Surrogate goals and private information

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 how surrogate goals interact with private information. I will focus on the threatener's private information about how costly it would be, e.g. in terms of reputation effects or reactions of third parties, to go through with threats against either the original goal or the surrogate goal. [1]

2 Mathematical framework

2.1 A simple threat game with private information

Consider a simple two-stage threat game: In the first state, the threatenee (player 1) decides whether or not to give in. If so, the game ends with payoffs $(-R, R)$. If not, the threatener decides whether to carry out the threat. If so, the resulting payoff is $(-T, -C)$, if not, the payoff is $(0, 0)$. Assume that $T > R > 0$ and $T > C$.

As long as $C > 0$, i.e. it is costly to go through with a threat, the only subgame-perfect Nash equilibrium is to not give in and not go through with the threat. In other words, the threat is not credible.

However, this changes if we instead consider a Bayesian game, where the payoff for the threatener when carrying out the threat is not constant, but described by a type parameter θ, representing private information.[2] Now, the threatener would carry out the threat if (and only if) $\theta \geq 0$, which solves the credibility problem if there is a non-negligible chance that $\theta \geq 0$ - that is, it is actually in the interest of the threatener to go through, e.g. because of reputation effects. In this framework, the threatenee would give in if $R < P(\theta \geq 0) \cdot T$, and not give in otherwise. This means that (if $R > P(\theta \geq 0) \cdot T$) there are equilibria where threats are carried out with non-zero probability.

2.2 Introducing a surrogate goal

Next, we will consider a case where there is both an original goal and the surrogate goal. The payoffs only differ in the branch where threats are carried out. For a threat against the original goal, the payoff is $(-T, \theta_O)$, for a threat against the surrogate goal, it is $(0, \theta_S)$, where θ_O, θ_S are type parameters that represent the threatener's private information about the cost of going through.

Let us further assume that the surrogate goal is perfectly credible. The threatener's strategy, then, is simply to threaten the surrogate goal if $\theta_S > \theta_O$, i.e. if harming the surrogate goal is less costly, and threaten the initial goal otherwise.

If p is the probability of the threat being carried out, then the adoption of the surrogate goal improve the threatenee's expected payoff by

$$p \cdot T \cdot P(\theta_S > \theta_O)$$

.

[1] Other forms of private information are also relevant. For example, the threatenee has private information about whether its surrogate is genuine, which gives rise to the credibility problem.

[2] The payoffs if the threat is carried out are $(-T, \theta)$.

2.3 Quantifying expected bargaining loss

The flip side of the surrogate goal is that it worsens the threatenee's bargaining position if $\theta_S > \theta_O$. To quantify this effect, we will consider the game as a bargaining problem. If going through with a threat results in payoff $-C$ for the threatener, then the disagreement point of the bargaining problem is $\alpha \cdot (-T, -C)$, where $0 \leq \alpha \leq 1$ represents how credible the threat is. (It is debatable what the right disagreement point is, but we are mostly interested in the qualitative behaviour, and the constant does not matter that much.) Normalising the disagreement point to $(0, 0)$, the Pareto frontier of the bargaining problem is $(\alpha T - R, \alpha C + R)$ for R between 0 and αT.

The Nash bargaining solution maximises the product $(\alpha T - R) \cdot (\alpha C + R)$. This yields

$$\frac{d}{dR}(\alpha T - R) \cdot (\alpha C + R) = -(\alpha C + R) + \alpha T - R \stackrel{!}{=} 0 \tag{1}$$

$$\Leftrightarrow R = \alpha \frac{T - C}{2} \tag{2}$$

Given this, the bargaining loss from adopting the surrogate goal is $\alpha \frac{\theta_S - \theta_O}{2}$ if $\theta_s > \theta_O$, and 0 otherwise. The expected bargaining loss is

$$\alpha \frac{\mathbb{E}[(\theta_S - \theta_O)^+]}{2},$$

using the notation $X^+ = max\{X, 0\}$.

3 Example

Let $\theta_O \sim N(0, 1)$ and $\theta_S \sim N(\mu, 1)$.[3] Then $\theta_S - \theta_O \sim N(\mu, 2)$.

The gain from using a surrogate goal with parameter μ is, as seen above, proportional to

$$P(\theta_S > \theta_O) = 1 - \Phi(-\frac{\mu}{\sqrt{2}}) = \Phi(\frac{\mu}{\sqrt{2}}),$$

where Φ denotes the cumulative distribution function of the standard normal distribution.

The loss of bargaining power is proportional to

$$\mathbb{E}[(\theta_S - \theta_O)^+] = \int_0^\infty x \frac{1}{2\sqrt{\pi}} e^{-\frac{(x-\mu)^2}{4}} dx \tag{3}$$

$$= \mu \frac{\text{erf}(\frac{\mu}{2}) + 1}{2} + \frac{e^{-\frac{\mu^2}{4}}}{\sqrt{\pi}}, \tag{4}$$

where erf is the error function.

Figures 1 and 2 show the qualitative behaviour of the resulting gain and loss as a function of how treatener-friendly the surrogate goal is (as expressed in the parameter μ).

4 Discussion

This analysis shows that there may be a genuine tradeoff between the benefit of a surrogate goal in terms of preventing disvalue from threats being carried out, and the loss of bargaining power. It is not possible, when taking private information into account, to make the bargaining loss arbitrarily small while also maintaining a high probability that the surrogate goal will be targeted.

Depending on the values of T, p, α, this may mean that adopting a surrogate goal is not considered worthwhile for any possible degree of threatener-friendliness. On the other hand, if T is large, it is still plausible that the gain from avoid very bad outcomes outweighs the cost.

[3] It is more realistic that θ_O is negative (with high probability), but we can normalise it to mean 0 without loss of generality.

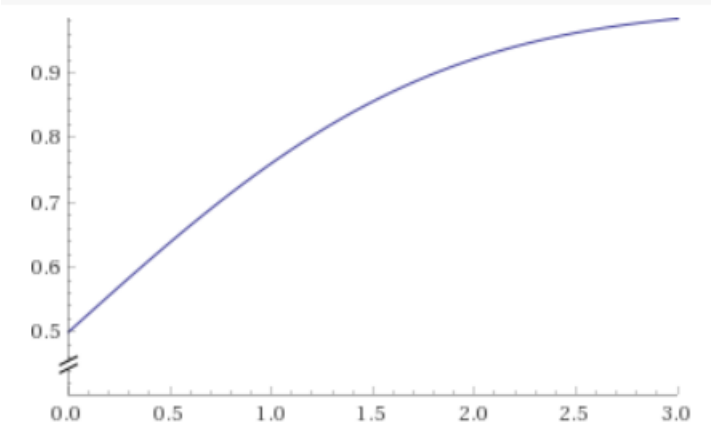

Figure 1: Expected improvement of payoff from using a surrogate goal (arbitrary units) as a function of mu

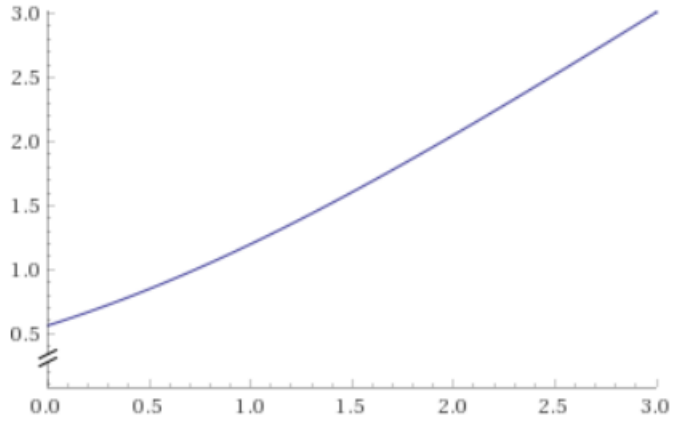

Figure 2: Bargaining loss from using a surrogate goal (arbitrary units) as a function of mu