Julien Panis-Lie
Self-sacrifice as social signal

supervisor: Jean-Louis Dessalles
Département Informatique et Réseaux
Télécom Paris

Defense target date: June
Language: English

Reviewer suggestion (non CP): Jean-Baptiste André
Reviewer suggestion (CP): Nicolas Baumard or Coralie Chevallier

Self-sacrifice as social signal

Modeling the motivations underlying an extreme example of prosocial behavior

1. Introduction

Throughout history, humans have been willing to lay down their lives for the sake of their groups (Whitehouse, 2018). Whether admired as heroes or reviled as terrorists, individuals who engage in such extreme pro-social behavior tend to be relatively well-off, educated and to display no appreciable psychopathology (Atran, 2003) – underscoring that self-sacrifice cannot simply be viewed as the result of unadaptive miscalculation or proximal causes (e. g. the contents of a particular ideology).

Could self-sacrifice therefore have a biological function? This project purposes to investigate the potential biological motivations that may underlie this behavior. Even though collective benefit is the displayed (moral) motivation for self-sacrifice, biological motivation will be understood in terms of benefits to the individual (Williams, 1996).

To do this, we will attempt to model these biological motivations, using a *social signaling framework* (Dessalles, 2014). Social signals are a specific case of costly signals (Zahavi, 1975; Grafen, 1990) whose purpose are to attract friends. In contexts where the signaled quality correlates with the fitness of friends, such a quality may be in social demand; when, in addition, the potential benefits in terms of increased social status upset the costs, social signaling can be expected to emerge.

This explanation cannot hold for self-sacrifice, however, since signalers would not survive to enjoy the advantages of their new status. We propose to add an additional hypothesis to our theoretical framework: *that social status be in part heritable for our species*, as the high status of one individual can raise that of every member of his or her family (Service, 1971). In a context where a martyr's family members are in high social demand, this could theoretically suffice to make self-sacrifice evolutionary stably – as long as it remains a low-frequency behavior (the fewer the heroes, the higher their status).

We are thus left with the need to find a plausible context where a martyr's family may be in high social demand. This is not immediately obvious, as the quality signaled by individuals who engage in self-sacrificial behavior need not be heritable. We propose to consider a context akin to *inter-group conflict*, where it may be crucial to ensure that one's friends carry no sympathy for the enemy (e. g. to avoid betrayal). In such a situation, *commitment to the group* is a highly desirable quality to have in friends. One way for (alive) individuals to signal this quality may be to honor the fallen martyrs, for instance by engaging in conspicuous ceremonies in their name. If and when such costly second-order signaling is evolutionary stable, honoring could entail indirect benefits for the family of martyrs, meaning that self-sacrifice may itself emerge – the two signals being expected to be mutually reinforcing, as would-be-patriots need martyrs to signal their commitment to the group.

Our main hypothesis, which we venture to investigate, is therefore the following: in a biological population where social status is in part heritable, and which is engaged in a context akin to inter-group conflict, self-sacrifice and honoring may emerge as first-order and second-order signals of individuals' commitment to the group, or patriotism. Using computer simulations, our project will study the conditions (if any) where these two signals may emerge (e.g. cost and probability of betrayal).

2. Methods

2. 1. Implementation of the model

The script is written in Python, and set in the Evolife framework, developed by Jean-Louis Dessalles to study various evolutionary phenomena (https://evolife.telecom-paristech.fr/). Evolife has been used to study social signals in general (Dessalles, 2014) and to model language as a particular social signal (Dessalles, 2017).

Evolife is based on a genetic algorithm. An individual's behavior is controlled by a binary vector (genome). Individuals live, reproduce sexually, and gain points in a (yearly) life game. Two modes of selection can be implemented:
- *ranking*: individuals are ranked following the points they have obtained, and are granted a number of potential children that is an increasing (non-linear) function of that rank ;
- *differential death*: individuals are granted life points (related to their yearly points) which protect them from life hazards, thus increasing their life expectancy (and reproductive opportunities).

As detailed below, the proposed script exploits this dichotomy in its treatment of the envisioned first-order and second-order signals. Evolife is modular in structure; the script would constitute a scenario implemented inside of Evolife. Source files (https://github.com/jlie10/SelfSacrifice) can be found on GitHub – on the "Internship" branch ("Master" being kept for a pending class validation).

2. 2. "Exogenous" base script

Before implementing the actual script, a simplified "exogenous" version was implemented (Exogenous.py). In this version, the social value of self-sacrifice is fixed exogenously, and depends on an 'Admiration' parameter. In order to make the link to the second script more obvious, we supposed that total admiration for heroes was equal to 'Admiration' times the number of individuals that don't sacrifice – seeing the low final yearly frequency of self-sacrificial behavior (see under), a truly fixed version (where total admiration does not depend on the number of heroes) would yield similar results.

This script serves as a basis for the full script. Self-sacrifice is treated as a genetically controlled trait: individuals in the scenario are endowed with a 'SelfSacrifice' gene (8 bits) whose relative value corresponds to an individual's probability of engaging in self-sacrificial behavior in a given year.

A heroes' sacrifice benefits his/her descendants. Each year, admiration for heroes "spills over" to the rest of the population. An individual's 'share' in this allocation depends on their ascendant's heroism: each time a parent engages in self-sacrificial behavior, his/her children gain 1 share point. Shares are inherited, as individuals are born with the average of their parent's share times a 'SacrificeHeredity' parameter, situated between 0 and 1. This parameter is generally kept under 1 (typically 0.5), to avoid a situation where individuals end up strongly related (as a given heroes' family invades the population).

The emergence and stabilization of self-sacrifice is studied following a ranking mode of selection. Admiration points indirectly gained by individuals are converted into reproductive points, by taking the integer part of these points divided by a 'ReproGainsThreshold' parameter (typically fixed at 10).

Keeping other parameters at their typical values (see after), self-sacrifice is, unsurprisingly, shown to

depend on ‘Admiration’ (relative to ‘ReproGainsThreshold’): when this parameter is too low (e.g. under 5 – tentatively), this gene’s value remains at 0 across the population; when the parameter is augmented, the equilibrium value of the gene grows with it – although it is quickly capped at 4-5 % (which, for a small population of 200, entails an average of 10 sacrifices per ‘year’ - which is far from negligible).

Full statistical analysis of this behavior has yet to be performed. The idea, in the following scenario, is to see if, in conditions where we would expect (alive) individuals to pay tribute to heroes on a scale comparable to the previous values of ‘Admiration’, self-sacrifice emerges as a stable strategy (as measured by gene value across the population, and as compared to results obtained in similar conditions for this base scenario).

2. 2. Actual script structure

This script (Sacrifice.py) builds on the previous one. The idea is to see if admiration for heroes, and therefore self-sacrificial behavior, can emerge “endogenously”, in a context akin to inter-group conflict, as described previously. The previous dynamic and ranking mode of selection is conserved for self-sacrifice, while other genes (see under) follow a differential death mode of selection (as points gained or lost in the social interactions detailed under are later translated into life points).

In addition to the ‘SelfSacrifice’ gene, individuals are born with a ‘Patriotism’ phenotype – which, for the sake of simplicity is equal to 1 or 0. Individuals are also endowed with a ‘Demand’ gene, as well as a ‘Patriot’ and a ‘NonPatriot’ gene, which respectively control their yearly investment in honoring heroes when they are patriots (1) or not (0).

When there are heroes to honor, this (costly) investment becomes a visible signal of their commitment to the group (patriotism). Individuals select their friends according to the potential signalers they encounter (if there are no heroes, every individuals signals at 0) and their ‘Demand’: any individual who signals above this value may be accepted as a friend (one interaction yields one new friend at the maximum).

Friendship is assumed to be mutually beneficial: an individual with a low ‘Demand’ will thus increase his/her chances of gaining from friendship. However, some ‘NonPatriot’ individuals may be traitors: befriending them ends up being extremely costly, as they betray their friends in the final stage of the year, to their own gain. Thus, in uncertain conditions where being betrayed is a probable and costly outcome, individuals with high ‘Demand’ may fare better.

In conditions where honoring is an honest signal of one’s future (absence of) betrayal, ‘Demand’ and ‘Patriot’ may thus potentially co-emerge, which may then (if ‘Patriot’ is high enough) allow ‘Self-sacrifice’ to emerge. This could be the case if non-patriots honoring potential is capped: because non-patriots are already committed to another group, or themselves, they may not have the time or resources to invest as much as patriots are able to.

2. 3. Important parameters

2. 3. 1) Main variable parameters

- ‘NbTraitors’: probability that a non-patriot actually ends up being a traitor. This is the main

parameter, which will vary between 0 and 100 ;
- 'MaxOffer': maximum investment that non-patriots can afford in honoring (out of 100). Fixed at 50 in an initial stage, although exploring other values – including 100 – will be useful (see after) ;
- 'FriendshipValue': value gained from being friends with a non-traitor – fixed at 10 in an early stage ;
- 'JoiningBonus': value gained from having been accepted as friend by someone else – fixed at 10 in an early stage ;
- 'Judas': what a traitor gains from betraying you – fixed at 20 in an early stage ;
- 'DenunciationCost': cost of being betrayed – fixed at 100 in an early stage.

Individuals start off the year with 100 points. The previous values for social points (last four parameters) were obtained via reasonable estimates from this scale, and after numeric simulations of honoring in isolation (see Honor.py). Given all these parameters, 'NbTraitors' controls the probability of betrayal. In order to differentiate with the cost of betrayal, the latter two parameters will also be made to vary around the previously specified values.
In this simplified version with only two levels of patriotism, 'MaxOffer' creates an arbitrary cutoff, above which a signal can be considered as honest. For this reason, 'MaxOffer' will be largely kept constant at the arbitrary 50, although looking at 100 could be useful to understand what happens when honoring is not honest.

2. 3. 2) Other fixed parameters (at least in a first stage)

With respect to the (expected) first-order signal:
- 'ReproGainsThreshold': corresponds to the value in terms of points acquired through honoring of one's descendants that yields an additional reproductive unit (for ranking) – kept at 10 ;
- 'Selectivity': controls how the degree of selection by ranking (how much expected number of offspring increases with rank) ;
- 'SacrificeHeredity': controls how shares are inherited from parents (see 2.1) – fixed at 50 in a first stage ;
- 'ReproductionRate': expected rate of children left each year. Fixed at 30%, a relatively high value, as (potential) equilibrium value for probability of self-sacrifice can be shown to be inferior to reproduction rate.

With respect to the (expected) second-order signal:
- 'HonoringCost': percent of investment in honoring that translates into cost – fixed at 100 ;
- 'SelectionPressure': maximum number of life-points that can be earned in a round – fixed at 6 (meaning that individuals with the highest scores have to be randomly selected 7 times before enough other individuals are selected between 1 and 6 times, in order to die that turn) ;
- 'EraseNetwork': indicates whether individuals' friendship networks are reinitialized each year. Fixed at 0 (False), mainly in order to avoid having to largely increase 'Rounds' and 'NbInteractions' (and hence computation time).
- 'Rounds': number of times friendship-forming interactions are launched in the population – fixed at 10 ;
- 'NbInteractions': number of such interactions per round, where one randomly chosen individual may accept a maximum of one friend – fixed at 100 ;
- 'SampleSize': size of the random sample of (live) people an individual thus interacts with – fixed at 5 ;
- 'MaxFriends': maximum number of friends one can hold – fixed at 10. Friendship bonds are assumed

to be symmetrical here ;

Genetic and general population parameters:
-'AgeMax': age after which individuals automatically die – fixed at 40 ;
- 'PopulationSize': fixed at 1000 ;
- 'MutationRate': fixed at 5 per 1000 ;
- 'NbCrossover': number of crossovers occurring during sexual reproduction – fixed at 1 ;
- 'GeneLength': fixed at 8, weighted bits (meaning that each gene value corresponds to a value between 0 and 2 to the power of 8 minus 1)
- 'AgeAdult': age before which reproduction is impossible – fixed at 0

It is anticipated that keeping 'EraseNetwork' at 0 and 'Rounds' and 'NbInteractions' relatively low is equivalent to re-initializing networks with large number of interactions – although this should be checked.
In addition, a 'DiffentialCosts' mode for honoring could be implemented, where dishonest signalers (non-patriots) would face a premium, rather than a capped offer.

2. 4. Predictions

Our main prediction is that, under plausible conditions corresponding to the default parameter values detailed above, honoring and self-sacrifice will co-emerge as second and first-order signals of patriotism (inability to betray). We expect this to be largely controlled by 'NbTraitors' (probability of betrayal).

When and if self-sacrifice emerges, we do not expect its probability to be high, for logical reasons (as gains in status decrease when there are too many perpetrators) as well as reasons relating to previous simulations (we do not expect equilibrium values to exceed that of the corresponding "exogenous" situations). In addition, we do not expect honoring to be stable in the long-term in the absence of heroes (as we envisioned it as a second-order signal based on the first-order signal).

More specifically, keeping all other parameter values constant, we expect:
- that for 'NbTraitors' equal to 0 or sufficiently small, 'Demand' should stay at 0 or non-significantly higher (as friendship is risk-free), thus inducing 'Patriot' and 'NonPatriot' to stay low, preventing the emergence of self-sacrificial behavior ;
- that when 'NbTraitors' is high enough for there to be a selective pressure for having 'Demand' above 'MaxOffer', honoring is a potentially honest signal. If, in addition, the benefits to patriots outweigh the costs, then honoring may emerge at such a level. Seeing as we chose 'MaxOffer' at a level which, in the "exogenous" script (for 'Admiration') was sufficient for self-sacrifice to emerge and stabilize, we that, provided that honoring can emerge, that self-sacrifice emerges as well ;
- if we vary 'DenunciationCost' instead of 'NbTraitors', we expect similar dynamics ;
- if we decrease 'JoiningBonus' under a certain threshold, attracting friends through signaling patriotism should become overly costly (risky), preventing any non-null behavior at equilibrium.

We expect 'MaxOffer' to play a decisive role, as it should correspond to the potential equlibrium state – when it is attainable. Testing 'MaxOffer' at 100 would be a way of seeing the consequences of dishonest signaling: under normal conditions we would expect that 'Demand' then rise to the maximum (as it is impossible to select patriot friends), precluding any signaling from emerging.

However, if ‘Judas’ is high enough, signaling may paradoxically remain interesting for non-patriots and we would expect a dishonest signaling equilibrium to emerge.

2. 5. Limits and possible refinements of the model

‘MaxOffer’ thus imposes a paradoxical arbitrary limit: if it is too low, we would not expect self-sacrifice to emerge, as honoring of heroes by patriots should not yield enough potential benefits for heroes’ children (see 2.1).

This is further motivation to explore a “differential costs” option, whereby dishonest signaling is assumed to be costlier than honest signaling (e.g. for the same type of reason, as non-patriots opportunity costs are higher, as they are engaged elsewhere / have other opportunities to invest in themselves…). This could allow for less straightforward results, as we would expect both signals to emerge above a certain threshold (for the ‘DishonestPremium’) and not under, without this threshold being immediately obvious.

However, we expect both treatments to be mathematically equivalent, as the poverty of our (expected) ability to modify final outcomes by varying parameter values derives from the poverty of the model itself. In a second-stage, the model could be refined:
- by allowing for more categories of patriotism, and the corresponding genes ;
- or, better yet, by allowing for continuous (or a large number of discrete values for) patriotism, and replacing genes by social learning of investment in honoring.

2. 6. Interpretation

The immediate outputs of the simulation are average (and individual) levels of gene values over time. If, as predicted, average value for the gene controlling self-sacrificial behavior stabilizes at a non-null value, and if the various parameters (and 'Honoring' genes) influence this according to our predictions, then we will have shown that biological motivations underlying self-sacrifice are a theoretical possibility. We would then study stability of the model itself over a certain range of parameters.

As detailed above, interpretation is however compounded by the current simplistic form of the model.

References

Atran, S. (2003). Genesis of Suicide Terrorism. *Science*, *299*(5612), 1534-1539. https://doi.org/10.1126/science.1078854

Dessalles, J.-L. (2014). OPTIMAL INVESTMENT IN SOCIAL SIGNALS: OPTIMAL INVESTMENT IN SOCIAL SIGNALS. *Evolution*, *68*(6), 1640-1650. https://doi.org/10.1111/evo.12378

Dessalles, J.-L. (2017). Language: The missing selection pressure. *arXiv:1712.05005 [physics, q-bio]*. Consulté à l'adresse http://arxiv.org/abs/1712.05005

Grafen, A. (1990). Biological signals as handicaps. *Journal of Theoretical Biology*, *144*(4), 517-546. https://doi.org/10.1016/S0022-5193(05)80088-8

Service, E. R. (1971). *Primitive social organization: An evolutionary perspective*. Random House, Toronto, CAN.

Whitehouse, H. (2018). Dying for the group: Towards a general theory of extreme self-sacrifice. *Behavioral and Brain Sciences*, 1-64. https://doi.org/10.1017/S0140525X18000249

Williams, G. C. (1996). *Adaptation and natural selection: a critique of some current evolutionary thought*. Princeton, NJ: Princeton Univ. Press.

Zahavi, A. (1975). Mate selection—A selection for a handicap. *Journal of Theoretical Biology*, *53*(1), 205-214. https://doi.org/10.1016/0022-5193(75)90111-3