sending it a *previous signal*.
Two part process:

- *initial deception signal*: Φ observes D (sent by attacker).
- *empirical exploit*: measure reduction in prediction power on subsequent observations (the *exploit sample*):

$$\overline{I_d} = \overline{I_e}(\Phi) - \overline{I_e}(\Phi|D)$$

Example: Bulk Mail Investment Con

- *target*: consider target IEM to be individual investor Φ predicting $p(X = win)$ for whether a particular investment will "win" (e.g. "stock market will go up next week").
- *initial deception signal*: starting with a large number of addresses (say 10,000), send half prediction that stock market will go up next week, other half down. Next week, reduce to addresses that got the correct prediction. Repeat 5 times.
- *exploit*: for the subset that received 5 correct predictions, now ask them to pay for your next prediction.
- *disinformation metric*: total money you receive, since your prediction actually has zero value. Note that disinformation for the target is information for the attacker.
- note: classic example of Model Selection

Example: Biological Species vs. Species

- *target*: set of all genotypes in one species Φ is a local-IEM.
- *attacker*: local-IEM for second species Δ.
- *initial deception signal*: say Δ is a non-poisonous butterfly species that mimics coloration of a poisonous butterfly species.
- *exploit*: say Φ is a bird species that predates on butterflies. The *exploit sample* will be its frequency of predation on species Δ (vs. non-mimic butterflies).
- *disinformation metric*: for local-IEM, I_e proxy is fitness. So disinformation is reduction in Φ 's fitness due to loss of a food source. Again, counts as increased *information* for Δ.

IEM vs. IEM Dynamic

- when payoffs for two IEMs in an evolutionary game are opposed, selection on "attacker" IEM Δ for increased empirical information, will induce disinformation on "target" IEM Φ (reduced fitness).
- in general, selection on set of possible "signal + exploit" behaviors will select for *control signals*, i.e. where Δ can most successfully induce a favorable (exploitable) target state.
- disinformation attack targeting a high fitness value feature τ will tend to induce many rounds of an arms race between Δ, Φ.
  - for Φ, strong negative selection for retaining the original function of τ, plus positive selection for escaping the specific disinformation attack.
  - for Δ, positive selection for beating the latest escape and regaining a successful disinformation attack.

What is the value of information in EGT?

- the best strategies in the field, e.g. TFT, Win-Stay-Lose-Shift, ZDR, are Markov-1; they respond *only* to the current game outcome, and remember no history.
- e.g. in historical tournaments, TFT beat all the more complicated strategies.
- ZD papers assert longer-term memory will *not* improve performance vs ZD players.
- Stewart & Plotkin (2013) offered a proof that generous ZD strategy (ZDR) is universally robust against all possible strategies as invaders -- if you accept their assumptions.

So information has no value?

Can "Information" Change Evolutionary Dynamics?

- EGT: first-order Markov ($p(X'|X)$) strategies such as Zero Determinant players are universally robust; players with longer memory or internal states (definition of an IEM) cannot help.
- i.e. whole population is a local-IEM, but no IEM vs. IEM dynamics.
- to investigate this claim, developed concept of an *Information Player*, where the individual player is itself an IEM: basic PL estimator of opponent Φ 's strategy vector $p(X'|X)$.
- tested IP_0 on classic EGT game (Prisoner's Dilemma: "Defect" vs. "Cooperate") vs. classic EGT strategies (e.g. TFT, WSLS, ZDR), but applicable generally.

IP0 Strategy: Infogain + Exploit

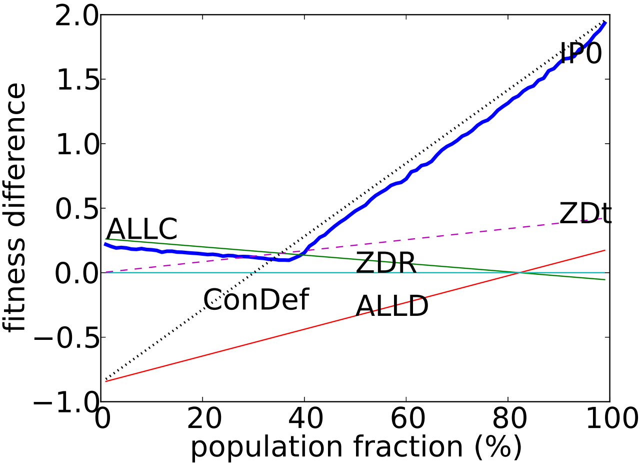
- two phase strategy
  - *infogain*: instead of maximizing score, maximize *information gain* by choosing move that will yield most information about opponent's strategy vector. Brief (10 moves).
  - *exploit*: choose optimal strategy vector against current inferred population mixture. Longer (average lifetime ~ 100 moves).
- "plays like me" = *self-recognition*: e.g. infogain "code" is highly distinguishable from Markov-1 strategies.
- NB: *infogain + exploit* meets our definition of a disinformation attack if opponent population Φ is a local-IEM.

The Power of Information?

"Information players" (IP) open two kinds of freedom that Markov-1 strategies lack:

- **self-recognition**: use a different strategy with yourself than with the competing type(s). E.g. Adami & Hintze (2013) outlined theoretical "tag player" with perfect knowledge of "type" of every player; cooperates with itself while defecting against "enemy" type (Conditional Defector, ConDef).
- **population sensitivity**: use different strategies with the competing type depending on your population fraction.

ZDR Resident



when IP_0 is in the minority, its optimal strategy is to play ALLC; when in the majority, switch to ConDef.

IP0 as Invader and Resident (Fixation Odds Ratio)

	IP0	ALLC	ALLD	TFT	WSLS	ZDR	ZDX
IP0		58.1	5.5	43.6	2.0	16.3	51.0
ALLC	0		0	49.5	0	21.1	54.8
ALLD	0	59.4		0	0.1	0	0
TFT	0	0	3.7		0	0	9.7
WSLS	0	34.7	0	7.1		0.3	21.2
ZDR	0	0	0.9	24.1	0		27.6
ZDX	0	0	1.6	0	0	0	

Values are $\rho/\rho_{neutral}$; donation game IPD, exp. imitation as in Stewart & Plotkin 2013. Assumptions: $\beta = 1, N = 100, i = 1, \varepsilon = 0.05$

- IP_0 was uninvadable (no successful invasions in 10,000 runs).
- IP_0 invades all ZIPs, generally better or almost as well as best existing invader.

Maximal Resident Advantage

- Define as the best lower-bound any resident strategy **R** can achieve vs. *all possible invaders* **I**:

$$(\overline{s_R - s_I})_{MRA} = \max_R \min_I (\overline{s_R - s_I})_{N,i}$$

- Information Player with perfect information (ConDef) achieves both maximal $s_{RR} = R$ (by cooperating with itself) and the best lower-bound interaction score $\min_I \Delta_{N,i}$ (because it plays ALLD vs. the invader). Hence it achieves MRA for $f < 0.5$, specifically

$$\min_I (\overline{s_R - s_I})_{N,i} \geq \frac{N-2i}{N-1}(R-P)$$

  since $s_{II} \leq R$ for all possible invading strategies **I**.

- IP_0 operates within a few percent of this theoretical limit, because it achieves that level of accuracy recognizing individual invaders.

MRA: The Transition from Coalition to Tyranny

- unique property of Information Players. Markov-1 cannot play different strategy with self vs. non-self, nor switch strategy based on population fraction.
- **dominance**: majority, favored for fixation. (Markov-1 do this)
- **tyranny**: not just above the opposition, but generic policy to **kill them all**.
- MRA is an objective obligation to tyranny.
- in the absence of MRA, even Satan himself should "play nice" (cooperate) sometimes, purely out of self-interest.
- one-party state: (MRA).
- two-party state: one side always has MRA.
- two-party state with strong independent block: can stay below MRA, if block is big enough, and independent enough.

The Case for Studying IEM Dynamics in EGT

- reveals new dynamics, e.g. frequency-dependent strategy optimization; MRA.
- many types of real-world players have IEM characteristics: e.g. species vs. species; brain vs. brain in biology; person vs. person.
- making disinformation an explicit subject for study seems like a ubiquitous and interesting phenomenon that EGT should come to grips with.
- But, low on the priority list for existing EGT researchers. So will only take off if people who share this interest coalesce as a group who will read & use each others' papers; collaborate etc.

Multiscale Information Evolving Machine

- define as IEM whose population elements are themselves IEMs, e.g. ecosystem of species.
- very different dynamic from *bounded information production* of simple IEM (e.g. statistical inference of a static target ω).
- E.g. strictly speaking, disinformation dynamic has no finite information bound.
- *information rate* is probably more important than *total bound*: here too, interaction dynamics such as disinformation seem like powerful "tunneling" mechanisms that reduce evolutionary leaps to a gradient of successive disinformation attacks (e.g. evolution of an immune system).
- is this all just a tedious replay of old arguments about "group selection" / *multilevel selection*?