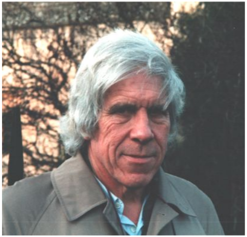
How to model the evolution of decision-making and individual preferences for social interactions

Laurent Lehmann

University of Lausanne

The starting point: the abstract of Hamilton (1964)

Species following the model should tend to evolve behavior such that each organism appears to be attempting to maximize its inclusive fitness.



W. D. Hamilton

- No proof given in the paper.
- There is no explicit formalization of behavior in the model.
- The paper contains important results and launched the field of social evolution theory.
- The paper also attracted a lot of criticism.

The question framing this course

How does one construct a compeling evolutionary model that can adress Hamilton's claim?

1. Formalize behavioral interactions and individuals.
2. Formalize evolution and adaptation.
3. Bring the two together and show in what sense Hamilton's claim holds.

In doing so we will go through many important concepts of social evolution theory. We will progress logically rather than historically.

Concept 1: action situation (or activity)

An action situation set of specific behaviours for a number of individuals and their associated outcomes that is mutually exclusive from another action situation.

Action situation

- Foraging.
- Escaping predators.
- Mating.
- Grooming.

- Action situations are generally recurrent.
- Action situations may or may not involve interactions with conspecific and we focus on those that do.

Action situation: three main components

1. A set of participating individuals.

2. A set of alternative feasible behaviours (action or stream of actions) to each such individual. Let's call the set of action $\mathcal{A}$ (assumed common to all individuals) and a particular action a.

3. A transformation (mapping) from the behaviours of individuals and the state of nature to some outcome to each participating individuals. It is convenient to call this outcome payoff.

Action situation: the volunteer's dilemma (or snow-drift game)

Action situation: the volunteer's dilemma (or snow-drift game)

Two player two action action-situation with $\mathcal{A} = \{\mathit{cooperate}, \mathit{defect}\}$ (or $\mathcal{A} = \{\mathit{contribute}, \mathit{shirk}\}$)

		Player 2	
		Cooperate	Defect
Player 1	Cooperate	B – C / 2	B – C
	Defect	B	0

Here B can be thought of as the benefit from escaping from the snow or be warned by a predator, while C is the cost of shoveling (or being caught by the predator).

Action situation and game theory

- More generally, action situations may involves streams of actions.
- By behavior we mean a realized stream of action and thus a description of what an individual does in any situation in which it may find itself.

Connection to game theory:

- This concept of behavior is equivalent to that of a "strategy".
- The concept of action situation is equivalent to "a game form".

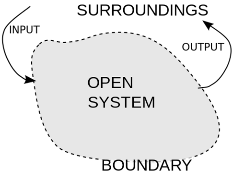
From game theory we know that whatever complex, we can conceptualise interactions as actions situations.

How do individuals take actions or express behavior?

- Individuals must have a biological mechanism to express behavior.
- The brain is the main organ in which information processing takes place and is the center of decision making in animals.
- How should we conceptualise decision-making?

Concept 2: behavior rules

All biological organisms are open systems exchanging energy, matter, and information with their surrounding.

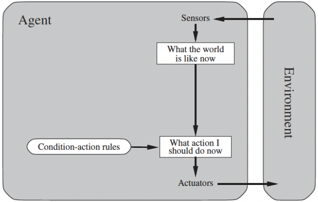


- Individuals thus act and react to each other and with their environment.
- Organism respond to changes of their internal factors as well as to changes in its external biotic and abiotic environmental conditions.

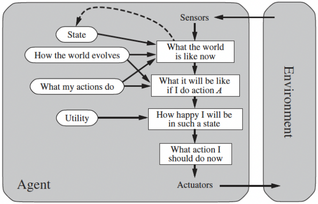
A behavior rule is a transformation (mapping) from the internal state of an individual and the (perceived) state of its surronding to some particular action.

Behavior rules and the computational sciences

Behavior rules range from the simple to the complex.



Reflex based agent


Sensors
State
What the world is like now
How the world evolves
What my actions do
What it will be like if I do action A
Utility
How happy I will be in such a state
What action I should do now
Agent
Actuators
Environment

Goal based agent

- The concept of a behavior rule should be thought equivalent to that of an algorithm or machine of the computational science (finite state machines, push-down automata, Turing machines).
- This is exactly how animal behavior is conceptualized and recurrent neural nets are Turing equivalent.

Behavior rules: pairs of interacting individuals

In equilibrium, we can regard the pair of actions $(a^*, b^*) \in \mathcal{A} \times \mathcal{A}$ expressed by two interacting individuals to be mutually interdependent in the sense that each individual responds to the other:

$$a^* = B(\mathcal{A}, b^*) \qquad b^* = B(\mathcal{A}, a^*)$$

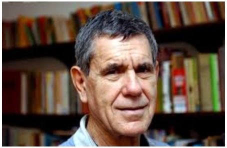
- Here $B(\mathcal{A}, x)$ is the behavior rule of an individual with action set $\mathcal{A}$ when its partner expresses action $x \in \{a, b\}$.
- The behavior rule thus maps the set of own actions and the partner action to one of own's action (formally a behavior rule is a correspondence).

We can now think of a feasible set of alternative behavior rules $\mathcal{B}$ and ask what type of rules are favored by evolution.

Concept 3: long-term evolution

- We can talk about Darwinian evolution only if we are prepared to consider different variants in the population and we just introduced the set of alternative behavior rules $\mathcal{B}$.
- Let us consider more generally that there is a set of heritable traits Θ.

The theory of long-term evolution aims to single out which trait in the trait set Θ will be favored by evolution. For this, we don't need to bother about genetic details.



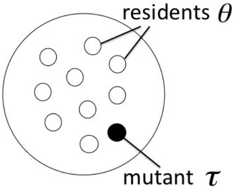
I. Eshel

Hence, we can consider that each trait $\tau \in \Theta$ is a replicator (or a clone).

Concept 4: resistance to invasion (or uninvadability)

A necessary condition for any trait (and thus behavior rule) to be observed in a population in the long-run is that it cannot be displaced by alternatives and thus needs to be resistant to invasion.

A resident trait θ is said to be uninvadable (or resistant to invasion) if it cannot be invaded by any alternative trait in the trait space $\tau \in \Theta$. i.e. all mutants go extinct with probability one.



One way to define the concept of adaptation is to equate it with uninvadability.

Concept 5: Lineage fitness

Let us now introduce the geometric growth ratio of a mutant τ in a resident θ population.

$$W(\tau, \theta) = \text{Lineage successful gene copy number per capita}$$

1. Demographic contexts: all possible physiological/life states.
2. Environmental contexts: all possible biotic and abiotic states.
3. Genetic contexts: all possible distribution of genotypes.

The theory of branching processes shows that type τ^* is uninvadable if

$$W(\tau, \tau^*) \leq 1 \quad \text{for all } \tau \in \Theta$$

This is a represenstation of Darwinian fitness.

The replicator as the unit of adaptation

An uninvadable trait τ^* thus maximizes the lineage fitness for a population expressing the uninvadable trait

$$\tau^* \in \arg\max_{\tau \in \Theta} W(\tau, \tau^*)$$

- Darwinian fitness is the gene's goal function and adaptation thus favor gene replication.
- Hence, we expect replicators "manipulate" the world and attempt to maximize their representation in the gene pool.

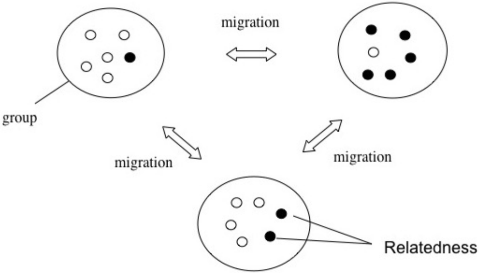
In this account, the replicator strives to maximize its survival and individuals and societies are byproducts.

Let's apply the machinery to pairwise interactions

- We now consider the evolution of behavior rules.
- We focus on pairwise interactions in a familly or group structured populations.

"Reading the theory of many-person games may seem to stand to that of two-person games in the relation of sea-sickness to a headache" (Hamilton 1975)

Concept 6: relatedness is induced by limited genetic mixing



Relatedness is defined as the probability *r* that two (homologous) genes randomly sampled in two different individuals in the same group (or familly) are identical-by-descent, i.e., they descend from a common ancestor.[1]

[1] Because the number of groups is large, the relatedness between individuals from two different is zero.

Within each family siblings face an action situation

Suppose two individuals, one with trait τ and the other with trait θ interact. Each individual is assumed to express some equilibrium action according to its behavior rule

$$a^* = B_\tau(\mathcal{A}, b^*) \qquad b^* = B_\theta(\mathcal{A}, a^*)$$

Because actions are interdependent we can write

$$a^*(\tau, \theta) \quad \text{and} \quad b^*(\theta, \tau)$$

- This emphasizes that behaviors depends on the behavior rules of interacting individuals.
- Actions determine outcomes and thus survival and reproduction.

Concept 7: individual fitness

We write the individual fitness of an individual of type τ when interacting with an individual of type θ as

$$w(a^*(\tau,\theta), b^*(\theta,\tau))$$

This is the expected number of succesfull offpring produced by an individual (including the surviving self)[2]. It is two-tiered fitness:

- The first layer emphasizes the relation between fitness and action/behavior.
- The second layer emphasizes the relation between action/behavior and behavior rules/choice rules.

[2] Strictly speaking, this is a proxy of fitness that does not account for population regulation that we neglect here.

Darwinian fitness in a structured population

The Darwinian fitness (or lineage fitness) is here given by

$$W(\tau,\theta) = (1-r)\, w(a^*(\tau,\theta), b^*(\theta,\tau)) + r\, w(a^*(\tau,\tau), a^*(\tau,\tau))$$

This consists of two parts:

- With probability $1-r$ an individual interact with a non-lineage member and hence with an individual that has a different behavior rule.
- With probability r an individual interact with a lineage member and hence with an individual that has the same behavior rule as self.

Nested levels of interactions to go from actions to Darwinian fitness

$$\underbrace{B_\tau}_{\text{behavior rule}} \quad \to \quad \underbrace{a(\tau,\theta)}_{\text{behavior}} \quad \to \quad \underbrace{\pi(a(\tau,\theta), b(\theta,\tau))}_{\text{payoff}}$$

$$\to \quad \underbrace{w(a(\tau,\theta), b(\theta,\tau))}_{\text{individual fitness}} \quad \to \quad \underbrace{W(\tau,\theta)}_{\text{replicator fitness}}$$

Can we say something general about the type of behavior rules that is favored by evolution?

Concept 8: strategy evolution, i.e. the benchmark

The bulk of evolutionary biology consider strategy evolution. i.e. actions are genetically determined so that $\mathcal{A} = \Theta$. Then,

$$a^* \in \arg\max_{a \in \mathcal{A}} W(a, a^*)$$

where

$$W(a, b) = (1 - r)\, w(a, b) + r\, w(a, a)$$

If $\Theta = \mathbb{R}$ is real valued, the first order selective effects on Darwinian fitness yield Hamilton's marginal rule.

This is the modern interpretation Hamilton's (1964) result.

Concept 9: behavior rule evolution

Consider now behavior rule evolution so that

$$\Theta = \mathcal{B}$$

Under strategy evolution, evolution selects the action, while under behavior rule evolution, the individual selects the action, while evolution selects the behavior rule.

Concept 10: preference evolution

Let us now restrict behavior rule evolution to preference evolution (the case implicitly presumed by Hamilton) and set

$$\Theta = \mathcal{U} \subset \mathcal{B}$$

where $\mathcal{U}$ is the set of (rational) preferences, and which can be taken as the set of utility functions, with element u taking value $u(a)$ for action $a \in \mathcal{A}$.

- This does not mean that the organism will posses mental representation of utilities, but that over long times and given enough repetition behavior should become observationaly indistinguishable of that of a rational decision maker.
- Learning dynamics typically converge towards equilibria given by maximizing behavior.

Concept 10: preference evolution

- What should be the form of the uninvadable utility function?
- What individuals do, $a^*(\tau, \theta)$, and $b^*(\theta, \tau)$, depends on the utility function of their partner and hence everything becomes interdependent.
- Informational assumptions turn out to be fundamental.

We make the assumption of incomplete information. Individuals cannot recognise the genotypes of their interaction partner, they only have information about the density distribution of genotypes in the population.

Concept 11: incomplete information

We assume that individuals have incomplete information about their partner's type. The fitness of a mutant under incomple information is

$$W(\tau,\theta) = (1-r)\, w(a^*(\tau,\theta), b^*(\theta)) + r\, w(a^*(\tau,\tau), a^*(\tau,\tau))$$

where

$$\begin{aligned} b^* &\in \arg\max_{b\in\mathcal{A}}\; u_\theta\left(b, b^*\right) \\ a^* &\in \arg\max_{a\in\mathcal{A}} (1-r)u_\tau\left(a, b^*\right) + r\, u_\tau\left(a, a^*\right) \end{aligned}$$

which defines a Nash-Bayesian equilibrium.

Semi-Kantian preferences are uninvadable

The uninvadable preference is given by

$$u_{\tau^*}(a,b) = (1-r)\,w(a,b) + r\,w(a,a)$$

- The first is the individual's realized fitness, given the action used by the players.
- The second term is the fitness that the individual would realize if – hypothetically – the opponent used the same strategy, since the individual thereby evaluates what would happen if others were to follow the same course of action as itself.

"Act as if the maxims of your action were to become through your will a universal law of nature" (one of Kant's categorical imperative).

Semi-Kantian preferences are uninvadable

The uninvadable preference is given by

$$u_{\tau^*}(a,b) = (1-r)\,w(a,b) + r\,w(a,a)$$

The firts order condition yields Hamilton's marginal rule:

$$\underbrace{\left.\frac{\partial w(a,b)}{\partial a}\right|_{a=b}}_{-c} + r\,\underbrace{\left.\frac{\partial w(a,b)}{\partial b}\right|_{a=b}}_{b} = 0$$

If we define inclusive fitness as "that property of an individual organism which will appear to be maximized when what is really being maximized is gene survival", then we addressed Hamilton's claim in a standard model.

Concept 11: preferences at the payoff level

Assuming small effects of payoff on fitness, we can obtain a representation of the preference at the payoff level.

$$u_{\tau^*}(a,b) = (1-\lambda)\left[(1-r)\,\pi(a,b) + r\pi(a,a)\right] + \\ + \lambda\,(1-r)\left[\pi(a,b) - \pi(b,a)\right]$$

This encapsulates three proximal motives:

1. $\pi(a,b)$: Self-interest.
2. $\pi(a,a)$: Kantian interest.
3. $\pi(a,b) - \pi(b,a)$: Rivalry or spite tends to favor anti-sociality.

λ is the coefficient of fitness interdepenence and measures the extent to which an increase in own fitness decreases that of it group neighbour, measure the intensity of kin competition.

Behavior rule evolution: beyond incomplete information

More generally, we have

$$W(\tau, \theta) = (1 - r)\, w(a^*(\tau, \theta), b^*(\theta, \tau)) + r\, w(a^*(\tau, \tau), a^*(\tau, \tau))$$

Here, the action $b^*(\theta, \tau)$ of the interaction partner depends on τ.

This covers social evolution models for the evolution of

- Preferences under incomplete information.
- Cooperation.
- Individual and social learning.
- Norms.

So this form of invasion fitness is pretty generic.

Concept 12: Hamilton's rule and general social evolution

For quantitative trait evolution $\Theta = \mathbb{R}$, we have that Hamilton's marginal rule holds at the trait level

$$\begin{aligned} S(\theta) &= \left.\frac{\partial W(\tau,\theta)}{\partial \tau}\right|_{\tau=\theta} \\ &= \left.\frac{\partial w(a^*(\tau,\theta), b^*(\theta,\tau))}{\partial \tau}\right|_{\tau=\theta} + r \left.\frac{\partial w(a^*(\tau,\theta), b^*(\theta,\tau))}{\partial \theta}\right|_{\tau=\theta} \end{aligned}$$

- This holds regardless of trait embeding and provides a general perspective on the selection pressure on any trait.
- Most generous interpretation of Hamilton's (1964) result.

All models of quantitative trait evolution with differential fitness can be put under this form (and this can be generalized to an arbitrary number of traits thus implementing any recurrent neural network).

Behavior rule evolution: beyond incomplete information

- Break down of Hamilton's rule at the action level.
- Close connection between preference evolution models, learning rules evolution models, and model for the evolution of cooperation under repeated interactions.
- Polymorphism in trait evolution is very likely.
- Interaction between relatedness and other mechanisms.
- Many interesting things to explore.

Thank you for your attention!

- Alger, Weibull and Lehmann 2020 “Evolution of preferences in structured populations: genes, guns, and culture” (Journal of Economic Theory).

- Alger and Lehmann 2023 “Evolution of semi-Kantian preferences in two-player assortative interactions with complete and incomplete information and plasticity” (Dynamic Games and Applications).