2. Internship objective and methods //+ predic

2.1. Objective and methods

2.1.1) Objective

The general objective of the internship is thus to study an alternative explanation for self-sacrifice, using a social signaling framework – as introduced above. Evolutionary signaling theory offers a path for avoiding the potential pitfalls (see 1.2.) of maladaptation and group benefit, as the previous hypothesis was formulated purely in terms of individual (inclusive) fitness.

Enlever “pitfall” ? / be more modest /// explain: more direct than… no ref to collective benefit...

The underlying objective of such an evolutionary approach to self-sacrifice is to go beyond its various manifestations and/or proximate causes, in order to provide the bases for developing a more integrated understanding of the phenomenon. By investigating these characteristics in relation to a potential biological function of self-sacrifice, one hopes to eventually be able to paint a picture of the causal relations between its various features, and tackle the question of its history (Tinbergen, 1963).

Of course, the actual objectives of the internship are not that lofty. The point is to attempt to provide a logical and reasonably robust argument (see next section), which may allow for limited reinterpretation of self-sacrifice. To begin, an evolutionary account in a general context (intergroup conflict) constitute an argument against the explanatory power of specific beliefs or ideology. Further, if self-sacrifice may be interpreted as a social signal of one’s commitment to the group, then group benefit should be understood as part of the signal, not its explanation. In such an interpretation, the fact that a group may benefit from an individual’s self-sacrifice is merely a consequence of that individual engaging in a behavior which has tended to augment his/her inclusive fitness in our evolutionary past.
~~[viré : donc biological motivations = … selfish genes] / qui pourrait etre lie a rituels aussi ci-dessous...~~

In addition, the paper aims to provide a stand-alone explanation. As with Blackwell’s (2008) model, a fundamental idea is that self-sacrifice may be beneficial to inclusive fitness, a martyr’s relatives may benefit from his/her celebrity. In contrast however, the hypothesis above aims to provide a mechanism by which may arise such a situation where martyr’s gain posthumous celebrity and this entails social (rather than material) benefits to their kin.

This can allow enable a re-framing of the role played by painful rituals in relation to identity fusion and self-sacrifice. As argued in the previous section, competition ensures the honesty of signals of one’s commitment to the group, whose cost grows with risk to the group (with demand). One is tempted to interpret painful rituals in this light, which allows to account for the correlation between their intensity and that of fusion, as felt by the group’s members (Whitehouse, 2019). Thus, “rites of terror” (Whitehouse, 1996) can be interpreted as an extreme example of a social signal, which may emerge

when groups face extraordinary threat. In addition, the proposed two-tier model offers a way to connect fusion and such rites with self-sacrifice, tentatively[10] suggesting that in such conditions, demand may allow extremely costly rights to emerge, which may warrant self-sacrifice – as mediated by the proximate feeling of "extreme" identity fusion.

2.1.2) Methods

Work during the internship relied first on *computer simulations*, using evolutionary paradigms. The basic idea behind these simulations is to achieve a *proof of concept*, meaning the simplest possible simulation that allows for the phenomenon under study to emerge in a robust manner (i.e. to be stable for a large array of parameter values). A credible proof of concept is an argument in favor of an explanation's logical and evolutionary plausibility, as pertaining to the objective outlined above – while failing to do so is indicative of important overlooked flaws.

The projected model was implemented into two Python scripts, and, in order to further assess the logic of the model and the validity of its implementation, predictions made on the basis of the above hypothesis were tested (see below).

Internship work was also grounded in analysis in terms of game theory. Using simplifying hypotheses, the objective was to obtain the simplest possible mathematical characterization of the phenomena under study, as well as the conditions under which they are stable.

2.2. Computer simulations

2.2.1) Evolife

Two (actually three) scripts were 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). Note that an individual's "genes" are understood here in a wide, evolutionary (or informational) sense[11]: they are *schemata*, portions of an individual's genome small enough to survive through evolutionary time and thus be considered as units of selection (Holland, 1975; Dawkins, 1976).

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 ;

10 This is merely a suggestion, one could probably play this argument the other way around.

11 This is different from the definition a molecular biologists would use, as it makes no reference to coding for proteines.

- *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 scripts exploit this dichotomy in its treatment of the envisioned first-order and second-order signals. Evolife is modular in structure; the script constituted a scenario implemented inside of Evolife. Both scripts can be found in this document's annex.

[GITHUB?]

2.2.2) "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. As such, this model can be seen as a social alternative to the "economic" model developed by Blackwell (2008). In order to make the link to the second script more obvious, total admiration for heroes was made 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 5).

2.2.3) 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’ should 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. As argued above (see 1.3.6)), this could be the case if non-patriots honoring potential is capped (‘MaxOffer’ parameter) or if they face a signaling premium (‘DishonestPremium’).

2.2.4) Main variable parameters

With respect to signal honesty:

- ‘DifferentialCosts’: specifies the “mode” – equal to 1 when non-patriots face a premium, and 0 when they face a cap (0 by default);

- ‘DishonestPremium’: premium (percentage) faced by dishonest signalers in differential costs mode;

- ‘MaxOffer’: maximum investment that non-patriots can afford in honoring (out of 100). Fixed at 50 in an initial stage, although exploring ther values – including 100 – will be useful (see after).

With respect to the social context, including the risk due to intergroup conflict (betrayal):

- ‘NbTraitors’: probability that a non-patriot actually ends up being a traitor. This is the main parameter, which will vary between 0 and 100;

- ‘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, simulations explored different values for the latter two parameters.

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.

[ZAPPER TOUTE LA PARTEI 232 ??]]

2.2.5) 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 2.2.2)) – fixed at 50 after an early analysis of results in the “exogenous” case;

- ‘ReproductionRate’: expected rate of children left each year. Fixed at 15%[12].

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;

12 30% in the preregistration document: changing the way results were visualized allowed to take a smaller, arguably more realistic, figure (see 3.1)

- ‘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 200[13];

- ‘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.

[Pareil on peut tej cette partie, vu qu’on va y repondre dans la partie d’apres….] // + attention aux “we”

2.3. Predictions

Specific predictions for these simulations are derived from those made in section 1.3.5). The hypothesis under study translates into: under plausible conditions corresponding to the default parameter values detailed above (representing a context of intergroup conflict), honoring and self-sacrifice should co-emerge as second and first-order signals of patriotism (inability to betray).

Both of these signals, as captured by individuals’ equilibrium gene values, should be *mutually reinforcing*. At equilibrium, neither should remain stable in the absence of the other, and large levels of one should correspond to large levels of the other.

When and if self-sacrifice emerges, however, it should remain a low-frequency behavior, captured here by *low probability of self-sacrifice* for all individuals at equilibrium, as controlled by relative gene value.

[=> limitation on honoring …. depend des RESULTS + interp]

More specifically, keeping all other parameter values constant, expectations are:

- 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’

13 1000 in the pre-registration document: a large population size had been anticipated to allow for easier visualization of results, but another more computationaly efficient solution was proposed (see 3.1)

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.