Surrogate goals under uncertainty

Tobias Baumann

December 3, 2019

1 Introduction

One common challenge to surrogate goals is that in real-world settings, both the threatener and the threatenee face considerable uncertainty about the other agent, which is why it can be hard to fulfil the desideratum of threatener-neutrality while also ensuring that threats target the surrogate goal rather than the initial goal.

In this document, I will develop a quantitative model of surrogate goals under uncertainty. Note that this is very different from the problem of private information, which is another challenge, and also from the credibility problem. For purposes of this document, we assume that the degree of uncertainty in the agents' estimates is common knowledge.

2 Mathematical framework

In the following, I will assume that there is some quantitative measure of *vulnerability to threats*. This could simply refer the extent to which the target cares about a certain goal, but it could also capture how costly it is to go through with a threat, the extent to which third parties would enact punishments, and other factors.

Given this, let X_i and X_s be random variables that represent the vulnerability of the initial goal and surrogate goal, respectively, as estimated by the threatener.

In addition, let $U(x)$ be a continuous, monotonically increasing function that denotes the threatener's utility if she threatens someone of blackmailability level x. Accordingly, the expected utility of threatening the initial goal is $\mathbb{E}[U(X_i)]$ and the expected utility of targeting the surrogate goal is $\mathbb{E}[U(X_s)]$. The threatener will only threaten the surrogate goal if

$$\mathbb{E}[U(X_s)] > \mathbb{E}[U(X_i)]. \tag{1}$$

This can be simplified in certain special cases:

- If X_i, X_s are constants, then (1) reduces to $X_s > X_i$ as U is monotonically increasing.
- If U is linear, then $\mathbb{E}[U(X_i)] = U(\mathbb{E}[X_i])$ and $\mathbb{E}[U(X_s)] = U(\mathbb{E}[X_s])$. Hence

  $$\mathbb{E}[U(X_s)] > \mathbb{E}[U(X_i)] \iff U(\mathbb{E}[X_s]) > U(\mathbb{E}[X_i]) \iff \mathbb{E}[X_s] > \mathbb{E}[X_i].$$

  The last step holds because U is monotonically increasing.
- Suppose X_i is a constant. (This is a relevant special case because it is plausible that there is much less uncertainty about the initial goal than about the surrogate goal.) If, in addition, U is convex, then $\mathbb{E}[X_s] > \mathbb{E}[X_i]$ is sufficient, but not necessary, to fulfil (1). This is because $\mathbb{E}[f(X)] \geq f(\mathbb{E}[X])$ for convex f, so

  $$\mathbb{E}[U(X_s)] \geq U(\mathbb{E}[X_s]) > U(\mathbb{E}[X_i]) = \mathbb{E}[U(X_i)].$$

  However, in general, $\mathbb{E}[X_s] > \mathbb{E}[X_i]$ is not sufficient. As long as $P(X_s < X_i) > 0$, there exists U such that (1) is violated.

3 Examples

Suppose that (the threatener's estimate of) vulnerability to threats is log-normally distributed, which is a plausible model if it is a product of many independent factors. That is,

$$X_i \sim \text{Lognormal}(\mu_i, \sigma_i^2)$$
$$X_s \sim \text{Lognormal}(\mu_s, \sigma_s^2)$$

Suppose, further, that $U(x) = x^a$ for $a > 0$. If $X \sim \text{Lognormal}(\mu, \sigma^2)$, then $X^a \sim \text{Lognormal}(a\mu, a^2\sigma^2)$. So

$$U(X_i) \sim \text{Lognormal}(a\mu_i, a^2\sigma_i^2)$$
$$U(X_s) \sim \text{Lognormal}(a\mu_s, a^2\sigma_s^2)$$

Therefore

$$\mathbb{E}[U(X_i)] = \exp\left(a\mu_i + \frac{a^2\sigma_i^2}{2}\right)$$
$$\mathbb{E}[U(X_s)] = \exp\left(a\mu_s + \frac{a^2\sigma_s^2}{2}\right)$$

(1) is therefore equivalent to

$$\begin{aligned} &\mathbb{E}[U(X_s)] > \mathbb{E}[U(X_i)] \\ \iff &\exp\left(a\mu_s + \frac{a^2\sigma_s^2}{2}\right) > \exp\left(a\mu_i + \frac{a^2\sigma_i^2}{2}\right) \\ \iff &\mu_s + \frac{a\sigma_s^2}{2} > \mu_i + \frac{a\sigma_i^2}{2} \end{aligned}$$

Depending on the value of a and σ_i, this can mean that the parameters μ_s and σ_s^2 must be chosen so that X_s is significantly higher (in expectation) than X_i, rather than just marginally higher, resulting in tradeoffs between the benefits of a surrogate goal and a loss in bargaining power. In particular, this is the case if a is close to zero and σ_i is small.

Note that I am bracketing the question of whether it is even possible to freely choose the parameters μ_s, σ_s^2 associated with (the threatener beliefs regarding) the vulnerability of the surrogate goal. This ties in to credibility issues and is likely only possible to a limited extent; in particular, it is plausible that one can mostly influence μ_s rather than σ_s^2, as the latter represents the variance in the threatener's estimate. This potentially exacerbates the loss in bargaining power that's necessary to make the surrogate goal work.

Another interesting example, which highlights the same issue, is the threatener utility function $U(x) = 1 - \frac{1}{x}$ with the same lognormal distributions of X_i, X_s. This represents a situation where it is very bad for the threatener if the victim's vulnerability to threats is near zero, but there are strongly diminishing returns to high levels of vulnerability.

A quick calculation (analogous to the above) reveals that, in this case,

$$\begin{aligned} &\mathbb{E}[U(X_s)] > \mathbb{E}[U(X_i)] \\ \iff &\mu_s - \frac{\sigma_s^2}{2} > \mu_i - \frac{\sigma_i^2}{2}. \end{aligned}$$

Now, suppose σ_i^2 is small, corresponding to a high degree of information about the vulnerability of the initial goal. If the variance σ_s^2 of the surrogate goal is large and we can only control μ_s, then μ_s must be very large, which represents a severe loss of bargaining power (if this is possible at all).

That said, the severity of this issue depends on the threatenee's utility as a function of vulnerability to threats, which is another variable that the above framework neglects for reasons of simplicity. Also, it is not clear whether real-world situations would involve U-functions (and X_i, X_s distribution) that result in strong tradeoffs.